Eleftheria Zeggini
Andrew Morris
Editors

Assessing Rare Variation in Complex Traits

Design and Analysis of Genetic Studies

Springer

Assessing Rare Variation in Complex Traits

Eleftheria Zeggini • Andrew Morris
Editors

Assessing Rare Variation in Complex Traits

Design and Analysis of Genetic Studies

Springer

Editors
Eleftheria Zeggini, Ph.D.
Wellcome Trust Sanger Institute
Hinxton, UK

Andrew Morris, Ph.D.
Department of Biostatistics
University of Liverpool
Liverpool, UK

ISBN 978-1-4939-2823-1 ISBN 978-1-4939-2824-8 (eBook)
DOI 10.1007/978-1-4939-2824-8

Library of Congress Control Number: 2015944081

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

In the last 10 years, genome-wide association studies (GWAS) have revolutionised our understanding of the genetic basis of a diverse range of inherited complex human traits of medical importance (for example, body mass index, blood pressure and lipid profiles) and prevalent disorders including type 1 and type 2 diabetes and coronary artery disease. However, despite the success of GWAS in identifying regions of the genome associated with these complex traits, the observed association signals typically account for only a small percentage of the heritability. One important limitation of GWAS regions is that they are most often characterised by common variant association signals, each with only modest effect on the trait. Consequently, there has thus been increased expectation that much of the "missing heritability" will be accounted for by rare genetic variation (typically defined to occur in less than 1 % of the population).

The gold standard approach to assaying rare genetic variation is through sequencing studies, which has been prohibitively expensive, until recently, on the scale of the whole genome. Furthermore, the traditional GWAS methods available for assessing the evidence for association with complex traits are suboptimal for the analysis of variants with frequency less than 1 % in the population. However, with improvements in the cost efficiency of next-generation sequencing technologies, and the development of novel powerful analytical techniques, empirical evidence is emerging for a role for rare genetic variants in many complex traits, including *NOD2* in Crohn's disease, *IFIH1* in type 1 diabetes, *MYH6* in sick sinus syndrome and *G6PC2* in regulation of plasma levels of fasting glucose. Ongoing population-based whole-genome sequencing initiatives, such as the 1000 Genomes and UK10K Projects, are providing invaluable insight into the distribution and characteristics of rare genetic variation across diverse population groups and, through improved imputation techniques, are enabling cost-effective assessment of the association of tens of millions of variants with complex traits.

In editing this book, we have been fortunate to be able to call on colleagues leading research in all aspects of the design and analysis of rare genetic variants in complex human trait association studies, and we are indebted to them for their

invaluable contribution. As in many areas of genomic research, the next few years promise an exciting period of rapid advancement of technology, and development of efficient and powerful analytical tools to accommodate and interpret the vast quantity of genetic data that will be generated. The findings of these studies will be instrumental in refining our understanding of the biological and physiological basis of heritable human traits, and will enable the development of novel therapeutic interventions in clinical care that have the potential to reduce the burden of disease on limited public health resources.

Hinxton, UK — Eleftheria Zeggini
Liverpool, UK — Andrew Morris

Contents

Contributors

Jennifer L. Asimit Wellcome Trust Sanger Institute, Hinxton, UK

Adam Auton Albert Einstein College of Medicine, Bronx, NY, USA

Marie-Claude Babron Inserm UMR-946, Genetic Variability and Human Diseases, Paris, France

University Paris-Diderot, Sorbonne Paris Cité, UMRS-946, Paris, France

Han Chen Boston University School of Public Health, Boston, MA, USA

Antonio Ciampi Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

Josée Dupuis Boston University School of Public Health, Boston, MA, USA

Paul Flicek European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Menachem Fromer Division of Psychiatric Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Emmanuelle Génin Inserm UMR-1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest Cedex 2, France

Centre Hospitalier Universitaire de Brest Morvan, Brest, France

Jacqueline I. Goldstein Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Celia M.T. Greenwood Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

Departments of Oncology, Epidemiology, Biostatistics and Occupational Health, and Human Genetics, McGill University, Montreal, QC, Canada

Audrey E. Hendricks University of Colorado Denver, Denver, CO, USA

Suzanne M. Leal Department of Molecular and Human Genetics, Center for Statistical Genetics, Baylor College of Medicine, Houston, TX, USA

Sébastien Letort Inserm UMR-1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest Cedex 2, France

Centre Hospitalier Universitaire de Brest Morvan, Brest, France

Dajiang J. Liu Department of Public Health Sciences, College of Medicine, Pennsylvania State University, Hershey, PA, USA

Anubha Mahajan Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

Andrew Morris Department of Biostatistics, University of Liverpool, Liverpool, UK

Loukas Moutsianas Wellcome Trust Centre for Human Genetics Research, University of Oxford, Oxford, UK

Dawn Muddyman Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Benjamin M. Neale Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

Shaun Purcell Division of Psychiatric Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Andy Rimmer Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, UK

Genomics PLC, Oxford, UK

Graham R.S. Ritchie European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

University of Edinburgh, Edinburgh, UK

Manuel A. Rivas Wellcome Trust Centre for Human Genetics Research, University of Oxford, Oxford, UK

Neil Robertson Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK

Tovah Salcedo Albert Einstein College of Medicine, Bronx, NY, USA

Ioanna Tachmazidou Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Yik-Ying Teo Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Life Sciences Institute, National University of Singapore, Singapore, Singapore

Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore

Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore

Chris Tyler-Smith The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Xu Wang Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

ChangJiang Xu Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

Yali Xue The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Ilenia Zara Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

CRS4 (Centre for Advanced Studies, Research and Development in Sardinia), Pula, ltaly

Eleftheria Zeggini Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Calling Rare Variants from Genotype Data

Jacqueline I. Goldstein and Benjamin M. Neale

Introduction

As discussed in Chap. 2, DNA microarrays have been successfully used for human genetics research to assay single nucleotide polymorphisms (SNPs) throughout the genome. DNA microarrays work by measuring the relative amount of binding of input DNA to a set of complementary oligonucleotide probes for each allele using a photometric assay. Once the raw data are collected, they need to be converted into a genotype call automatically and with high accuracy. Over the past decade, many groups have published calling algorithms that are able to achieve greater than 99.5 % accuracy. However, these algorithms work best for common SNPs and are not as accurate for low-frequency and rare variants (minor allele frequency <5 %). With the widespread usage of microarrays targeting rare variants such as Exome Chip and MetaboChip, new calling algorithms that accurately call rare variants have been published over the last year. In this chapter, we will describe how DNA microarrays work (see section "Microarray Technology"), give a brief overview of genotype calling algorithms (see section "Genotype Calling Algorithms"), and summarize the different algorithms designed for rare variants and how well they perform (see section "Application to Rare Variants").

J.I. Goldstein • B.M. Neale (✉)
Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
e-mail: bneale@broadinstitute.org

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_1

Microarray Technology

Historical Background

The first DNA microarrays were created in the late 1980s with the advent of photolithography techniques (Fodor et al. 1991). By repeatedly applying light, masks, and modified nucleotides at specific locations on an array, custom nucleotide sequences were directly synthesized on a glass slide in a highly parallel and customizable manner. In the mid-1990s, the array technology was adapted for genotyping by synthesizing probes, which are 25 base pair nucleotide sequences complementary to an allele of a SNP of interest, and adding a labeling scheme to quantify the amount of DNA binding to probes (Chee et al. 1996). Affymetrix, a biotechnology company founded on this technology, went public in 1996 and started producing DNA microarrays for biomedical research. Since then, other companies have produced DNA microarrays including Illumina, Agilent, and Applied Biosystems. However, in this section, we focus on how Affymetrix and Illumina genotyping arrays work as they are the most commonly used for human genetics research.

Affymetrix

The earlier version of the Affymetrix array, the GeneChip, consists of tiled 25mer oligonucleotide probes directly synthesized onto a glass slide (Liu et al. 2003). Each SNP is assayed by 6–8 probe sets consisting of an octet of probe sequences: a quartet of a perfect match (PM) and mismatch (MM) probe for both the A and B alleles on the forward strand and a corresponding quartet for the reverse strand. The position of the SNP in each probe set differs by varying offsets to ensure adequate signal is obtained for each SNP. For example, the SNP in one probe set is the 9th nucleotide in the probe sequence and in another probe set is the 13th nucleotide. The role of the mismatch probes, which have identical sequences to the perfect match probe, except for the SNP of interest, is to measure unspecific hybridization or background noise. Biotin-labeled DNA is hybridized to the array and fluorophores then bind to the biotin. Next, a laser is shined onto the array and the amount of light intensity emitted by the fluorophores is measured for each square on the array consisting of millions of copies of the same probe. Finally, a genotype call is made based on which probes fluoresce in each quartet.

Recently, Affymetrix has developed a new array technology called Axiom, which has replaced the GeneChip family (Hoffmann et al. 2011). Like the GeneChip arrays, the Axiom arrays consist of tiled 30mer oligonucleotide probes bound to a glass slide that are complementary to a target DNA sequence. However, instead of the probe containing the SNP of interest, the probe terminates the nucleotide before the SNP. The assay begins when unlabeled DNA fragments hybridize to the probe. Next, short 9 base pair oligonucleotides are added to the array. There are four oligonucleotide sequences per SNP that are identical except for the first nucleotide corresponding to

the location of the SNP. Oligonucleotides beginning with an A or T are labeled with a red fluorophore and oligonucleotides beginning with a G or C are labeled with a green fluorophore. The oligonucleotides that are not complementary to the bound DNA fragment are washed away and the bound oligonucleotides are ligated to the glass-bound probes. Finally, the assay is finished when a laser is shined onto the array and the amount of light emitted is measured for each square.

Unlike GeneChips, the Axiom arrays do not have mismatch probes or offset probes for every SNP because they were found to not provide any additional information over the perfect match probes (Korn et al. 2008). In addition, for SNPs that are not strand ambiguous, both alleles can be measured by one probe, as opposed to GeneChips where each allele has to have its own feature on the array. Therefore, Axiom arrays are more efficient and allow more SNPs to be assayed on one chip.

Illumina

Illumina BeadChip microarrays consist of probes covalently bound to microscopic silica beads (Gunderson and Martin 2009). The beads are randomly assembled into microwells that have been etched into a glass slide. On average, each chip will have 20 beads with a given probe sequence. Probes consist of a 30 base pair decoding sequence, which is used to determine the identity of the bead in a particular microwell, followed by 50 base pairs that are complementary to the DNA sequence of interest. The probe either terminates at the SNP of interest (Infinium I technology) or one base pair before the SNP being assayed (Infinium II technology). After the DNA binds to the probe, a single base pair extension reaction occurs with nucleotides covalently bound to either biotin (adenine and thymine) or dinitrophenol (guanine and cytosine). Red fluorescently labeled antibodies bind to the dinitrophenol, and green fluorescently labeled antibodies bind to the biotin. When a fiber optic laser is shined onto a bead, the light intensity in both the red and green channels is measured, which corresponds to the nucleotide added to the probe sequence. The raw intensity data for each bead on the chip is compiled into a binary intensity data file (IDAT). Illumina's GenomeStudio software then reads the IDAT files and returns the raw red and green intensity values for each probe calculated from the intensities of all beads. GenomeStudio also automatically normalizes intensities by using a six degree of freedom affine transformation for each bead pool in order to facilitate comparison of intensity values across arrays (Teo et al. 2007).

Technical Confounders

The hybridization of a complementary probe to a target DNA sequence of interest is the key step in obtaining accurate genotype calls for a given SNP. If the probe sequence is not unique, and is complementary to multiple locations in the genome,

the resulting intensity data will be a mixture of all potential combinations of SNPs. If the probe is not perfectly complementary to its intended target, then the probability that the probe will hybridize to the target DNA is substantially lower, and the resulting intensity output will be extremely low or undetectable. If the GC content of a probe is too high, then the probe will not hybridize efficiently to the DNA resulting in low intensities that can be the same intensity as unspecific hybridization. In addition, poor DNA quality (e.g., highly degraded DNA) may also lead to both spurious genotypes for that individual sample. All of these technical confounders affect the quality of the assay and thus the ability to make accurate genotype calls.

Genotype Calling Algorithms

Once the raw intensity data have been obtained, a calling algorithm is used to determine a sample's genotype for each SNP on the array. For Affymetrix GeneChips, the amount of light intensity measured from the A and B allele probes indicates which alleles are present in a sample. Likewise, for Affymetrix Axiom arrays and Illumina BeadChips, the amount of light intensity measured in the red and green intensity channels for each probe indicates which copies of the allele are present in a sample. For example, if a sample has two copies of the A allele and no copies of the B allele (AA), then the amount of light intensity measured for the A allele will be much larger than that measured for the B allele. Similarly, if a sample has two copies of the B allele and no copies of the A allele (BB), then the amount of light intensity measured for the B allele will be much larger than that measured for the A allele. For samples with one copy each of the A and B alleles (AB), the amount of light intensity measured for both the A and B alleles is approximately equal. When the A and B intensity measurements are plotted in two-dimensional space, three distinct clusters are formed corresponding to each of the three possible genotypes (AA, AB, BB) (Fig. 1). Points are then assigned genotypes based on which cluster they are closest to. If genotype calls were only needed for one SNP, it would be easy to make calls based on a visual inspection of the cluster plot. However, most arrays contain at least 100,000 SNPs, making it impossible to call each SNP by hand. Therefore, for genotyping arrays to be useful for large-scale association studies, it is imperative to have a calling algorithm that automatically determines the location of each cluster and makes accurate genotype calls.

A variety of machine learning approaches and statistical models have been used to automatically detect cluster locations and assign genotypes to points. The earliest Affymetrix algorithm, MPAM (Liu et al. 2003), classified points by minimizing the distance between points of the same class and maximizing the distance between points of different classes. MPAM was later replaced by Affymetrix's DM algorithm (Di et al. 2005), an extension of ABACUS (Cutler et al. 2001), which classifies points based on likelihood scores. The likelihood score is derived from the expected intensities for each genotype class and the observed intensity values for a sample. Illumina's genotype caller, GenCall (Illumina, Inc. 2005),

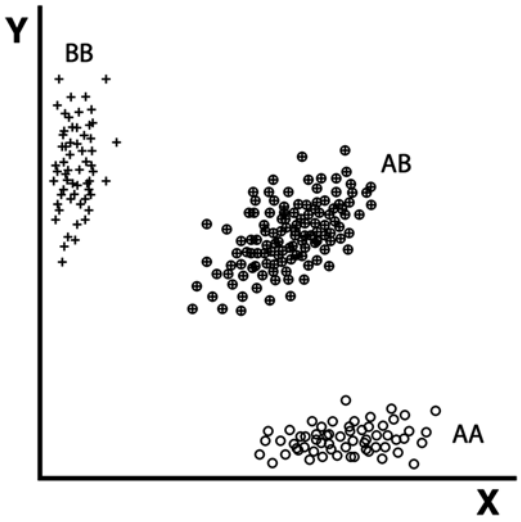

Fig. 1 Example of an intensity cluster plot. This plot demonstrates the two-dimensional intensity profile for a common SNP. The *x*-axis represents the amount of red intensity detected corresponding to the presence of the A allele while the *y*-axis represents the amount of green intensity detected corresponding to the presence of the B allele. The three genotype clusters corresponding to the two homozygous genotypes (AA, BB) and the heterozygote cluster (AB) are separated based on how much intensity is detected from each fluorophore

uses a proprietary clustering algorithm. RLMM (Rabbee and Speed 2006) uses a point's Mahalanobis distance from each cluster's predetermined bivariate distribution to assign genotypes. GEL (Nicolae et al. 2006), BRLMM (Affymetrix Inc. 2006), and CRLMM (Carvalho et al. 2007) use a Bayesian statistical framework to find the likelihood a point belongs to a genotype class based on bivariate cluster distributions. Illuminus (Teo et al. 2007) and Birdseed (Korn et al. 2008) use a Gaussian mixture model to determine the most likely position of the cluster distributions based on the observed data. CHIAMO (The Wellcome Trust Case Control Consortium 2007) uses a Bayesian four-class mixture model. The accuracy of genotype calling algorithms is limited by the quantity and quality of the data input into them. If there are not enough data points to form well-defined clusters, then the aforementioned statistical methods will not work. Therefore, two approaches are used to aggregate enough data points to allow for sufficient cluster formation: within-array and population-based.

Within-array algorithms leverage intensity information for all probes on the array simultaneously. The advantages of this approach are that it doesn't require additional samples to be genotyped for calls to be made, is easy to parallelize, and is not susceptible to batch effects and differences in DNA quality when aggregating data across many arrays. The disadvantages are that probes behave differently due to GC content and the uniqueness of the probe in the genome as described in the section "Technical Confounders." In addition, the absolute light intensities measured vary across the

array based on where probes are with respect to the camera. As a consequence, when data are aggregated for all probes on the array, the resulting clusters are more widely distributed making it more difficult to distinguish them. In contrast to within-array algorithms, population-based algorithms aggregate data across many arrays (on the order of 100–10,000) and cluster data for each SNP separately. The advantage of this approach is that it is not as sensitive to how well a specific probe hybridizes to its target as long as the signal to noise ratio is high. However, it is more sensitive to differences in DNA quality and experimental conditions across arrays that can cause the clusters to not be as easily differentiated. Finally, other algorithms such as MAMS (Xiao et al. 2007), M(3) (Li et al. 2012), and optiCall (Shah et al. 2012) utilize information from both within-array and population-based clustering in order to incorporate the advantages of both approaches.

For many projects, information besides the intensity data is available that can be used to improve the accuracy of the calls. For example, BRLMM and Birdseed use training datasets with samples of known genotypes in order to predetermine the bivariate cluster distributions. HapMap samples are the most commonly used training dataset because their genotypes are known with high confidence from the HapMap Project (International HapMap Consortium 2005). SNPCaller (Lin et al. 2008) uses the pedigree structure of genotyped samples in order to avoid Mendelian errors for projects with trios. Chiamante (O'Connell and Marchini 2012) uses sequencing data, if available, to inform where the genotype clusters lie and make calls for points even when the array data are ambiguous. BeagleCall (Browning and Yu 2009) utilizes the linkage disequilibrium structure of the genome to impute calls for data points that do not match a specific genotype cluster. Birdseed and BRLMM use a sample's reported gender to make more accurate genotype calls for the sex chromosomes. zCall (Goldstein et al. 2012) and BRLMM utilize calls from another genotype calling algorithm (GenCall and DM, respectively) in order to have a starting point for the position and bivariate distribution of each cluster.

When assessing how well a genotype calling algorithm works, the three most widely used metrics are the concordance with known genotypes, the call rate, and the percentage of SNPs that are consistent with Hardy–Weinberg equilibrium (HWE). Most studies have calculated concordance by comparing an algorithm's calls to HapMap data, or another dataset, where the genotypes are known. However, family data (with or without known genotypes) can also be used to determine a calling algorithm's accuracy by calculating the number of Mendelian errors. Also, it is important to assess the accuracy of a caller for each genotype class individually. If a calling algorithm has significantly different accuracies for each genotype class, the association tests will be biased and false positives and false negatives will be introduced (Nicolae et al. 2006). Like accuracy, the call rate of an algorithm is also important. For example, if a calling algorithm is extremely accurate when it makes a call, but only assigns calls to 75 % of the data, this leads to a dramatic loss in power for the call set. Many calling algorithms have confidence scores available that allow the user to make trade-offs between the accuracy of calls made and the call rate. Finally, deviations from HWE are used because if the resulting calls from an algorithm are out of HWE, then this is indicative of a lot of inaccurate calls being made or a problem with the assay (Chap. 5).

In summary, genotype calling algorithms use the relative amount of intensity measured for the A and B alleles to make a genotype call for each sample based on a statistical model for where the genotype clusters lie in two-dimensional space. Some genotype calling algorithms also incorporate information besides the intensity data to make more accurate calls. Finally, evaluating the accuracy and call rate of an algorithm against previously established benchmarks is essential before the algorithm's genotype calls can be used for association studies.

Application to Rare Variants

As described in Chap. 2, genotyping arrays specifically targeting lower-frequency variation (minor allele frequency <5 %) have been produced by both Affymetrix and Illumina and utilized in a variety of human genetics association studies. However, the genotype calling algorithms that were used for older arrays with only common SNPs do not work well for rare variants because they assume three genotype clusters exist when clustering data, which is not the case for rare variants. For example, for a variant with a minor allele frequency of 1 % in the population, one would need to genotype 100,000 samples in order to have an expectation of ten points in each genotype class if the variant is in Hardy–Weinberg equilibrium. In addition, it is more difficult to benchmark how well an algorithm works for rare variants. Most samples, including HapMap samples, have not been assayed to the same extent for rare variants as they have for common variants unless they have sequencing data available. Using family data does not necessarily circumvent this problem; if the rare genotype is not detected and everyone in the family is called a common allele homozygote, no Mendelian error will be made. Analogously, calculating the overall concordance of an algorithm to known genotypes is uninformative because it does not capture how well genotypes with rare alleles are called. For example, a calling algorithm that called all points common allele homozygotes would be correct ~99.8 % of the time for a SNP with a MAF = 0.1 %. Therefore, finding an appropriate comparison dataset and assessing a calling algorithm's accuracy is more challenging for rare variants than it is for common SNPs. For the remainder of this section, we will describe a number of calling algorithms that were written for rare variants and consider how well they work in comparison to existing methods.

GenCall

GenCall is the default algorithm in Illumina's GenomeStudio software (Illumina, Inc. 2005). It uses a proprietary algorithm to do initial clustering. However, when less than three clusters are observed after the initial clustering, neural networks are used to predict the locations of the unobserved cluster(s). Although Illumina did not test how well GenCall works for rare variants using actual data, they used simulations

with real data to quantify how well they expect it to work. They found that they could achieve greater than 99 % call rate for all genotype classes while maintaining an accuracy of 99.9 % for all genotype classes at a MAF = 0.5 % (http://www.illumina.com/documents/products/technotes/technote_genotyping_rare_variants.pdf).

Birdseed

Birdseed is an algorithm developed for the Affymetrix SNP 6.0 array (Korn et al. 2008). It uses a two-dimensional Gaussian mixture model that requires an input training dataset with known genotypes (usually HapMap samples) in order to initialize the location of the clusters. Next, a series of expectation–maximization (E–M) steps are done until parameters in the Gaussian mixture model converge. For rare variants, Birdseed imputes the position of missing clusters if they were not observed. The authors of Birdseed compared their algorithm to BRLMM for MAF = 5 % using HapMap samples. They found both Birdseed and BRLMM called common allele homozygote genotypes accurately (~99.87 % and 99.90 %, respectively). However, they both were significantly worse at calling heterozygote genotypes (~99 % and 98.7 %) and minor allele homozygote genotypes (≈95.5 % and 95 %). The authors did not report how well Birdseed works for MAFs less than 5 %.

GenoSNP

GenoSNP is a solely within-array genotype calling algorithm that is able to accurately genotype rare variants without the need for a large reference population (Giannoulatou et al. 2008). The algorithm works by clustering data for all SNPs within one bead pool on a single array in order to minimize intensity differences due to probes in different bead pools being manufactured at different times and located in different locations on the array. A four-component mixture model of Student's *t*-distribution is used to model the intensity data, and either an E–M algorithm or a modified version of variational Bayes E–M algorithm is used to find the optimal model given the input data. Genotype calls are made based on which genotype has the highest likelihood calculated from the optimal model. The authors compared their method to GenCall and Illuminus using 120 HapMap samples genotyped on the Illumina HumanHap300Duo array. The accuracy of the methods was determined by comparing the calls made by each algorithm to those publically available in the HapMap database. GenoSNP had the highest accuracy for heterozygote calls among the three methods (99.738 %), but had the lowest accuracy for homozygote calls (99.264 % for GenoSNP while 99.823 % for GenCall). Although GenoSNP did not explicitly assess the performance of their method for rare variants, they would have achieved comparable accuracy rates for each genotype class as they do for common SNPs due to their within-array clustering approach.

ALCHEMY

ALCHEMY is a genotype calling algorithm intended for projects with small numbers of samples or highly inbred populations where most samples are homozygotes (Wright et al. 2010). Briefly, the algorithm utilizes a Bayesian statistical framework where the parameters are the probabilities that A allele and B allele are present based on the data and the expected allele frequencies. To get the probability of the A and B alleles being present in a sample, a bivariate mixture of Student's *t*-distributions is used to model the mean signal and noise components of the intensity distribution for each allele across all samples for a given SNP. An E–M algorithm is then used to find the best parameters for the mixture model, which are then used to calculate likelihoods for each genotype class. Inbreeding coefficients and expected minor allele frequencies can also be incorporated into the model as priors in order to make more accurate calls. The authors found that BRLMM-P performed slightly better than ALCHEMY when comparing the accuracy of calls to 270 HapMap samples genotyped on the Affymetrix Human 500 K GeneChip. To test whether their algorithm worked better than BRLMM-P for small sample sizes, the authors genotyped two distinct rice lines and the resulting progeny of the cross on a custom array. Accuracy was determined by only looking at SNPs where the consensus ALCHEMY calls from the full dataset followed a Mendelian mode of inheritance and calculating how many genotype calls were correct when calling between 1 and 72 samples concurrently. ALCHEMY performed better than BRLMM-P for smaller sample sizes (99.2 % accuracy for 18 samples compared to 88.7 % for BRLMM-P). However, the authors note that ALCHEMY does not perform as well for heterozygote calls compared to homozygote calls with small sample sizes. The authors did not report how using ALCHEMY calls as their truth dataset biases their results.

M³

M^3 uses a two-stage calling procedure in order to take advantage of calling genotypes within an array and across SNPs (Li et al. 2012). The first stage clusters each SNP in a population-based manner using a Gaussian mixture model. The second stage recalls SNPs with a low minor allele frequency or poor average posterior likelihood scores calculated from the original genotype calls using cluster properties derived from another reference SNP that has similar intensity properties to the SNP being recalled. To determine how well M^3 performs for rare variants compared to other methods, the authors obtained calls for 141 Illumina Omni 1 M arrays from 38 unique HapMap samples using GenCall, GenoSNP, and M^3. When requiring that no errors were made in classifying homozygotes compared to known HapMap genotypes for all three callers (in order to minimize errors due to reference strand errors), they found that M^3 had the highest accuracy for all three MAF bins analyzed

(99.20 %, 99.11 %, and 98.77 % for MAF < 0.1 %, MAF < 0.05 %, and MAF < 0.01 %, respectively). M^3 also had the highest call rate among the three algorithms. No differentiation was made between the accuracy of different genotype classes.

optiCall

optiCall utilizes information from both across samples and across SNPs simultaneously in order to call SNPs of all minor allele frequencies accurately (Shah et al. 2012). The algorithm works by first creating a four-class mixture model of Student's *t*-distributions where the input intensity data to the model are randomly sampled from across SNPs and samples. Next, each SNP is clustered and called separately in a population-based approach using a separate four-class mixture model informed by priors from the across SNP/across sample mixture model. The authors compared the accuracy of optiCall to GenCall, Illuminus, and GenoSNP for rare variants by hand-calling 600 SNPs from an Illumina ImmunoChip that were called as monomorphic (MAF = 0 %) by one algorithm, but had a minor allele frequency between 4×10^{-4} and 0.01 in two other algorithms. They found that Illuminus had misclassified the most SNPs as being monomorphic (354), while optiCall only misclassified 1 SNP compared to 3 for GenoSNP and 13 for GenCall. However, the authors did not note the overall accuracy of each algorithm that correctly called a SNP as variant (MAF > 0 %) among the 600 rare SNPs that were manually called and did not differentiate between the calling accuracies of common allele homozygotes and rarer genotypes. The authors state that their method is sensitive to intensity outliers and therefore recommend removing them from the data before running optiCall.

Chiamante

Chiamante is a genotype calling algorithm for microarrays that has the ability to incorporate information from sequencing data to make more accurate genotype calls using a Bayesian framework (O'Connell and Marchini 2012). They used Illumina Omni 2.5S microarray data and 4× sequencing data from the 1000 Genomes Project to test how well their method works. Affymetrix Axiom data with genotype calls from the Axiom GT1 algorithm (modified version of BRLMM for Axiom arrays) is used as the truth dataset. For rare variants (MAF < 5 %), the authors found that Chiamante with the addition of sequencing data outperformed GenoSNP, Illuminus, and GenCall. The overall concordance of Chiamante + sequencing with Axiom genotype calls for major allele homozygotes was 99.78 %, for heterozygotes was 98.11 %, and for minor allele homozygotes was 95.63 %. However, the authors note that because they used the Axiom GT1 genotypes as a truth dataset and the Axiom GT1 calling algorithm doesn't work as well for rare variants, the accuracies reported are overestimates.

zCall

zCall is implemented as a post-processing step for GenCall in order to recover rare variation that GenCall called as No Calls using outlier detection (Goldstein et al. 2012). To assess the performance of their algorithm for rare variants, the authors used Exome Chip data and compared the calls they obtained from both GenCall and zCall to whole-exome sequencing calls for singleton SNPs in the sequencing data. They used the SNP-wise concordance (SWC), which is the percentage of passing SNPs divided by the total number of SNPs, to determine how well a caller works for rare variants. A SNP is considered passing if all genotypes are concordant with the sequencing calls. However, they did allow for common allele homozygotes to be called as a No-Call because this error mode wouldn't affect association test statistics very much. The SWC of GenCall for singleton SNPs was 92.49 % and 99.12 % for zCall. They also compared zCall to optiCall and found using an independent dataset that the SWC for singleton SNPs in exome sequencing data for GenCall was 93.12 %, optiCall was 98.21 %, and zCall was 99.27 %. zCall works well at recovering missed variation from GenCall. However, it is also very sensitive to data quality. If the data have a low signal to noise ratio or batch effects exist in the data, then zCall will not work as well.

iCall

iCall is a multi-sample, single SNP calling algorithm designed to work for both common and rare variants (Zhou et al. 2014). iCall utilizes the framework of the Illuminus algorithm where parameters representing the likely cluster positions are estimated from an E–M algorithm and a three-component Student's *t*-mixture model is used to make genotype calls (Teo et al. 2007). iCall expands on Illuminus by using a set of novel penalty functions to guide the E–M algorithm to account for situations in which a SNP has a low minor allele frequency, the dataset has a small number of samples, or the intensity profile of the assay is significantly different than what is expected. To assess the performance of iCall, the authors compared their algorithm to GenCall, optiCall, Illuminus, and GenoSNP using 81 samples that had been genotyped on the Illumina Exome Chip and whole-genome sequenced. The performance of each algorithm was estimated from 13,542 common SNPs, 1,356 low-frequency variants, and 1,530 rare variants that were polymorphic in the sequencing data and had been genotyped on the array. The authors found that iCall had the highest overall concordance rate and the highest minor allele concordance rate for rare SNPs, correctly identifying 97.435 % of heterozygous and minor allele homozygous calls from sequencing with an input sample size of 1,000. However, iCall was more conservative when making genotype calls for rare variants and consistently had a higher missed minor allele call rate than optiCall (3.422 % compared to 3.076 % for an input sample size of 1,000). The authors do not state how often

iCall calls a true major allele homozygote call as a heterozygote or minor allele homozygote. Finally, the authors applied zCall to genotype calls made from optiCall, GenCall, and iCall and found that the application of zCall always improved the concordance for all three callers. The combination of GenCall + zCall performed slightly better for common SNPs (97.683 % versus 97.681 %), while iCall + zCall performed better for low-frequency and rare SNPs (97.656 % versus 97.608 %).

Conclusion

In this chapter, we have described the technology used in genotyping arrays, provided an overview of genotype calling algorithms, and surveyed the literature for how well existing genotype calling algorithms work for rare variants. We would like to note that the focus from genotyping common variants to rare variants has happened quite recently. Therefore, more research will be done in the near future as genotyping arrays with rare variants are more widely adopted.

References

Affymetrix, Inc.. BRLMM—an improved genotype calling method for the GeneChip® Human Mapping 500K Array Set. 2006. pp. 1–18.

Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85:847–861. doi:10.1016/j.ajhg.2009.11.004

Carvalho B, Bengtsson H, Speed TP, Irizarry RA (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. Biostatistics 8(2):485–499

Chee M et al (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–614

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678. doi:10.1038/nature05911

Cutler DJ et al (2001) High-throughput variation detection and genotyping using microarrays. Genome Res 11(11):1913–1925

Di X et al (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays

Fodor SP et al (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251:767–773

Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC (2008) GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics 24:2209–2214. doi:10.1093/bioinformatics/btn386

Goldstein JI et al (2012) zCall: a rare variant caller for array-based genotyping: Genetics and population analysis. Bioinformatics 28:2543–2545. doi:10.1093/bioinformatics/bts479

Gunderson, KL (2009) Whole-genome genotyping on bead arrays. In: Dufva M (ed) DNA microarrays for biomedical research: Methods and protocols. Humana Press, a part of Springer Science+Business Media, LLC. doi:10.1007/978-1-59745-538-1_13

Hoffmann TJ et al (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98:79–89. doi:10.1016/j.ygeno.2011.04.005

Illumina, Inc. Illumina GenCall Data Analysis Software. Technology Spotlight. 2005. http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf%3E.

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320. doi:10.1038/nature04226

Korn JM et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40:1253–1260. doi:http://www.nature.com/ng/journal/v40/n10/suppinfo/ng.237_S1.html.

Li G, Gelernter J, Kranzler HR, Zhao H (2012) M(3): an improved SNP calling algorithm for Illumina BeadArray data. Bioinformatics 28:358–365. doi:10.1093/bioinformatics/btr673

Lin Y et al (2008) Smarter clustering methods for SNP genotype calling. Bioinformatics 24:2665–2671. doi:10.1093/bioinformatics/btn509

Liu W-M et al (2003) Algorithms for large-scale genotyping microarrays

Nicolae DL, Wu X, Miyake K, Cox NJ (2006) GEL: a novel genotype calling algorithm using empirical likelihood

O'Connell J, Marchini J (2012) Joint genotype calling with array and sequence data. Genet Epidemiol 36:527–537. doi:10.1002/gepi.21657

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12. doi:10.1093/bioinformatics/bti741

Shah TS et al (2012) optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics 28:1598–1603. doi:10.1093/bioinformatics/bts180

Teo YY et al (2007) A genotype calling algorithm for the Illumina BeadArray platform

Wright MH et al (2010) ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. Bioinformatics 26:2952–2960. doi:10.1093/bioinformatics/btq533

Xiao Y, Segal MR, Yang YH, Yeh RF (2007) A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. Bioinformatics 23:1459–1467. doi:10.1093/bioinformatics/btm131

Zhou J et al (2014) iCall: a genotype-calling algorithm for rare, low-frequency and common variants on the Illumina exome array. Bioinformatics 30(12):1714–1720. doi:10.1093/bioinformatics/btu107

Calling Variants from Sequence Data

Andy Rimmer

Introduction

Detecting genetic variation from high-throughput sequencing data (variant calling) is a difficult and computationally intensive task and continues to be the subject of much research. Current sequencing technologies produce short (~100 base pair) reads; these are then typically matched to a reference sequence, before being used to identify potential variation between the reference and the sample being sequenced. There a number of methods which can be used to identify variants, and different methods are more effective for different types of variant. In the following sections, we will introduce several of the most popular methods. First though, we need to discuss sequencing data more generally and introduce a few of the various file formats that exist for storing this data.

Next-Generation Sequencing Data

The last decade has seen a vast increase in the amount of DNA sequence data being generated around the world. Various technologies exist, but the current dominant platform is the Illumina HiSeq machine, and we will be focusing almost exclusively on this for the purposes of this chapter, though most of what is said will also be relevant to other platforms. The Illumina machines produce millions (in some cases billions) of short reads, normally in pairs, of about 100 base pairs each. The volume of data produced by a typical run of these machines is large (tens of gigabytes), and

A. Rimmer (✉)
Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK

Genomics PLC, Oxford, UK
e-mail: andy.rimmer@genomicsplc.com

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_2

so we need an efficient way of storing this data. The favoured data formats (and the ones you will most likely have to deal with) for storing these reads are called BAM and FASTQ.

DNA Sequence Reads

Sequencing data is produced in the form of *reads*. A read is a string of nucleotide sequences. Each read also has a sequence of *quality scores*, one for each sequenced nucleotide. The nucleotide sequence tells us the machine's interpretation of the biological sequence, and the quality score tells us how accurately the machine thinks its output represents the biological sequence. Low-quality scores mean low confidence in the nucleotide sequence.

Single and Paired Reads

Reads can be produced either in pairs or one at a time. You need to know which, in order to deal correctly with the data. Sequencing reads in pairs give more information, as the pair of reads is known to be close together in the biological sequence (typically within 100–1,000 bp), and this can be used to help significantly in later analysis. The read pairs can normally be selected during sequencing, so that the distance between reads is similar across pairs; this distance (the distance on the biological sequence between the start of one read and the end of the other) is known as the *insert size*.

The FASTQ File Format

FASTQ is the standard file format for representing raw sequencing data. It stores both the nucleotide sequence and the corresponding sequence of quality scores for each read. There is a description of the format on Wikipedia (http://en.wikipedia.org/wiki/FASTQ_format). Roughly, a FASTQ file contains a string of bases for each read, along with a quality score for each base, which describes how confidently that base was identified by the sequencer. FASTQ files are human readable (although the quality scores are reported using ASCII codes (http://en.wikipedia.org/wiki/ASCII) rather than simple numbers and are difficult for most people to interpret). If your data arrives in FASTQ format, then it has not been aligned to a reference sequence, and you will need to do this before you can start identifying variants (unless you are using an assembly-based approach, but see later for more on this).

FASTQ files may also come in a compressed form, in which case the file names will normally end in *.gz* (if compressed with gzip) or *.bz2* (if compressed with bzip2). Compressed FASTQ files are not human readable unless you un-compress them first (see www.gzip.org or www.bzip.org for more information on these standard compression tools).

The BAM File Format

BAM is a dedicated file format designed to store next-generation sequencing reads that have been aligned to a reference sequence. It is a good format for efficiently storing and accessing this sort of data, particularly if the reads are sorted by reference coordinate. It does, however, require special programs to read, write and otherwise manipulate it. If you are happy with Linux command-line tools, then you might want to look at Samtools (http://samtools.sourceforge.net/), which has a set of basic BAM-processing utilities, and also a variant caller (more on that later). If you want a visual display of the data in a BAM file, then try the Integrative Genomics Viewer (IGV, available from http://www.broadinstitute.org/igv/). There are many other tools for manipulating BAM files, which we will not discuss further here.

The BAM format also has a human-readable equivalent, called SAM, which contains exactly the same data but can be read using standard text-reading programs *less* or *more* if you are using Linux/Mac OS X, but do not try to open these files in Excel as they can be huge.

Header and Alignment Sections

A BAM file is divided into two parts. There is a *header* section, which contains meta-information about the content of the BAM file, and an *alignment* section, which contains the actual read data.

Other Sequencing Technologies

There are several commonly used next-generation sequencing platforms. Thankfully, the output from all of these can be stored in BAM or FASTQ format, as these formats are designed to be platform agnostic. Because of this, most of the tools we will discuss can be used on data from any sequencing platform. Results may vary however, since some tools will be optimised for specific platforms, and will deal more effectively with the particular kinds of problems which occur on that platform.

Experimental Design

Before discussing data processing in more detail, it is worth going over a few aspects of data generation which really need to be decided early. If you want to detect certain kinds of variation in your data, then you need to generate the right kind of data; otherwise it may be difficult or impossible to perform the desired analysis. Here are a few things you may wish to consider.

Platform: Which sequencing platform to use? Is quality or quantity of data the main priority? Do you want longer reads or more accurate reads?

Paired or Single: Do you want to sequence reads in pairs (the answer is probably yes)? If so, how large should the insert size be?

Genome, *Exome or Targeted Sequencing*: Do you want to sequence the complete genome of your samples (known as *whole-genome sequencing* or *WGS*), or is it enough to sequence just the exons of known genes? (*whole-exome sequencing* or *WES*). Perhaps you only want to sequence a specific set of genes/contigs. These decisions will have a significant impact on the scope and accuracy of variant detection which is possible with your data. Obviously you cannot detect variants in parts of the genome you did not sequence, but there are other effects. Detecting large variants is much easier with whole-genome sequencing, as this gives quite even, contiguous coverage across the genome, except in some very repetitive regions. If you are interested in large insertions, deletions, copy-number variants or large re-arrangements of any kind, then you may need to consider whole-genome sequencing: these events can be detected with exome or targeted sequencing, but it is much harder.

Pooled or Multiplexed: Are you sequencing multiple samples together? If so, you can either use bar-coded multiplexing or pooled sequencing. With multiplexing, you can sequence many samples together but then separate them afterwards, using a sample-specific tag. Pooled sequencing results in the loss of sample information: you just get many reads and have no idea which read came from which sample. This makes it impossible to infer per-sample genotypes. Pooled sequencing is not generally recommended, unless absolutely necessary because of financial constraints, for example.

Processing Sequence Data

This section explains how to get from raw sequencing data to something which is ready to be used for variant calling. We introduce some new data formats and summarise some of the techniques used.

Removing Adapter Sequences

If you are sent raw FASTQ files, then it may be necessary to do some basic processing on the data before anything else. In particular you should check for contamination by adapter sequences. When the insert size of a fragment is smaller than the length of a read, the sequencer will continue sequencing into the *adapter*. The adapter is a short sequence tag that is attached to the fragment during sequencing. If the adapter is sequenced, then sequence of nucleotides in the contaminated read will be part biological sequence and part adapter sequence. This is not desirable and can lead to problems with downstream processing (particularly mapping and

variant calling). Consequently, we recommend that you use a program to strip out any such contamination at the earliest possible stage. There are several programs for this; a popular program is *cutadapt* (http://code.google.com/p/cutadapt/).

Mapping and Aligning Sequence Data to a Reference Sequence

Once you have your sequence data and have established exactly what it is, the first task (if this has not already been done for you) is to map and align the sequence data to a reference sequence. A fairly reliable way to tell if your data has already been mapped is to check the data format. If your data is in FASTQ format, then it is not aligned to a reference sequence. If your data is in BAM/SAM format, then it is.

For mapping, you will need your data, a mapping program and the latest version of the standard human (or relevant species) reference sequence. Reference sequences are available from NCBI (http://www.ncbi.nlm.nih.gov/RefSeq/) in FASTA format.

The FASTA Format

The FASTA format (http://en.wikipedia.org/wiki/FASTA_format) is the standard way of storing and accessing reference DNA sequences. It is simple and human readable, although FASTA files may come compressed with gzip or bzip2.

Mapping Software

There are many programs designed to map and align human DNA to the reference sequence. Most will take FASTQ sequence files as input, along with a reference sequence in FASTA format, and output a BAM or SAM file (remember these are the same format, just one is text and one binary and you can convert between them using programs such as Samtools). The most widely used programs for this are BWA (http://bio-bwa.sourceforge.net/) and Stampy (http://www.well.ox.ac.uk/project-stampy). It is extremely important to use a good mapping program; otherwise you may lose vital information, making certain variant-calling tasks impossible. BWA and Stampy are fairly reliable tools: BWA is faster, and Stampy is better for dealing with sequences that are significantly diverged from the reference.

The BAM format contains a value for each read, called the *mapping quality*, which tells you how accurate the mapping is for that read. Specifically, the mapping quality encodes the probability that the read is mapped to the wrong position as a *PHRED* score (http://en.wikipedia.org/wiki/Phred_quality_score). A low mapping quality means that the read may be mapped to the wrong place, i.e. that there are other locations in the genome which contain similar sequence. The mapping quality is computed by the read mapper and output into the BAM/SAM file.

Make sure to check the documentation of the mapping program that you use, so you understand what the output contains.

Sorting and Filtering Data and Quality Checks

Once you have a BAM file (or many BAM files, as it is best to keep each sample in its own BAM file, but this depends on your experiment), with your data aligned to the reference sequence, you need to make sure that it is ready to be used for variant calling. Exactly what is required will depend on the variant caller you are using, but a standard requirement is that the reads in the BAM file be in coordinate sorted order and that each BAM file has an associated index file. It may also be necessary to merge data from several sources into a single BAM file or to filter the data in the BAM file to remove low-quality reads and/or certain sequencing artefacts.

If the read mapper you used to create the BAM file(s) does not output the reads in coordinate sorted order (most do not), then you can use a program like Samtools or Picard (http://picard.sourceforge.net) to sort the reads. This is straightforward, and most of these programs work out of the box with no problems. The only issue you may need to consider is disk space: BAM files are large, and sorting a BAM file requires copying the data, so you need at least twice as much space for the sorting to work. Make sure to check this first, as sorting can take quite a long time, and you do not want it to fail right at the end due to lack of space.

After sorting, you will need to create an index file for each BAM file. The index file can only be created from a sorted BAM file, so you must do the sorting first. The index file is a utility that allows programs to quickly locate reads by coordinate. Many programs rely on the existence of an index file and will not work without one. The convention is to name index files in the same way as the BAM files, but with a *.bai* extension, e.g. *test.bam* has index file *test.bam.bai*. You can create an index file using *Samtools*.

If you have data from several different sources that you want to merge into a single BAM file, then use *Picard*. You can also use Samtools for this, but it does not correctly deal with merging the *header* components of multiple BAM files, so *Picard* is recommended instead.

Checking BAM Data for Quality

It is a good idea to check your BAM files at this stage, to make sure there are no serious problems with the data. In particular you should make sure that you have sequenced the right samples, that you have coverage of the desired genomic locations and that there is minimal contamination. You could use a dedicated program such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for this. Here are a few things to check:

Coverage: Do your samples have coverage in the expected genomic locations? If so, is it enough (as a rough guide, at least 10× coverage per base is recommended for basic variant calling, but this depends on your experimental design)?

Duplication Rate: There will be a certain fraction of read pairs that are identical to other pairs in the same sample. This is often due to sequencing the same fragment multiple times. It is a good idea to remove duplicate reads. Check the proportion of reads in your data that are duplicates.

Contamination: Check for the presence of unexpected sequence. If there is a large amount of sequence that cannot be mapped to the reference, then something may be wrong.

Insert sizes and Read lengths: Check the distribution of insert sizes in your data. Does this match the expected distribution?

Base and Mapping Qualities: Check for reads with very low mapping qualities or very low base qualities.

Variant Calling

In this section we discuss variant calling and describe the different types of variant that can be detected with current methods and describe some of those methods.

Variant Types

Variant-calling algorithms are often optimised for specific types of variant, so we will review these now, in roughly ascending order of size.

SNPs: Single nucleotide polymorphisms (SNPs) are probably the best studied class of variant. Most variant-calling tools presented in the literature are optimised for finding these. SNPs are single base changes with respect to the reference sequence and occur in roughly one base per thousand in normal human genomic DNA. After alignment to the reference, base differences can be easily spotted.

MNPs: Multi-nucleotide polymorphisms are sequences of two or more bases which are adjacent and different from the reference sequence. These occur less frequently than SNPs. MNPs are not difficult to identify, but some variant callers report them as multiple individual SNPs, which can complicate downstream processing.

Short Insertions: Short (defined here as less than 100 base pairs although there is no consensus) insertions are less common than SNPs, but still occur roughly every 10,000 base pairs or so (check numbers). They are much more common in certain parts of the genome, depending greatly on the sequence context. Specifically, insertions are very common in long repetitive sequences, e.g. homopolymers and dinucleotide repeats. Insertions can be difficult to identify, since they contain sequence which may not be present in the reference. For accurate insertion calling, it is very important to use a good read mapper (e.g. Stampy, BWA).

Short Deletions: These occur at a similar frequency to short insertions. Deletions are easier to deal with than insertions, but still require accurate read mapping and alignment.

Complex Replacements: A loosely defined category, which encompasses all small variants which do not fit neatly into the previous categories, for example, variants which include both an insertion and one or more base changes. These are often reported as multiple SNP/indel variants

Inversions: Inversions are sequences of DNA that are reversed with respect to the reference. These come in various sizes. Large inversions are difficult to identify correctly. Small inversions are often called multiple variants, e.g. one insertion (of reversed sequence) plus one deletion (of normal sequence).

Large Insertions: Large insertions (more than 100 base pairs) are very difficult to call correctly, as the inserted sequence will often span a whole read, and that read does not map anywhere in the reference sequence. These are best identified with an assembly-based approach (see later).

Large Deletions: Large deletions are somewhat easier to spot, as the surrounding sequence can be mapped to the reference. Again, an assembly-based approach is probably the best way to find these.

Tandem Repeats: A special class of small insertion/deletion that occurs in locally repetitive sequence. These can be difficult to identify correctly if the total length of the repeat is much longer than a read length.

Copy Number Variants: Similar to tandem repeats, but with a larger repeated unit. These can span large genomic intervals (many kb) and are difficult to call accurately.

Structural Variants: A general term for large events which do not fit neatly into the other categories. These include whole chromosomal duplications, translocations, gene-fusion events, etc. These are all difficult to call from sequencing data.

The Variant Call Format (VCF)

Thanks to work carried out as part of the 1000 Genomes Project, there is now a standard data format for reporting variant calls, which is used by most variant-calling programs and many downstream processing tools. This is the Variant Call Format (VCF) (http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-formatversion-41). VCF allows the representation of all kinds of small variants and, in principle, large variants as well, though the files can become large. The format has a number of fixed fields, i.e. data which must be present, but also allows for arbitrary annotations to be added to each call. Genotype information can be present (but is not required), and multiple samples can be included in a single VCF file. Both phased and unphased genotypes are supported.

Reporting Indel Alleles in VCF

There is substantial room for ambiguity when reporting insertion or deletion alleles in repetitive sequence. It is often possible to report the same allele in a number of different positions without changing the overall sequence (e.g. an *A* insertion in the sequence *AAAAAAA* could be reported anywhere in the sequence). The convention is to report all alleles as far to the left as possible; this is not a requirement of the VCF standard, but most variant callers now do it, and it is a good idea; otherwise it becomes very difficult to compare indels; check your variant caller's documentation.

Complex Variants

As mentioned before, the reporting of MNPs and complex substitutions is not standardised, and one variant caller might report the substitution ACAC → AAAA as 2 SNPs, whilst another reports it as a single replacement, and yet another reports a deletion of ACAC followed by an insertion of AAAA.

Genome VCF

A recent innovation is for some variant callers to report, in VCF, not just the variant calls, but also hard reference calls, i.e. regions for which the caller was confident that only the reference sequence exists. This is useful for discriminating cases where there is little or no coverage from cases where there is coverage, and it clearly supports the reference. There is not currently a standard way of reporting this.

SNP Calling

SNPs are probably the most studied form of genetic variation, and SNP calling is the most mature form of variant detection. In principle, once the reads are correctly aligned to a reference sequence, then SNPs can be identified by simply reading off the base changes between the reads and reference. In practice it is not quite this simple, as various problems result in fake SNP candidates appearing, and a naïve approach to SNP calling will result in a large number of false SNPs being called. Here are a few things that can cause problems:

Sequencing Errors: Put simply, the sequencer gets it wrong, and an incorrect base is present in the read. This is sometimes reflected in a low-quality score for the base in question, but not always.

PCR Errors: Copying errors occur during PCR amplification, resulting in reads with incorrect bases. This can be exacerbated by sequencing the same PCR fragment multiple times, which is one reason for getting rid of duplicate read pairs.

Contamination: Your data contains a small amount of DNA from another sample or something else which was sequenced at the same time (e.g. bacteria/virus).

Bad Mapping/Alignment: If the read mapper puts reads in the wrong place, this can result in apparent base mismatches. This may happen systematically at certain sites. This is why you must use a good mapper.

Unidentified Insertion/Deletion nearby: If there is a real variant which has not been identified (e.g. a large insertion/deletion), then reads flanking that variant may not be aligned correctly, resulting in apparent SNPs. Even good mappers will not be able to identify all insertions/deletions correctly, so this is likely to be a problem.

Software for SNP Calling

There are many SNP-calling programs out there and quite a few good ones. The standard and most used program is the Genome Analysis Toolkit (http://www.broadinstitute.org/gatk), developed at the Broad Institute. Other popular tools include Samtools (http://samtools.sourceforge.net/) and Platypus (http://www.well.ox.ac.uk/platypus).

How to Check SNP Calls for Quality

There are a number of simple measures that can be used to quickly check a set of SNP calls, to see if it is sensible. Checking individual SNPs can be done either by visual inspection of the relevant part of the BAM file (using IGV (http://www.broadinstitute.org/igv/home) or a similar tool) or by follow-up sequencing with an independent technology, but it is a good idea to check some overall metrics first. Here are a few commonly used quality checks:

Number of SNPs: SNPs occur at the rate of roughly 1 per 1,000 bases in human DNA, so a quick count should tell you if anything is very wrong. A good idea is to check the number of SNPs called per chromosome.

Transition/Transversion Ratio: The rate of transition SNPs (C → T, T → C, A → G, G → A) occurring naturally is known to be significantly higher than the rate of transversion SNPs (everything else). A good benchmark is that the rate should be above 2. Figure 1 shows this rate computed for human chromosomes 1–22.

Allele Frequency: Real SNPs should occur at an allele frequency of either 0.5 or 1.0 in a single sample. Check the distribution of allele frequencies. There should be two large peaks and then a certain amount of noise. If there are not pronounced peaks at 0.5 and 1.0, then something is wrong.

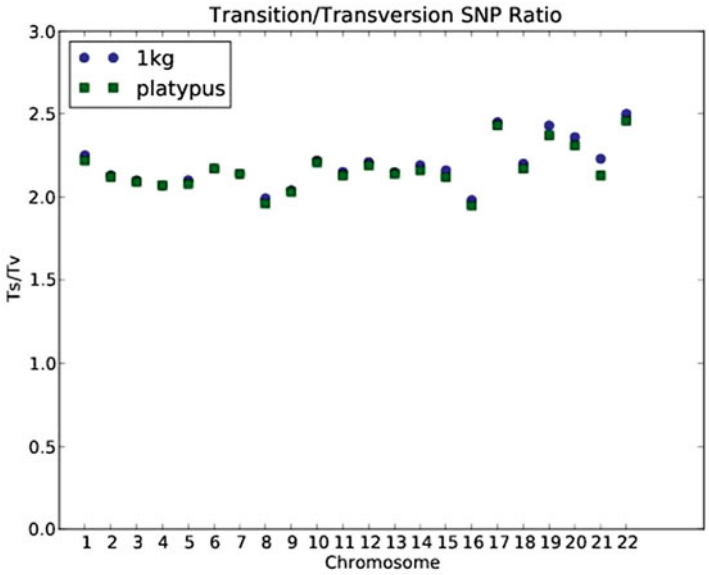

Fig. 1 Transition/transversion rate for SNPs across the human genome, from 1000 Genomes Project Phase 1 data

Indel Calling

Insertion and deletion (Indel) calling has received somewhat less attention than SNP calling, but methods for this are gradually becoming mature. Currently there is still a substantial difference between the outputs of various indel calling programs running on the same data, whereas SNP callers are now giving fairly similar results. The reason is that indel calling is inherently more difficult, requiring in the worst case a reassessment of the alignments and mapping positions of reads. Also, the sequencing error modes which cause insertion/deletion errors during sequencing are less well understood and so less well modelled in the tools. Naive indel calling by simply counting sites where the read mapper has flagged insertions or deletions will result in massive over-calling of indels and is not recommended. Here are a few sources of indel errors:

Sequencing Errors: The sequencing mechanism goes wrong, and sequence is either skipped or sequenced multiple times, resulting in an incorrect nucleotide sequence in the affected read. This is particularly common when sequencing repetitive sequence, e.g. homopolymers or short tandem repeats.

Mapping Errors: The read mapper either puts the read in the wrong place or puts it in the right place and aligns it badly. Either of these results in any real indel being missed and an incorrect indel being seen.

Adapter Contamination: Sequencing the adapter sequences (see earlier) can result in spurious insertions being seen at the end of reads. This is why you should trim the adapter sequences.

Unidentified Large Variants: Not calling a large insertion/deletion/inversion correctly could result in a series of incorrect small indel calls.

Software for Indel Calling

This is less clear-cut than for SNPs. There is not yet a huge weight of evidence to support any one program being the best, but good tools include GATK, Platypus, Dindel and Samtools. Assembler-based approaches such as Cortex are also good choices for calling in particular large indels and more complex variation, although such approaches can be less sensitive than mapping-based approaches particularly for small variants.

How to Check Indels for Quality

This is also less clear than for SNPs, but there are a few useful metrics:

Indel Length: Long indels are less common than short ones, so the length distribution should be heavily peaked around 0. If this is not the case, then something is probably wrong.

Insertion/Deletion Ratio: What little good-quality data there is suggests that, for human genomic DNA, the real insertion/deletion ratio is close to 1.0. Typically the observed ratio is higher than this due to over-calling of fake insertions.

Triplet Enrichment: In exonic/coding sequence, there should be an apparent enrichment of triplet indels (i.e. those lengths which are multiple of 3), as these do not cause frameshifts in the coding sequence.

Calling Other Small Variants

There are no standard protocols for the calling of other small variants. They are reported differently by different variant callers, and there is little data so far on the performance of standard tools in identifying them. Tools like recent versions of GATK and Platypus, which use haplotype calling, should do a reasonable job of calling these.

Calling Large Variants

As mentioned previously, detecting large variants is difficult. There is not much reliable data on how well the various available tools perform, but this is definitely not a solved problem yet. Detecting large variants requires a completely different approach

to detecting small variants, and it is more difficult to represent large variants in VCF, so not all programs output VCF. Tools you might want to try include Cortex (http://cortexassembler.sourceforge.net/index_cortex_var.html) and Pindel (http://bioinformatics.oxfordjournals.org/content/25/21/2865.long).

Different Approaches to Variant Calling

This section will go briefly through the most common approaches to variant calling, used by standard programs such as GATK, Platypus, samtools and Genome STRiP, as well as a number of lesser-used tools, and attempt to summarise the strengths and weaknesses of these approaches.

Mapping Based

The most common method is to map and align sequence data to a reference sequence (see above) and then identify variants by comparing the aligned reads directly to the reference sequence. The accuracy of these approaches depends greatly on the mapping software putting the reads in the right place. These tools can be split into two categories: those that do naïve allele counting based on the mapper alignments and those that perform realignments.

Naive Allele Counting

The simplest methods just look at the aligned reads and count how often a base change or insertion/deletion is observed at a particular location. These methods are normally quite sensitive to SNPs and short indels (conditional on good original mapping/alignment), but must rely on good filtering strategies to reduce the false call rate. Examples of tools that work in this way are Samtools and SYZYGY (http://www.broadinstitute.org/software/syzygy/).

Realignment Based

A more accurate but also more computationally intensive approach is to perform a realignment of reads in certain regions in order to improve the alignments (read mappers typically only see one read at a time, so it is possible to improve alignments by considering all reads within a small genomic window). These newly aligned reads are then used to decide which variants are present. Realigning the reads in this way helps to avoid the problem of having, for example, spurious SNP calls generated by actual insertion/deletion events which were not correctly identified by the read mapper. Programs which follow this approach include GATK, Dindel and Platypus.

Assembly Based

An entirely different approach to variant detection is to avoid read mapping completely and assemble contigs from scratch, using the raw, unmapped reads. This can be done in a way which is entirely reference-free or using information from the reference sequence to scaffold the assembly.

A good example of a program which uses assembly (reference based or reference-free) to identify variants is Cortex (http://cortexassembler.sourceforge.net/). Assembly-based methods can access much larger variants than can be found using standard mapping-based approaches, but may be somewhat less sensitive for small variants.

Local Assembly Based

A more recent approach is to use a combination of read mapping followed by local assembly of small genomic intervals (100–1,000 bp). This approach hopefully gives more power to detect larger variants whilst retaining accuracy and sensitivity for SNPs and short indels. This method is currently implemented by GATK and Platypus.

Split Read Based

Another approach designed to detect larger variants is split-read mapping. This starts from aligned read data and takes read pairs where one read did not map to anywhere on the reference sequence. These unmapped reads are assumed to be reads spanning the break points of large variants (deletions or other structural variants). These are then remapped around the position of their 'mate' using a different algorithm which allows large gaps in the middle of the read. Once this is done, the remapped reads are used to identify structural variant break points. This approach is implemented in PRISM (http://bioinformatics.oxfordjournals.org/content/early/2012/07/31/bioinformatics.bts484.abstract).

Insert Size and Coverage

The last method we discuss involves using information about the insert-size distribution or read pairs and coverage to detect large deletions. These methods look for clusters of read pairs with unusually large insert sizes and use these to infer the presence of large (i.e. much larger than the typical insert size) deletions. This approach is implemented in various tools such as Genome STRiP (http://www.broadinstitute.org/software/genomestrip/genome-strip).

Things to Look Out For

This final section outlines a few common problems and pitfalls encountered in variant calling. None of the methods mentioned above are perfect. Even SNP calls from good-quality data can still have a significant error rate, and the problem is much worse for indels. So here are a few things to watch out for.

Repeats in Reference

The current (37) build of the human genome contains a large amount of repetitive sequence. There are (sometimes quite large) regions which are duplicated in several places. As a result, it is often impossible to map read pairs unambiguously. Most mappers will flag reads that fall in these regions with either a very low or zero mapping quality. Large numbers of reads mapping to the wrong location may cause fake SNP or short indel calls. A good way to avoid this problem is to filter your BAM files and remove reads with low mapping qualities.

The reference sequence also contains a number of collapsed repeats. These are regions which are only represented once in the reference sequence, but are really duplicated many times in real DNA. These can cause problems, as some of the copies may not be exact, and all the reads from all copies will map to the same location in the genome. This can cause large clusters of fake SNPs and indels to be called. A good way to spot this is by looking at the coverage distribution: regions like this will often have many times higher coverage than average (note that this only works for whole-genome data or for exome/targeted when looking across many samples, as it relies on having even coverage across the genome).

Centromere

The centromeres and telomeres are very difficult regions to call variants in, as they tend to be very repetitive. It may be sensible to either skip calling in these regions or simply filter out calls made there.

HLA and MHC

The HLA regions can be very divergent from the reference sequence. Most standard variant callers expect a lower rate of variation than is present in these regions and so do not give accurate calls here. In addition, mapping to the HLA is difficult, and many read mappers will not place reads in the right place here.

Sex Chromosomes and Autosomes

Calling variants on the human sex (X and Y) chromosomes needs some special consideration. Obviously there may be either 1 or 2 copies (or more) of the X chromosome, which will affect coverage and the rate of homozygous calls, so any filters applied to your calls should be modified accordingly.

Large-Scale Abnormalities

It is always worth checking for large-scale abnormalities before looking too closely at things like SNPs. Checking coverage and rates of homozygosity across the target region is always a good idea, as this may show up things like whole-chromosome loss/duplication or deletions of large parts of a chromosome. Rare events such as uniparental isodisomy (two copies of the same chromosome inherited from one parent), which can result in an entire chromosome being homozygous for all variants, are easy to miss if you are not looking at the right scale.

Calling Variants on Different (Non-European) Populations

Some variants are fixed in certain populations. The reference sequence can obviously not represent all populations, so in some cases the reference allele may be the minor allele in the population you are studying.

Allele-Biased Variants

Variants may be well supported by the data but still be present at lower-than-expected allele frequencies. For example, a called SNP may be present in 100 out of 1,000 reads in a single diploid sample. There is always the chance that this is a real, somatic variant, but the most likely explanation is that it is an artefact of some kind. It is worth looking at the estimated frequencies and the read counts. Most variant callers will filter on this automatically, but it is always worth checking.

Strand Bias

Similarly, you should check what fraction of supporting reads are from the forward/reverse strands. The balance should normally be around 50/50, with some spread depending on various things including local sequence context. The forward/reverse

read counts will be reported by most callers, but you can always go back to the original BAM file and check with IGV or a similar tool. Variants which are heavily strand biased are likely to be artefacts. Again, these are normally filtered, but you should check.

Somatic Variants

If you are interested in somatic variants, which may be present at allele frequencies of substantially lower than 50 %, then make sure to check whether or not the variant caller you are using has filters on the allele frequency.

Summary

Variant calling from high-throughput sequence data is a challenging task, requiring a good understanding of the error modes of modern sequencing platforms and robust statistical modelling, as well as fast and memory-efficient software to deal with the huge volumes of data. There are now a variety of good programs to deal with this problem and a variety of approaches to dealing with it. We have seen how to move from raw sequence data to filtered variant calls. This is a fast-changing field, new methods are being actively developed and these methods will need refining as the data changes.

Rare Variant Quality Control

Anubha Mahajan and Neil Robertson

Introduction

The success of any association study depends crucially on the implementation of a rigorous quality control (QC) procedure. Typically, these QC measures involve identification and removal of individuals and variants with high error rates that confound the analysis and results in spurious association signals. Here we provide a protocol detailing the data quality assessment and control steps that are carried out during the QC of a data set, with special reference to rare variants that have propensity for spurious genotypes.

Once genotype data are generated, there are several steps required to process the data into high-quality genotypes for each individual. Sample mix-ups are an ever present menace and need to be guarded against. Mix-ups may occur in the lab or may be the result of erroneous digital manipulation, such as the incorrect sorting of pick-sheets. At basic level the consistency of samples can be checked by comparing estimated sex with sex reported by the phenotypes, where sex is usually present. However, if earlier genotypes are available for a set of samples, then these can help to identify sample mix-ups with far greater efficacy than using sex estimations alone. Some labs may run a panel of SNPs as a set of "DNA fingerprints" to aid traceability. Typically these will be a few dozen common polymorphisms genotyped using the Sequenom platform. Otherwise results of earlier single genotyping or chip genotyping can be used where there is an overlap. Many chips also include a subset of common SNPs; the Illumina Human Exome chip included a set of 274 SNPs for this purpose. Even if there are not enough overlapping SNPs to verify a sample's identity beyond reasonable doubt, it may still assist in isolating negative

A. Mahajan (✉) • N. Robertson
Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine,
University of Oxford, Oxford, UK
e-mail: Anubha@well.ox.ac.uk; n.r.robertson@me.com

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_3

matches. Genotype comparison can best be accomplished by running an identity-by-descent estimation, though interpretation of pairwise results can be involved. Essentially, nominally identical samples will either match, match another sample, or match no one even though the counterpart sample is nominally present. All but the former set need to be excluded. Any issues should be followed up with the originating lab in order to pinpoint earlier errors and correct downstream sample sheets.

Next, QC steps are most often dealt with at two levels: the sample- and the variant-specific level. Ideally, sample level QC should be implemented prior to performing QC on the variant level. This ensures that bad quality DNA samples do not influence variant QC matrices and consequently that the maximum number of variants is carried forward for downstream association analyses.

Sample Quality Control

It is useful to examine per sample quality metrics, including the following steps:

1. Identification of individuals with ambiguous genomic sex or discordant gender information
2. Identification of individuals with high missing genotype rate and/or outlying heterozygosity
3. Identification of duplicated or related individuals
4. Identification of individuals of divergent ancestry
5. Identification of individuals with high genotype error rates
6. Identification of individuals with outlying singleton content
7. Identification of individuals exhibiting batch effects

Ambiguous or Discordant Gender Information

One of the first procedures that should be implemented in any genome-wide association study (GWAS) QC protocol is checking for potential sample swaps and/or contamination, both of which can arise from handling errors. One of the easiest ways to discover such errors is by using genotype data from the X-chromosome. Since males have only one copy of the X-chromosome, they cannot be heterozygous for any marker outside the pseudo-autosomal region of the Y-chromosome. Therefore, mean homozygosity across all X-chromosome variants can be used to infer gender. Typically, homozygosity rates are expected to be 1 for males and less than 0.2 for females. Any discrepancy detected between the genetically determined and the reported gender is often indicative of a sample swap and should be reviewed to identify the source of error. Such samples should be removed from further analysis unless it can be correctly identified, or it can be confirmed that sex was recorded incorrectly.

Gender assignment based on homozygosity rate between 0.2 and 0.8 is ambiguous and is often indicative of sample contamination. Samples with ambiguity in genetically determined gender should therefore be eliminated from downstream analysis.

Sample Genotyping Failure Rate and Heterozygosity

The genotype failure rate and mean heterozygosity per individual are both indicators of DNA sample quality. Genotyping of samples of suboptimal DNA quality and/or low concentration could lead to a high proportion of missing genotypes and aberrant calling. Samples with high genotype failure rate should thus be eliminated from further analysis. A widely recommended threshold is genotype failure rate of 3–5 %. However, this threshold is subjective and may vary from study to study depending on the genotyping platform used (rare variant content) and the quality of the DNA samples genotyped. The distribution of missing genotype rates across all genotyped samples should be inspected to determine the most appropriate threshold, attaining a balance between eliminating minimum samples to achieve maximum genotyping efficiency. It is also good practice to investigate genotype failure rates of individuals on a per-plate basis. If the genotype failure rate of samples for any plate within a study is >10 % (once platform failures have been excluded), the whole plate should be repeated or excluded from subsequent analyses.

The heterozygosity of a sample is the fraction of non-missing genotype calls (autosomal variants only) that are heterozygous. Mean heterozygosity differs between ethnicities and depends heavily on variants genotyped (e.g. due to variability in allele frequencies). As for genotype failure rates, the distribution of mean heterozygosity across all individuals should be reviewed to determine reasonable thresholds at which to exclude the most extreme samples (Fig. 1). It is recommended that mean heterozygosity estimates are generated separately across common and rare variants. Individuals with an excessive or reduced proportion of heterozygote genotypes, which may be indicative of DNA sample contamination or inbreeding, respectively, should be excluded from further analysis.

Sample Relatedness

The next step in running sample QC in GWAS is to look for related individuals in the study. This not only helps in estimating the number of related samples (recorded as well as cryptic relatedness) in the dataset, but is also another way of identifying both potential sample mix-ups and pedigree integrity based on discrepancies between genetic information and self-reported relationships (if accessible). To identify duplicate and related individuals, pairwise kinship estimates, identity by state (IBS), are calculated for individuals in the study, based on average proportion of alleles shared in common at genotyped autosomal variants. Regions of extended linkage disequilibrium (LD), such as the HLA, and highly correlated variants (typically $r^2 > 0.2$) are removed when calculating these estimates. Related individuals share more alleles IBS than expected by chance, with the extent of increased sharing proportional to the degree of relatedness. IBS estimates can further be used to calculate the degree of recent shared ancestry for a pair of individuals (identity by descent, IBD).

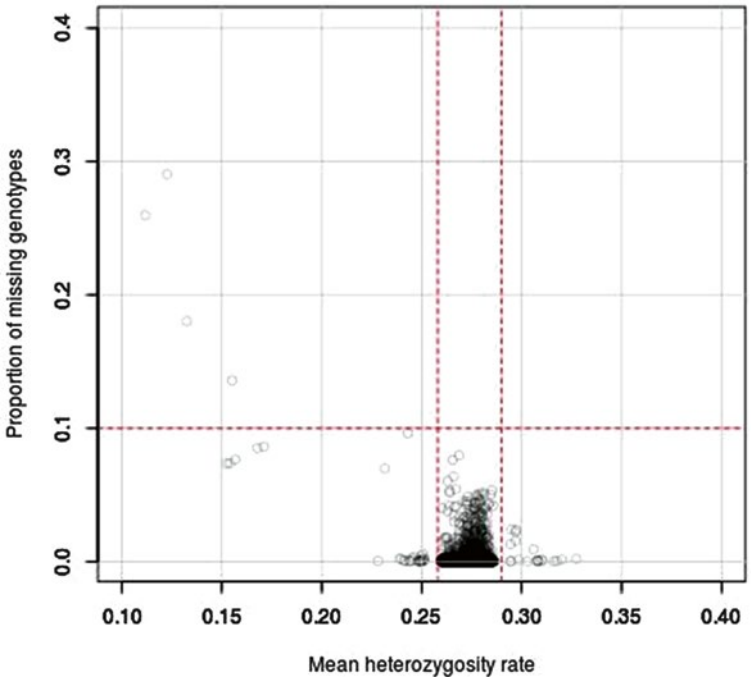

Fig. 1 Genotype failure rates vs. heterozygosity across all individuals the study. Shading indicates sample density, and dashed lines denote QC thresholds

Theoretically, samples that share two alleles IBD at every locus (IBD = 1) are either duplicates or monozygotic twins, whilst IBD = 0.5, 0.25, and 0.125 is expected for first-degree, second-degree, and third-degree relatives, respectively (Fig. 2). Often, genotyping error, LD, and population structure introduce some variation around these theoretical values, and IBD > 0.9 are considered to be indicative of a duplicate sample (or monozygotic twin).

Duplicates and/or related samples introduce bias in population-based association studies, unless they are accounted for in the analysis. For these same reasons, it is customary to remove one individual from each pair with an IBD > 0.1875, which is halfway between the expected IBD for third- and second-degree relatives. For family-based data sets, any discrepancies detected between pedigree records and estimated IBD would be indicative of non-paternity, adoption, sample mix-up, or duplicate processing of the same individual and thus should be investigated further to attempt to identify the problem. Kinship estimates are also very informative in detecting identical or related individuals recruited from multiple centres, in studies where datasets from different collection sites are combined.

Each study should also deliberately include duplicate pairs on plates to determine genotyping error rate. Despite great improvements in genotyping technologies, rare variant genotyping is still difficult, and these checks are very important in establishing the quality of rare variant calls.

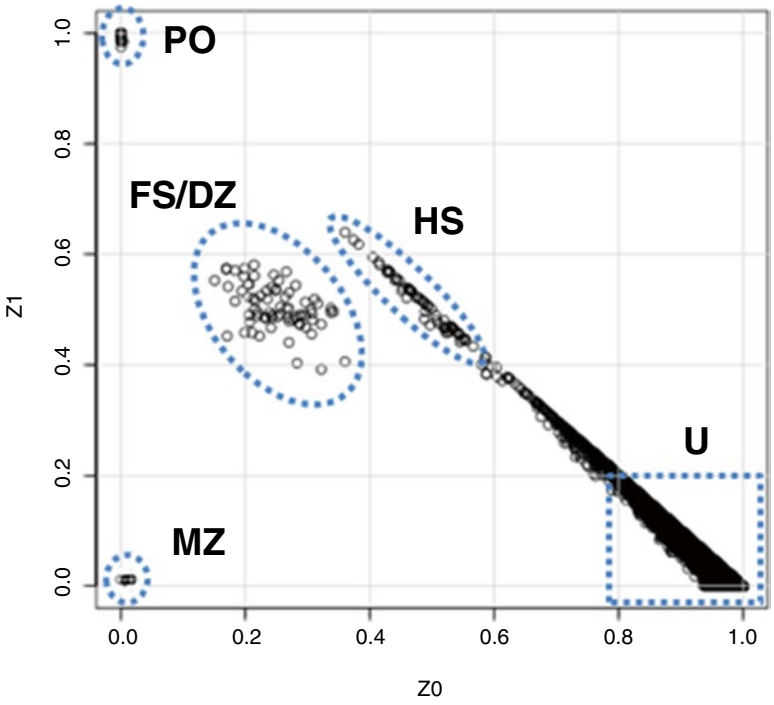

Fig. 2 Relatedness inference from pairwise IBD coefficients, Z_0 and Z_1. Each point represents a pair of samples, and the diagonal line is $Z_0+Z_1=1$. The *dotted lines* in the plot highlight the estimated relatedness: *PO* parent–offspring ($Z_1=1$), *FS/DZ* full sibling/dizygotic twin ($Z_0=0.25$, $Z_1=0.5$), *HS* half-sibling-like ($Z_1=0.5$, $Z_0=1-Z_1$), *MZ* monozygotic twins/duplicates ($Z_0=Z_1=0$), and *U* unrelated

Population Stratification

Population stratification arising due to systematic differences in genetic ancestry of individuals included in a study can be a major source of bias in population-based association studies and can lead to false-positive signals. Evolutionarily, since rare variants occurred in recent human history, they are expected to be population specific and show greater population diversity than common variants (Li and Leal 2008). It is therefore very important to ensure that samples included in inferring rare variant associations are drawn from a relatively homogenous population. Often, even after giving careful consideration to matching of cases and controls on population origin, some level of stratification may still be detected in samples. Efforts should be made to remove or reduce the effect of population stratification by removing individuals of divergent ancestry from downstream analyses. Several statistical methodologies have been developed to detect and adjust for population structure in association analysis. A more detailed description of population structure analysis and the adjustments required for robust rare variant association analysis are provided in Chap. 19.

The frequency differences of variants between individuals from different ancestries may introduce bias in association studies and may also introduce apparent deviations from Hardy–Weinberg equilibrium (HWE; see below). Therefore, individuals from diverse ancestry should be ideally removed, or at least treated separately, whilst performing variant QC.

Genotype Error Rates

Variant genotyping concordance should be checked on duplicate samples to confirm the robustness of genotyping across sites and/or different platforms, wherever applicable. Duplicate samples with high genotyping error rates are indicative of bad DNA sample quality, and consequently both should be excluded from further analysis. Samples should also be tested for concordance with previous genotype data (wherever available), and samples with high degree of discordance should be excluded.

Singleton Content

Singletons are variants found in only one of the genotyped samples. Singleton status is known to be affected by various factors genome-wide, such as recombination, selection, and mutation. In genotyped samples, the number of singletons observed per sample would depend on the genotyping array content, especially the rare variant content, and the number of samples genotyped. Random errors in genotyping of rare variants are likely to end up as singletons. Consequently, an elevated number of singletons per sample could be indicative of genotyping errors. The distribution of singletons across all individuals should be reviewed to determine outlying samples that should be excluded from downstream analysis.

Batch Effects

As a general practice, samples are partitioned into small batches for processing during genotyping. Ideally, these batches should contain random sets of samples with respect to sex, ethnicity, and other potential confounders. However, despite all the care, systematic differences in the composition of samples in a batch, and the per-plate genotyping error and efficiency, can result in batch effects and could, in effect, lead to detection of false-positive associations. Examining the average minor allele frequency and genotyping failure rate across all variants for each plate and batch is often a good measure to determine batch effects. Plates or batches with significantly different estimates should be further investigated for genotyping or composition problems and, if unresolved, should be removed from further analysis.

Variant Quality Control

The success of an association study depends crucially on using a high-quality set of variants. Suboptimal variants not only introduce false positives, but also reduce the ability to identify true associations with traits. The criteria for filtering out low-quality variants are study specific, and the utmost care should be taken when defining these thresholds since every removed variant is a potentially missed causal variant. Generally, two variant QC thresholds can be used: a more stringent threshold at which variants are removed from the analysis and a second liberal threshold for which variants are flagged and re-examined later for potential QC-related bias. The calling of rare genotypes is subject to higher error rates, and therefore they are typically removed using increased levels of stringency with decreasing frequency thresholds.

Variant QC consists of the following steps:

1. Identification of variants with high missing genotype rate
2. Identification of variants demonstrating significant deviation from HWE
3. Identification of variants with significantly different missing genotype rates between cases and controls (if you have cases and controls in your data set)
4. Identification of variants that have significantly different allele frequency distributions between sample batches/sub-cohorts
5. Identification of variants with bad cluster plots
6. Identification of variants for which allele frequency differs significantly from that reported in the 1000 Genomes Project (2012)

Genotyping Rate

The proportion of samples with a genotype call for each variant is a good indicator of its quality. Variants that fail genotyping on a large proportion of samples are poor assays and consequently might result in spurious associations. A recommended threshold for excluding variants with high genotype failure rates is 5 % for common variants and a more stringent threshold of 1 % for rare variants. However, as mentioned in sample QC, these thresholds may vary from study to study, and the distribution of genotype failure rates should be reviewed before implementing any thresholds (Fig. 3).

Deviation from HWE

Testing for HWE is commonly used for quality control of genotyping because departure from equilibrium is considered to be an indicative of potential genotyping errors or population stratification. However, deviations from HWE may also indicate

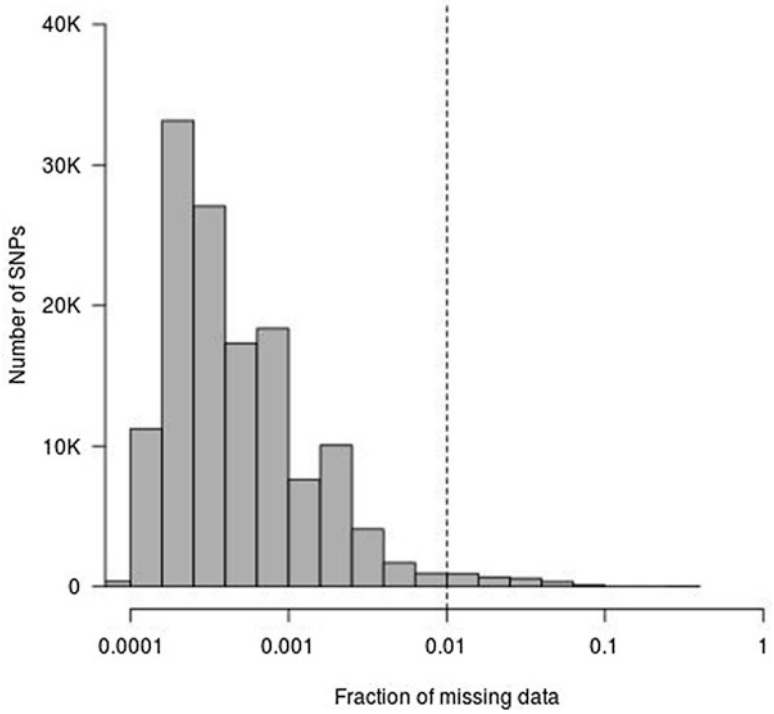

Fig. 3 Histogram of missing genotype rate across all QC-passed individuals. The *dashed lines* represent threshold at which SNPs were removed from further analysis

selection and therefore may in fact reflect a true association to a disease (Lee 2003; Nielsen et al. 1998). Removing these variants from further investigation would be counterproductive; therefore it is common practice to not omit them from the analysis but to flag for close scrutiny post association analysis. It is also, therefore, beneficial to use only control samples when testing for deviation for HWE. If the study sample has been drawn from diverse ethnicities, it is necessary to test for HWE within each ancestry group separately. Typically, deviations from HWE towards an excess of heterozygotes reflect technical errors on assay, such as nonspecific amplification of the target region. Examination of genotype cluster plots for variants can be useful in screening for technical origins of HWE deviations. Like other QC parameter thresholds, the significance thresholds for declaring variants to be out of HWE vary greatly between studies (p-value thresholds between 0.001 and 5.7×10^{-7} using exact test). It is recommended that when lower thresholds are used, genotype cluster plots should be examined manually for quality for variant showing any evidence of deviation.

Differential Missing Genotype Rates

Significant differences in missing genotype rate between cases and controls can be another confounding factor in association analysis. Calling cases and controls together greatly reduces this confounding, but significant differences in genotyping failures may still exist and lead to spurious associations. Variants should be tested for differential missing genotype rate and excluded from subsequent analysis. A recommended threshold is p-value less than 1×10^{-4} and should be defined on study-wise basis.

Allele Frequency Distribution

In studies where samples are collected from multiple sites, it is recommended that tests for significant differences in call rate, allele frequency, and genotype frequency between sites are conducted to ensure homogeneity of samples. This check is also very important when cases and /or controls are drawn from diverse sources.

Cluster Plots

Large number of variants being tested in a GWAS precludes looking at intensity plots of all of the variants. However, statistical measures of separation between the three genotype clusters, such as those provided by Illumina's clustering algorithm (cluster separation score and GenTrain score; varying from 0 to 1), provide an easy way to detect poorly clustering variants Typically, a threshold of 0.4 and 0.6 for cluster separation and GenTrain score, respectively, is recommended for detecting poor clusters. However, even after following very stringent individual and variant QC, genotyping errors may persist. Inspecting cluster plots manually of at least any significant variants is the best way of ensuring the robustness of genotype calls and association. Good quality variants should have three clearly defined tight clusters, with the homozygotes falling along the vertical or horizontal axis and the heterozygotes at a 45° angle (Fig. 4). Poorly performing variants often present overlapping clusters from one allele and from the heterozygote or more than three clusters or a split cluster.

Concordance with 1000 Genomes Project Data

After all QC steps, it is recommended that allele frequency checks be performed between the genotyped data and the relevant 1000 Genomes Project population. Variants for which allele frequency differs significantly from this reference should be omitted from further analysis.

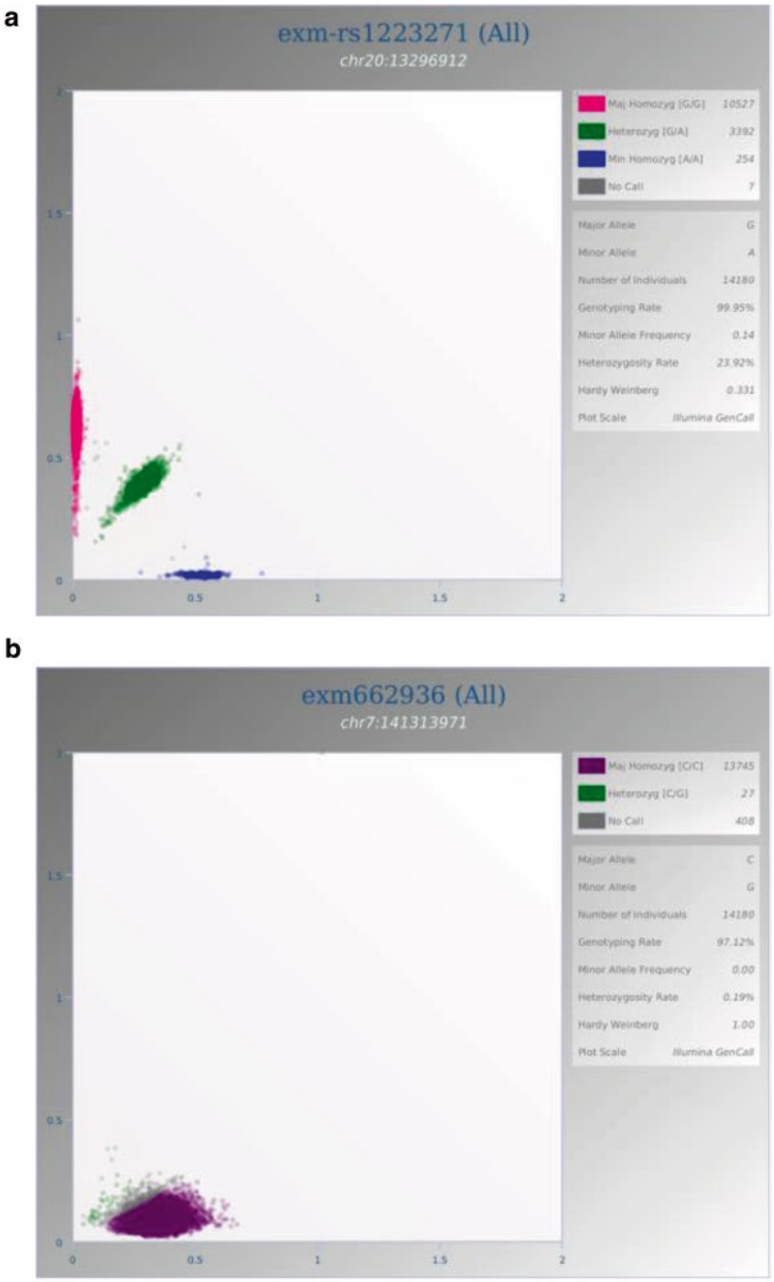

Fig. 4 Examples of genotype cluster plots. Cluster plots for two SNPs from genotype array. Each spot corresponds to one individual's genotype. Individuals with major and minor homozygous genotypes are *pink/purple* and *blue*, respectively. Heterozygous individuals are shown in *green*; individuals with missing genotypes are *grey*. (**a**) An SNP with a good cluster plot; (**b**) represents a badly separated cluster plot where samples have been erroneously classified as *purple* and *green*

X-Chromosome Quality Control

Quality control for X-chromosome variants should be performed by following the same pipeline as for autosomal variants. However, genotype failure rate, HWE, allele frequency, and concordance checks should typically be based on female samples only.

Post Hoc Confirmation of Quality Control

As a final step to confirm the quality of the QC process, it is advisable to examine the square of the GWAS test statistics for any correlation with residual genotype call rate, HWE, and minor allele frequency of the surviving variants.

The sample size of the heterozygote and rare homozygote clusters makes rare variants difficult to call, which thus frequently present as false positives in case–control association tests. It is valuable to re-genotype any significantly associated rare variants using a different technique to insure robustness of any finding. Furthermore, even when well called, associations at these rare variants are less robust because they are driven by the genotypes of only a few individuals.

Unfortunately, even with the most stringent QC, poor quality rare variants slip through the net and generate false-positive association signals, either in single variant or gene-based tests; however, these can easily be recognised through careful inspection of cluster plots when using genotyped data. Furthermore, there are continued developments in efficient and accurate algorithms for calling rare variants, and these show promise for reduced error rates (refer to Chap. 3 for more details).

References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Lee WC (2003) Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. Am J Epidemiol 158:397–400

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy Weinberg disequilibrium at a marker locus. Am J Hum Genet 63:1531–1540

Rare Structural Variants

Menachem Fromer and Shaun Purcell

In this chapter, we will briefly touch on the historical discoveries of large abnormalities in the structure of the human genome. It is now clear that more subtle structural variants are in fact ubiquitous and key to understanding the spectrum of risk for many human diseases. While many of these changes are individually rare, the aggregate burden in the population is significant. With this in mind, we give an overview of the technologies developed to assay these variants in a high-throughput manner at ever-increasing granularity, including array-based platforms and next-generation sequencing. We then focus on whole-exome sequencing, since many disease studies to date have adopted this approach. Throughout, we review some of computer software and algorithms available for extracting structural variant information from experimental data. We conclude with a comparison of the strengths and weaknesses of the various current technologies and provide a small sampling of emerging methods for investigating the range of structural variation in more detail.

M. Fromer (✉) • S. Purcell
Division of Psychiatric Genomics and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
e-mail: fromer@broadinstitute.org; shaun@atgu.mgh.harvard.edu

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_4

Structural Variation

Historically, geneticists have detected large changes in the structure of chromosomes using a microscope or cytogenetic staining techniques to perform a karyotype analysis. In doing so, they detected marked and unusual changes in DNA quantity and structure, many of which were related to disease, including aneuploidies, euchromatic variants, and rearrangements. Our growing knowledge of structural changes to the genome is reviewed in a number of recent papers (Sharp et al. 2006; Feuk et al. 2006). A well-known example of such a change is the association of an additional copy of chromosome 21 with Down's syndrome (Carter 2007). However, over the past two decades, researchers first deciphered more detailed global features and then the exact sequence of the human genome. During that time, the pervasiveness of even finer-scale genomic structural rearrangements has become apparent, with as much as 5 % of the genome varying in structure across individuals in the population (Sharp et al. 2006; Perry et al. 2008; Zhang et al. 2009). These structural changes, most of which are submicroscopic, include novel sequence insertions, duplications, deletions, inversions, repeated sequence motifs, and translocations. Together, deletions and duplications form a subclass of structural variation that is known as copy number variants (CNVs).

There are numerous working hypotheses regarding the many ways in which structural variation (SV) in the genome evolves and diversifies within the population. For example, the existence of segmental duplication regions (low copy repeats) clustered throughout the human genome has implicated nonallelic homologous recombination (NAHR) as playing a mechanistic role in the generation of new structural variants in the population (Sharp et al. 2006). Specifically, NAHR can result in deletion, duplication, or inversion of nearby genomic sequence. Moreover, it has been suggested that preexisting sequence inversions can reduce the frequency of recombination during meiosis, resulting in greater rates of chromosomal rearrangements (Feuk et al. 2006).

Large structural variants (thousands of DNA bases or more) have long been known to be associated with certain Mendelian forms of disease, spanning a wide range of phenotypes, including hemophilia (inversion), α-thalassemia (deletion), and glucocorticoid-remediable aldosteronism (duplication) (Lupski 1998). Moreover, as technologies have expanded to enable discovery of novel structural variants (see below), many structural variants have been associated with phenotypic variation including increased risk for various diseases (Sharp et al. 2006; Feuk et al. 2006). These variants may mediate disease risk through changes to gene expression levels (directly via dosage effects for deletions and duplications or indirectly through the effect of the variant on transcriptional mechanisms), disruption of the coding sequence of a gene, the fusion of two genes into a new "gain-of-function" gene, or even predisposing to further deleterious structural changes during DNA replication (Feuk et al. 2006; Zhang et al. 2009). For a recent review of how structural variation can impact phenotype at the molecular and cellular levels, see (Weischenfeldt et al. 2013).

By necessity, structural variants that are common in the population cannot play a large role in the etiology of a rare disease. On the other hand, as numerous studies have shown that common variation tagged by single-nucleotide polymorphisms (SNPs) can definitively play a role in the risk for common diseases (Burton et al. 2007), the role of common structural variants has been debated. To address this, a recent large-scale study of common CNV was conducted (Craddock et al. 2010). It was found that CNVs that can be genotyped on existing array-based experimental platforms (see next section) are "unlikely to contribute greatly to the genetic basis of common human diseases" or have already been tested for association via common SNPs that tag them. Thus, the focus of much disease research has been in the realm of rare structural variants, which collectively could account for disease in a few percent of individuals, including for both more common diseases (International Schizophrenia Consortium 2008; Shlien and Malkin 2010) and rare conditions such as structural birth defects (Southard et al. 2012). We follow that theme in this chapter and focus mainly on rare structural variation.

Arrays

As many large-scale single-nucleotide genotyping efforts were undertaken for the purpose of performing genome-wide association with disease, high-throughput studies of structural variants have often "piggybacked" these, to utilize the same technologies for a different, but additional, purpose (Zhang et al. 2009). Specifically, SNP genotyping arrays were constructed to assess the diploid genotypes of individuals at loci of common single-base variation. However, by considering the intensity of the signals at these sites across genomic regions (Conrad et al. 2006), a number of algorithms were able to convert relative shifts in these intensities for an individual into CNV information, thus providing both SNP and CNV data from a single experimental platform (Korn et al. 2008; Wang et al. 2007). Similarly, it has recently been demonstrated that "exome arrays" intended for genotyping rare protein-altering sequence variation can still also be utilized to detect large CNVs (400 kilobases or longer) that overlap protein-coding genes (Szatkiewicz et al. 2013).

Two commonly applied methods for calling CNV from genotype array data are Birdsuite and PennCNV (Wang et al. 2007), both of which use a hidden Markov model (HMM) to smooth out noisy patterns across genomic regions. Many other methods exist, including circular binary segmentation (CBS), wavelets, expectation maximization, and clustering, and these have been extensively compared and tested on benchmark data sets over the past few years (Karimpour-Fard et al. 2010). A critical finding of such studies was that different bioinformatic tools applied to the same raw data can yield CNV calls with less than even 50 % concordance (Pinto et al. 2011), emphasizing the need for appropriate filtration of CNV calls (particularly in complex genomic regions) for both research and clinical applications.

A second widely used array-based method of CNV identification that predates the use of SNP arrays is array comparative genomic hybridization (aCGH), also known as chromosomal microarray analysis (CMA). In this approach, test and reference DNA are each labeled using a different fluorophore and then hybridized to the same array of "tiled" (overlapping) genomic probes, and fluorescence comparisons provide relative copy number levels across a genomic region (Carter 2007; Heidenblad et al. 2008). The advantage of this technique is that it suffers considerably less from experimental noise than genotyping arrays since the simultaneous hybridization of test and reference DNA samples should automatically account for batch effects and variability between arrays. In addition, the genomic resolution of detected CNV is typically a few kilobases for aCGH (Heidenblad et al. 2008) vs. tens to hundreds of kilobases for genotyping arrays, which also often allows aCGH sufficient resolution for the mapping of structural breakpoints. However, a key disadvantage is that, in research settings, this method still requires an extra array to be run in the lab that is tailored only for finding copy number variation. Nonetheless, aCGH has fast become the de facto standard and initial clinical diagnostic tool for tumor cytogenetics and for prenatal and postnatal screening of babies with anomalous defects, augmenting or even replacing traditional cytogenetic techniques (Shinawi and Cheung 2008).

A major disadvantage of almost all array-based approaches for detection of chromosomal structural variation is that they are typically blind to genomic rearrangements that have no effect on copy number, such as balanced translocations and inversions. With the advent of next-generation sequencing, researchers have started to systematically explore these and other variations. We now give an overview of these technologies in the context of structural variation detection.

Next-Generation Sequencing

For the first time, the advent of next-generation sequencing (NGS) permits the simultaneous assessment of many forms of genetic variation present in individuals, including single-nucleotide variants, small insertions and deletions (indels), as well as larger structural variants. While the identification of single nucleotide and small indels is not always straightforward and requires careful consideration (DePristo et al. 2011; Li et al. 2009), there are even more hurdles to using the same sequencing data for the detection of structural variation, resulting from, e.g., sequencing efficiency, formation of chimeric DNA fragments, and more.

Many methods have been devised to extract structural variation from NGS data, and each has strengths and weaknesses in trying to deal with very large amounts of data with a considerable degree of noise. Some of the well-known tools include PEMer (Korbel et al. 2009), VariationHunter (Hormozdiari et al. 2009), Pindel (Ye et al. 2009), SVDetect (Zeitouni et al. 2010), CNVer (Medvedev et al. 2010), Genome STRiP (Handsaker et al. 2011), AGE (Abyzov and Gerstein 2011),

CNVnator (Abyzov et al. 2011), BIC-seq (Xi et al. 2011), GASVPr (Sindi et al. 2012), and cnvHiTSeq (Bellos et al. 2012); for a recent review comparing and contrasting these and other methods, see (Xi et al. 2012). Typically, these tools use one or more of the following sequencing features to infer the existence of structural variation and the genotypes for particular individuals: a) larger (or smaller) than expected distances between mapped read pairs (for paired-end sequencing), b) reads split in half by the breakpoint of a structural variant, c) aberrant read depth reflective of copy number changes, d) correlation in a population sample of the presence or absence of nearby SNPs (linkage disequilibrium, LD), and e) sharing of structural variants, and hence sequence features indicative of these variants, among individuals bearing more common (though still rare) variation; see, e.g., Handsaker et al. (2011). The main strength of these NGS-based approaches is their potential for mapping structural variation at nucleotide-level resolution, in addition to being sensitive to detecting smaller structural changes (on the order of hundreds or thousands of base pairs) than those that can be reliably called from the microarray-based approaches described above (on the order of 10,000–100,000 base pairs). Moreover, many of these tools can detect deletions and duplications (CNV), as well as other structural variation, including novel sequence insertions, inversions, and translocations.

Using a subset of these tools, the 1000 Genomes Project Structural Variation group has thus far constructed a fine-scale map of CNV including deletions, insertions, and tandem duplications (Mills et al. 2011), with over half of these mapped at base-pair resolution. As expected, complete or partial deletions of genes were observed at lower than expected frequencies in the population, indicative that significant negative selection is acting against such variation and suggesting the potential for future disease research. The researchers were also able to localize genomic "hot spots" for the formation of new structural variation, consistent with the hypothesis that genomic architecture can contribute to structural instability and thus have implications for disease risk. In this large-scale analysis, it was found that Genome STRiP, which integrated all of the multiple categories of evidence listed above in order to detect structural variation, performed the best in terms of both accuracy and sensitivity. Furthermore, by separating the "discovery" stage from the "genotyping" stage, Genome STRiP was well positioned to integrate the structural variation found by all other algorithms tested and genotype those events in all individuals to produce the final list of refined structural genotypes.

While the tools described here were almost all exclusively developed for the identification of structural variation from whole-genome sequencing data, there are additional biases and data issues that need to be accounted for when dealing with sequencing that targets only a subset of the entire genome. A prototypical example of this is whole-exome sequencing, which has recently become popular for the study of disease. We now turn to this class of data and discuss how it critically differs from whole-genome data, as well as the approaches tailored to call structural variation from such exome data.

Whole-Exome Sequencing in Disease Studies

In whole-genome shotgun sequencing, genomic DNA is sheared to create random fragments, and these fragments are then sequenced and mapped back to the reference genome and used to call SNV and indels, as well as structural variation (as outlined above). While many biases exist (including GC content biases), the overwhelming majority of the genome is, in principle, accessible for interrogation of structure using these methods. On the other hand, in targeted sequencing, particular genomic fractions are enriched (by array-based hybridization, or "capture"), and only these subsets of DNA fragments are sequenced. Prominent among this class of targeted experiments is whole-exome sequencing, which aims to sequence the approximately 1 % of the genome that codes for protein sequence. The advantage of whole-exome sequencing is the decreased cost for even very high coverage, as well as the greater interpretability of the effect that the discovered genetic variation will have on protein-coding regions. Along with these factors, the expectation from Mendelian disease that rare coding variation plays a major role in human disease has propelled exome sequencing to the forefront as a state-of-the-art approach in genetic studies of human disease (Teer and Mullikin 2010).

Various methods have been formulated to specifically address the needs of detection of structural variation (mostly CNV) from exome sequencing. Specifically, these approaches must account for the fact that the sequencing reads from targeted sequencing are both nonuniformly clustered in the genome and cover only a small fraction of the genome. This implies that the chances that the breakpoints of any particular structural variant lay within a read are small. In addition, the distance between mate pairs generated by paired-end sequencing cannot easily be used to infer the presence of a structural variant as the pairs will be biased, for example, toward those derived from fragments with both ends in the exome (Fromer et al. 2012). Thus, most (but not all) methods for detecting CNV from exome data rely almost exclusively on sequencing read depth information. In addition, any such method must cope with the additional genomic biases introduced by the exomic capture step, which often exacerbates sample- and target-specific patterns of read depth depletion and enrichment resulting from PCR amplification, sequencing efficiency, and experimental conditions, all of which act together to obfuscate the quantitative relationship between underlying copy number and the observed sequencing depth (Fromer et al. 2012).

The advantage of detecting CNV specifically from exome data is that all variants can be assigned to one or more genes, allowing researchers to develop mechanistic and causal hypotheses of how disease-associated CNV can actually bring about that disease. Moreover, as for structural variation derived from whole-genome data, whole-exome data can yield CNV data at finer scale than the resolution provided by array-based approaches (tens of thousands of base pairs), with exome-based CNVs approaching the level of single exons or even possibly base-pair resolution.

The methods developed for detecting CNV from whole-exome sequencing data include CONDEX (Ramachandran et al. 2011), SeqGene (Deng 2011), ExomeCNV

(Sathirapongsasuti et al. 2011), the approach in Lonigro et al. (2011), exomeCopy (Love et al. 2011), VarScan 2 (Koboldt et al. 2012), CONTRA (Li et al. 2012), CoNIFER (Krumm et al. 2012), ExoCNVTest (Coin et al. 2012), XHMM (Fromer et al. 2012), ExomeDepth (Plagnol et al. 2012), the technique in Wu et al. (2012), exome2cnv (Valdés-Mas et al. 2012), CoNVEX (Amarasinghe et al. 2013), and FishingCNV (Shi and Majewski 2013). The algorithms presented in Nord et al. (2011), as well as the Splitread algorithm (Karakoc et al. 2012), also detect CNV breakpoints at base-pair resolution when they are present in the targeted coding regions.

An important distinction between the above methods is that some utilize "comparative" data normalization based on the relative values of read depth between one sample and another sample(s), particularly those geared for cancer applications, where samples can naturally be grouped as pairs of somatic tumor DNA and matched normal DNA from the same tissue (Sathirapongsasuti et al. 2011; Lonigro et al. 2011; Koboldt et al. 2012; Valdés-Mas et al. 2012; Amarasinghe et al. 2013). On the other hand, some of the methods perform more upfront and general data-driven normalization for all samples simultaneously, using either principal component analysis (Fromer et al. 2012; Krumm et al. 2012; Coin et al. 2012; Shi and Majewski 2013) or with explicit modeling of potential biases (Li et al. 2012; Plagnol et al. 2012; Wu et al. 2012). For calling the CNV, a typical approach is to use hidden Markov models (HMM) or segmentation algorithms, both of which have been adapted from the array-based CNV calling procedures described above.

Thus far, some of the above and similar approaches have been applied in studies of disease, including in tracking the genomic evolution of metastatic cancers in response to therapy (Murtaza et al. 2013), for association with autism (Krumm et al. 2012; Lim et al. 2013), and in the search for Mendelian mutations in a family with nonsyndromic hearing loss (Park et al. 2013). As larger cohorts of cases and controls, and family structures, are exome sequenced, the ability to detect CNV and other structural variants from the resulting exome sequencing data holds much promise for gaining insight into disease. Specifically, as with structural variation inferred from whole-genome sequencing, the unprecedented fine-scale resolution of CNV inferred from exome sequencing can allow researchers to see previously unobservable phenomena, finding both better breakpoints and ever-smaller CNV. One such example is the potential to highlight subclasses of CNV with potentially higher functional impact, e.g., those that delete or duplicate only parts of a whole gene, as demonstrated using the XHMM method (Fromer et al. 2012). The XHMM algorithm (Fromer et al. 2012) has also been used to find de novo CNV (those arising in children but not found in their parents) by filtering out the many spurious instances of poorly called CNV in either the child or the parents. The latter two results are based on the extensive range of quality scores associated with each CNV call that is output by XHMM, allowing the user to query the dataset for instances of high certainty regarding the existence of a CNV, the lack of a CNV in another individual, and the points where a CNV starts or stops.

Conclusions and Future Directions

In this chapter, we have discussed how structural variation has moved from being thought of as extremely rare and responsible for only very particular genetic diseases to in fact being detectable at much wider and more refined scales (in terms of variety and size). Collectively, structural variation makes up a nontrivial portion of the genetic differences between individuals in the population and their risk for both rare and complex genetic diseases and other phenotypes. As newer technologies have allowed scientists to access more varied forms of structural changes in human genomes, each new discovery has demonstrated the complexity of the genome. This complexity, both a result of evolution and sometimes a driver of evolution, also has the potential to cause or increase risk for disease when it gets modified sufficiently to adversely affect cellular functioning.

In this chapter, we have discussed the technologies that currently exist for assessing structural variation, as well as the computational algorithms applied to each data type. The experimental technologies are summarized and compared in Table 1, which is intended to be a rough guide as how one may choose to proceed for a particular research project.

Some newer developments that will provide even richer data for analysis and detecting potential association with disease and other complex traits include the ability to detect more complex structural rearrangements by devising clever and specific techniques based on next-generation sequencing technologies (Talkowski et al. 2011; Sobreira et al. 2011). Preliminary reports have already demonstrated the true power of these technologies and the seeming ever-expanding complexities potentially present in the human genome, e.g., in the clinical diagnosis of a prenatal sample (Talkowski et al. 2012). Also, the ability to detect structural variation in

Table 1 Comparison of technologies for detecting rare structural variation

Category	Technology	Genomic scope	Structural variation classes detected	Minimum size of variation reliably detected (kb)	Minimum number of samples
Microarray	SNP array	Genome	CNV	~50	1
	Exome chip	Exome	CNV	~400	1
	aCGH	Genome	CNV	~1	1
Next-generation sequencing	Whole-genome sequencing	Genome	CNV, inversions, translocations, small and large indels	Base-pair resolution	1
	Whole-exome sequencing	Exome	CNV, small and large indels	Exon-level resolution	20–50
PCR		Targeted	CNV	Base-pair resolution[a]	1

[a]A hypothesized CNV can be confirmed at base-pair resolution using PCR. But if the breakpoints are incorrect, then other technologies will typically be required

single cells can usher in a new era of understanding with respect to the somatic genetic variability of human cells (Konings et al. 2012). Another challenge in the accurate detection of structural variation is that it becomes quite difficult to estimate the exact copy number of multi-copy CNV in a particular individual (Aten et al. 2008). Existing technologies have been adapted and new ones developed for this purpose, including quantitative PCR (qPCR), droplet digital PCR (Whale et al. 2012), and NanoString's nCounter technology (Geiss et al. 2008), originally devised for gene expression measurement.

As with most classes of large-scale genetic association study, those based on rare structural variants require thoughtful deliberation regarding possible sources and causes of artifact, particularly those resulting in false-positive variant calls. This is all the more true as next-generation sequencing data has the potential to provide a comprehensive look at a wide spectrum of structural variation present in each individual, with the caveat that the large and growing quantities of data will also contain noise, batch effects, and other experimental and bioinformatic artifacts that may disguise themselves as "interesting" signal. Nonetheless, with the development and benchmarking of many tools tailored to specifically address these issues, geneticists can now be empowered to judiciously apply these tools to both large and small whole-genome or whole-exome sequencing studies of disease and other complex traits in order to gain new insights into disease etiology and mechanisms.

References

Abyzov A, Gerstein M (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics 27(5): 595–603. doi:10.1093/bioinformatics/btq713

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21(6):974–984. doi:10.1101/gr.114876.110

Amarasinghe KC, Li J, Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. BMC Bioinformatics 14(Suppl 2):S2. doi:10.1186/1471-2105-14-S2-S2

Aten E, White SJ, Kalf ME, Vossen RHAM, Thygesen HH, Ruivenkamp CA, Kriek M, Breuning MHB, den Dunnen JT (2008) Methods to detect CNVs in the human genome. Cytogenet Genome Res 123(1-4):313–321. doi:10.1159/000184723

Bellos E, Johnson MR, Coin LJ (2012) CnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. Genome Biol 13(12):R120. doi:10.1186/gb-2012-13-12-r120

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661–678. doi:10.1038/nature05911

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21. doi:10.1038/ng2028

Coin LJM, Cao D, Ren J, Zuo X, Sun L, Yang S, Zhang X et al (2012) An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. Bioinformatics 28(18):i370–i374. doi:10.1093/bioinformatics/bts379

Conrad DF, Daniel Andrews T, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38(1):75–81. doi:10.1038/ng1697

Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D et al (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464(7289):713–720. doi:10.1038/nature08979

Deng X (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. BMC Bioinformatics 12(1):267. doi:10.1186/1471-2105-12-267

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498. doi:10.1038/ng.806

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7(2):85–97. doi:10.1038/nrg1767

Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM, Handsaker RE et al (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am J Hum Genet 91(4):597–607. doi:10.1016/j.ajhg.2012.08.005

Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Perry Fell H et al (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26(3):317–325. doi:10.1038/nbt1385

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43(3):269–276. doi:10.1038/ng.768

Heidenblad M, Lindgren D, Jonson T, Liedberg F, Veerla S, Chebil G, Gudjonsson S, Borg A, Mansson W, Hoglund M (2008) Tiling resolution array CGH and high density expression profiling of urothelial carcinomas delineate genomic amplicons and candidate target genes specific for advanced tumors. BMC Med Genomics 1:3. doi:10.1186/1755-8794-1-3

Hormozdiari F, Alkan C, Eichler EE, Cenk Sahinalp S (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19(7):1270–1278. doi:10.1101/gr.088633.108

International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237–241. doi:10.1038/nature07239

Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE (2012) Detection of structural variants and indels within exome data. Nat Methods 9(2):176–178. doi:10.1038/nmeth.1810

Karimpour-Fard A, Dumas L, Phang T, Sikela JM, Hunter LE (2010) A survey of analysis software for array-comparative genomic hybridisation studies to detect copy number variation. Hum Genomics 4(6):421–427. doi:10.1186/1479-7364-4-6-421

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22(3):568–576. doi:10.1101/gr.129684.111

Konings P, Vanneste E, Jackmaert S, Ampe M, Verbeke G, Moreau Y, Vermeesch JR, Voet T (2012) Microarray analysis of copy number variation in single cells. Nat Protoc 7(2):281–310. doi:10.1038/nprot.2011.426

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10(2):R23. doi:10.1186/gb-2009-10-2-r23

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E et al (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40(10):1253–1260. doi:10.1038/ng.237

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, Quinlan AR, Nickerson DA, Eichler EE (2012) Copy number variation detection and genotyping from exome sequence data. Genome Res 22(8):1525–1532. doi:10.1101/gr.138115.112

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. doi:10.1093/bioinformatics/btp352

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL (2012) CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28(10):1307–1313. doi:10.1093/bioinformatics/bts146

Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, Neale BM et al (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77(2):235–242. doi:10.1016/j.neuron.2012.12.029

Lonigro RJ, Grasso CS, Robinson DR, Jing X, Wu YM, Cao X, Quist MJ, Tomlins SA, Pienta KJ, Chinnaiyan AM (2011) Detection of somatic copy number alterations in cancer using targeted exome capture sequencing. Neoplasia 13(11):1019–1025

Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA (2011) Modeling read counts for CNV detection in exome sequencing data. Stat Appl Genet Mol Biol 10(1). http://www.degruyter.com/view/j/sagmb.2011.10.issue-1/1544-6115.1732/1544-6115.1732.xml.

Lupski JR (1998) Genomic Disorders: Structural Features of the Genome Can Lead to DNA Rearrangements and Human Disease Traits. Trends Genet 14(10):417–422. doi:10.1016/S0168-9525(98)01555-8

Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M (2010) Detecting copy number variation with mated short reads. Genome Res 20(11):1613–1622. doi:10.1101/gr.106344.110

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59–65. doi:10.1038/nature09708

Murtaza M, Dawson S-J, Dana WY, Tsui DG, Forshew T, Piskorz AM, Parkinson C et al (2013) Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 497(7447):108–112. doi:10.1038/nature12065

Nord AS, Lee M, King M-C, Walsh T (2011) Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics 12(1):184. doi:10.1186/1471-2164-12-184

Park G, Gim J, Kim A, Han K-H, Kim H-S, Seung-Ha O, Park T, Park W-Y, Choi BY (2013) Multiphasic analysis of whole exome sequencing data identifies a novel mutation of ACTG1 in a nonsyndromic hearing loss family. BMC Genomics 14(1):1–9. doi:10.1186/1471-2164-14-191

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C et al (2008) Copy number variation and evolution in humans and chimpanzees. Genome Res 18(11):1698–1710. doi:10.1101/gr.082016.108

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC et al (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29(6):512–520. doi:10.1038/nbt.1852

Plagnol V, Curtis J, Epstein M, Mok KY, Stebbings E, Grigoriadou S, Wood NW et al (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28(21):2747–2754. doi:10.1093/bioinformatics/bts526

Ramachandran A, Micsinai M, Pe'er I (2011) CONDEX: copy number detection in exome sequences. In: 2012 IEEE international conference on bioinformatics and biomedicine workshops, 0:87–93. IEEE Computer Society, Los Alamitos, CA. doi:10.1109/BIBMW.2011.6112359

Sathirapongsasuti JF, Lee H, Basil AJ, Horst GB, Cochran AJ, Binder S, Quackenbush J, Nelson SF (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27(19):2648–2654. doi:10.1093/bioinformatics/btr462

Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. Annu Rev Genomics Hum Genet 7(1):407–442. doi:10.1146/annurev.genom.7.080505.115618

Shi Y, Majewski J (2013) FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data. Bioinformatics 29(11):1461–1462. doi:10.1093/bioinformatics/btt151

Shinawi M, Cheung SW (2008) The array CGH and its clinical applications. Drug Discov Today 13(17–18):760–770. doi:10.1016/j.drudis.2008.06.007

Shlien A, Malkin D (2010) Copy number variations and cancer susceptibility. Curr Opin Oncol 22(1):55–63. doi:10.1097/CCO.0b013e328333dca4

Sindi SS, Önal S, Peng LC, Hsin-Ta W, Raphael BJ (2012) An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol 13(3):R22. doi:10.1186/gb-2012-13-3-r22

Sobreira NLM, Gnanakkan V, Walsh M, Marosy B, Wohler E, Thomas G, Hoover-Fong JE, Hamosh A, Wheelan SJ, Valle D (2011) Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. Genome Res 21(10):1720–1727. doi:10.1101/gr.122986.111

Southard AE, Edelmann LJ, Gelb BD (2012) Role of copy number variants in structural birth defects. Pediatrics 129(4):755–763. doi:10.1542/peds.2011-2337

Szatkiewicz JP, Neale BM, O'Dushlaine C, Fromer M, Goldstein JI, Moran JL, Chambert K et al (2013) Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. Mol Psychiatry 18(11):1178–84. doi:10.1038/mp.2013.98, http://www.nature.com/mp/journal/vaop/ncurrent/abs/mp201398a.html

Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A et al (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet 88(4):469–481. doi:10.1016/j.ajhg.2011.03.013

Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, Hanscom C et al (2012) Clinical diagnosis by whole-genome sequencing of a prenatal sample. N Engl J Med 367(23):2226–2232. doi:10.1056/NEJMoa1208594

Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet 19(R2):R145–R151. doi:10.1093/hmg/ddq333

Valdés-Mas R, Bea S, Puente DA, López-Otín C, Puente XS (2012) Estimation of copy number alterations from exome sequencing data. PLoS One 7(12), e51422. doi:10.1371/journal.pone.0051422

Wang K, Li M, Hadley D, Liu R, Glessner J, Struan FA, Grant HH, Bucan M (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17(11):1665–1674. doi:10.1101/gr.6861907

Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14(2):125–138. doi:10.1038/nrg3373

Whale AS, Huggett JF, Cowen S, Speirs V, Shaw J, Ellison S, Foy CA, Scott DJ (2012) Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. Nucleic Acids Res 40(11):e82–e82. doi:10.1093/nar/gks203

Wu J, Grzeda KR, Stewart C, Grubert F, Urban AE, Snyder MP, Marth GT (2012) Copy number variation detection from 1000 genomes project exon capture sequencing data. BMC Bioinformatics 13(1):305. doi:10.1186/1471-2105-13-305

Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD et al (2011) Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci 108(46):E1128–E1136. doi:10.1073/pnas.1110574108

Xi R, Lee S, Park PJ (2012) A survey of copy-number variation detection tools based on high-throughput sequencing data. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR (eds) Current protocols in human genetics. Wiley, Hoboken, http://www.currentprotocols.com/WileyCDA/CPUnit/refId-hg0719.html

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865–2871. doi:10.1093/bioinformatics/btp394

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26(15):1895–1896. doi:10.1093/bioinformatics/btq293

Zhang F, Wenli G, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10(1):451–481. doi:10.1146/annurev.genom.9.081307.164217

Functional Annotation of Rare Genetic Variants

Graham R.S. Ritchie and Paul Flicek

Overview

Genome-wide association studies have successfully identified a growing number of common variants that robustly associate with a wide range of complex diseases and phenotypes. In the majority of cases though, the variants are predicted to have small to modest effect sizes, and, due to the technologies used, many of the signals discovered so far may not be the causal loci. As rare variation studies begin to explore the lower ranges of the allele frequency spectrum, using whole genome or whole exome sequencing to capture a larger proportion of variants, we expect to find variants with a more direct causal role in the phenotype(s) of interest. Interpreting possible functional mechanisms linking variants with phenotypes will become increasingly important.

Experimental investigation is the most direct way to establish if a candidate variant is causally involved in some phenotype, but it is a costly and time-consuming process, and so it is important to try to use as much existing relevant information as possible to prioritise variants for follow-up and to help formulate specific hypotheses

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-1-4939-2824-8_19

G.R.S. Ritchie (✉)
European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

University of Edinburgh, Edinburgh, EH16 4UX, UK
e-mail: sigraham.ritchie@ed.au.uk

P. Flicek
European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
e-mail: flicek@ebi.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_5

about functional mechanisms to inform subsequent experiments. The genome is complex and different classes of variants may have a wide range of, possibly tissue-specific, effects depending on their genomic context. In this chapter, we review some important classes of genome annotation and highlight some relevant computational tools and databases to help interpret and prioritise candidate variants depending on their genomic context. These resources may also play a role in the discovery of rare variant signals, as association techniques based on collapsing multiple rare variants together (reviewed in Chaps. 13 and 14) may use annotation of genes and regulatory elements to select biologically meaningful groups of variants, and other techniques can use prediction scores to upweight likely functional variants to increase statistical power. In this chapter, we focus on smaller-scale variants such as single nucleotide variants (SNVs) and short sequence insertions and deletions (indels), though some of the approaches we discuss may also be applied to larger structural variants.

Mapping Variants to Annotated Features

An obvious first step in trying to interpret possible functions of sequence variants is to identify overlapping genomic features that may be affected. Features of particular interest include protein-coding and non-coding genes, transcription factor binding sites and other potential regulatory regions. There are a wide range of resources and databases that can be used to identify likely functional genomic features, from very specific resources on a single class of feature such as the miRanda databases of microRNA (miRNA) target sites (Betel et al. 2007) to broad collections of annotations such as the Ensembl (Flicek et al. 2012) and UCSC (Meyer et al. 2013) databases.

For small numbers of variants, looking up the relevant loci in a genome browser, such as Ensembl or UCSC, is a convenient way to find overlapping or nearby features and to visualise variants in their genomic context. Both browsers contain a wealth of information on genes, regulatory regions and informative local genomic properties such as conservation, GC content and co-located or nearby variants (all of which we discuss in more detail later). For larger numbers of variants, automated approaches are clearly required. For simply identifying features overlapping variants, software packages such as BEDTools (Quinlan and Hall 2010) and BEDOPS (Neph et al. 2012a) provide powerful and efficient tools for computing overlaps and proximity (among other useful metrics) between large numbers of genomic loci and can read common variant file formats such as VCF and GVF and annotation files in widely used formats such as BED, GFF, GTF and SAM (more details on these formats are given in the Appendix). More variation specific tools such as the Ensembl Variant Effect Predictor (McLaren et al. 2010) and ANNOVAR (Wang et al. 2010) also identify a wide range of features overlapping variants, but can also make more specific predictions depending on the affected feature.

For many available annotations, especially those in non-coding regions, our understanding of the importance of specific genomic sequences is still in its infancy, and all we can report is that the variant overlaps the relevant annotation. For several classes of feature, such as genes and transcription factor binding sites, we have a more

detailed understanding of the importance of particular nucleotide sequences and so can make reasonably specific predictions about the effect of an allele on the element, as we discuss below. Even when we cannot take this further step, these overlaps provide some indication of the genomic context of the variant locus, and several studies, including the ENCODE consortium (Consortium, The ENCODE Project 2012), have found significant enrichments of trait-associated variants in less well-characterised regions, such as DNase1 hypersensitive sites, suggesting that these variants, or those nearby, affect some as-yet uncharacterised functional elements.

Variants Falling in Protein-Coding Genes

Protein-coding genes are perhaps the best understood genomic features, and given that a variant falls somewhere in an annotated gene structure, there are a number of predictions that can be made about its possible effect on gene function, such as whether it is predicted to change the amino acid sequence of the encoded protein, introduce premature stop codons or affect mRNA splicing. There are several computational tools that are designed to make these predictions that work mainly by first identifying annotated genes overlapping the variants and then applying various biologically informed rules based on both the variant location and allele sequences.

The Ensembl VEP uses a set of standardised consequence terms defined in the sequence ontology (SO) (Eilbeck et al. 2005) to describe the predicted effect of a genetic variant. The use of a standardised term set is important as it allows comparison between the results of different annotation systems, and the ontology structure supports biologically informed grouping and querying of annotation results. The VEP also provides a wide range of ancillary annotation such as cDNA and protein relative coordinates, predicted amino acid substitutions (AASs) and SIFT and PolyPhen predictions for missense variants (discussed below). Several other similar tools such as ANNOVAR and VAT (Habegger et al. 2012) work in a similar way but have different performance characteristics and vary in the amount of ancillary information available.

Variants that are predicted to have the most severe effects on coding genes include those that introduce premature stop codons, disrupt essential mRNA splicing signals and indels that change the translational reading frame. These are collectively termed “loss of function” (LoF) variants and are typically expected to be highly deleterious as they have been implicated in a number of severe diseases (MacArthur et al. 2012). Stop codons introduced early in the transcript mean that the mRNA is likely to undergo a cell surveillance process known as “nonsense-mediated decay” (NMD) (Isken and Maquat 2007) where the aberrant mRNA is degraded to avoid the production of deleterious protein isoforms and so may effectively knock-down the affected transcript. However, stop codons towards the end of the transcript may escape this process and only truncate a few amino acid residues and therefore have minimal effect on protein function, so not all premature stop variants should be considered functionally equivalent.

Frameshifting variants may lead to an entirely different translated sequence and substantial elongation or truncation of the protein product. As with premature stop

codons, the position of the variant in the coding sequence will clearly affect the severity of the variant. Hu and Ng (2012) present a new tool that aims to identify frameshift variants that are likely to be truly deleterious and find that variants that affect fewer and less conserved residues are more likely to be tolerated. Hu and Ng (2012) also find that proximal frameshift variants are frequently compensatory in that a nearby downstream variant restores the reading frame disrupted by an upstream variant, highlighting the importance of considering the haplotype background of a variant.

Variants that disrupt the essential two nucleotide donor and acceptor splice sites at either end of introns are also typically expected to severely disrupt the protein product. While these essential positions are indeed highly conserved, there is also substantial sequence conservation in the flanking nucleotides and in the branch site towards the 3′ end of the intron, so variants in these regions may also affect accurate splicing (indeed, this is one way in which "synonymous" variants in coding sequence might still have functional effects). Desmet et al. (2009) introduce a tool called the Human Splicing Finder which uses position weight matrices to predict the effect of different alleles on splicing motifs in all these relevant regions.

It is important to note that despite the expected severity of loss of function variants, there are still a substantial number of common LoF variants in human populations, and each individual is predicted to carry up to 20 such variants in a homozygous state (MacArthur et al. 2012). This observation implies that we should be cautious about the interpretation of LoF variants without further phenotypic evidence. MacArthur et al. (2012) use their extensive survey of LoF variants found in the 1000 Genomes Project data to develop a classifier that can identify genes that are likely to be tolerant of LoF variants based on conservation and protein network information, and so this approach may be used to filter LoF variants to identify those more likely to have some phenotypic effect.

Other forms of coding variant that have been the subject of substantial research are missense variants predicted to result in a single AAS; these are an interesting class of variant as it appears that some AASs do not have any noticeable effect on protein function and the underlying variants are common in human populations, while others have been implicated in a wide range of diseases—around half of the mutations implicated in human disease from the Human Gene Mutation Database (HGMD) are classified as missense (Stenson et al. 2009). Several computational techniques have been developed to try to discriminate damaging AASs from apparently benign variants. These approaches can be divided into two main classes: those that make predictions based on some biologically informed assumptions about properties of important residues and those that are trained by machine learning methods to discriminate between benign and damaging substitutions. A widely used example of the first class is an algorithm called SIFT (Ng and Henikoff 2001) which makes predictions based entirely on a protein multiple sequence alignment (MSA) by looking for evidence that a substitution at a specific residue might be tolerated because, for example, the mutant residue (or one with similar physico-chemical properties) is found at that position in a related protein from another species, or conversely if a substitution is likely to be damaging because the affected residue is highly conserved. A popular example of the second class of approaches is PolyPhen-2

(Adzhubei et al. 2010) which uses a set of missense variants annotated in the UniProt database (UniProt Consortium 2011) as involved in human disease and trains a naïve Bayes classifier to discriminate between these damaging variants and a control set of common, polymorphic variants. PolyPhen uses a set of 12 predictive features for each variant, including a similar conservation metric from an MSA as used by SIFT, three-dimensional structural data, whether the residue is in a transmembrane region or a protein domain *inter alia*. There are also a number of other tools that take similar approaches but use different sets of annotations. Thusberg et al. (2011) provide a recent review and performance comparison of several AAS prediction tools, and Liu et al. (2011) present a database called dbNSFP which contains precomputed predictions from four tools for all possible AASs in the human genome.

Given the wide variety of these AAS effect prediction tools, a few methods have recently been proposed that combine predictions from a number of different tools to try to improve performance over any single technique. One of the first such methods is known as Condel (González-Pérez and López-Bigas 2011) and integrates scores from five different predictors using a weighted average which the authors show gives a substantial improvement in sensitivity and specificity on some test sets. CAROL (Lopes et al. 2012) integrates predictions from SIFT and PolyPhen using a weighted Z-method, and the authors find that this method can outperform Condel on their test set. There are plug-in modules available for the Ensembl VEP to compute both Condel and CAROL scores for missense variants.

Proteins are typically composed of one or more functional domains, and when considering the effect of any coding variant, it is also useful to check if it might disrupt any important protein domains. There are a number of databases of well-characterised protein domains, such as Pfam (Punta et al. 2011) and InterPro (Hunter et al. 2012), and Ensembl (among other resources) provides a mapping of these domains to gene annotations.

Variation in other gene regions, such as introns and the 5′ and 3′ untranslated regions (UTRs), is typically currently annotated by tools such as the Ensembl VEP and ANNOVAR simply as an overlap. However, these regions are known to contain important signals for gene regulation and may also affect mRNA structural stability. Regulatory features in the UTRs include miRNA target sites found in the 3′ UTRs of many genes. These short sequences are bound by specific miRNAs which typically serve to suppress translation of the mRNA and act as a form of post transcriptional gene regulation. The miRanda algorithm for miRNA target prediction (John et al. 2004) can be used to identify variants that disrupt likely target sites and may also be applied to identify variants that introduce novel target sites. As well as important sequence signals for mRNA splicing, intronic regions may also contain many of the regulatory elements discussed later, such as transcription factor binding sites and enhancers.

An important consideration when interpreting all forms of genetic variants is that many human genes are subject to alternative splicing and may give rise to a number of possible transcripts, frequently depending on tissue or developmental stage. A single variant may therefore be predicted to have a number of different effects depending on which transcripts it falls in—an apparently highly deleterious premature stop codon may have little consequence if it is found in an exon that is

rarely included in any transcript. Rich and detailed annotation of alternatively spliced transcripts is therefore very important for accurate variant interpretation, and the GENCODE gene set (Harrow et al. 2012) represents perhaps the most detailed set of manually annotated transcript models available for human.

Even if a variant is predicted to affect an important transcript, it appears that even severely deleterious genetic variants may be tolerated as long as they are in a heterozygous state and so only disrupt one copy of the gene, although it appears that for some genes (termed haploinsufficient), a single functional copy is not adequate to maintain function (Huang et al. 2010). Huang et al. develop a predictive model of genes that are likely to be haploinsufficient based on a number of gene-level annotations and which can be used to further prioritise variants and highlight the importance of considering variant annotations at the organismal level.

Variants in Non-coding Genes

There is increasing interest in transcribed regions of the genome that do not give rise to protein-coding mRNAs, and a number of different classes of non-coding RNA genes have now been identified and are extensively annotated in the GENCODE resource. There has been less work on interpreting the possible effects of variants in non-coding genes, but some of the approaches described above, such as annotation of variants affecting splicing, may also be applied to these.

The function of many RNA genes depends on the secondary structures formed after the RNA has been transcribed from genomic sequence. Intra-strand base pairing is an important factor in determining this structure, and sequence variants that disrupt base complementarity may thus affect the function of RNA genes. The RNAsnp server (Sabarinathan et al. 2013) uses RNA structure prediction algorithms from the Vienna package (Hofacker 2003) to predict the possible effect of variants on RNA secondary structure.

Some specific classes of RNA genes have other well-characterised functional sequence regions. As discussed above, miRNAs serve an important role in gene regulation, and they do so by binding specific sequences in the UTRs according to base pair interactions. Sequence variants in the binding regions of mature miRNA transcripts may therefore have potentially complex downstream effects on regulatory networks.

Intergenic and Regulatory Variants

Genetic regions remain the most well-characterised regions of the genome, but recent large-scale efforts such as the ENCODE and the NIH Roadmap Epigenomics projects have made available substantial amounts of information about biochemical activity in the ~98 % of the genome that does not encode protein. These data are varied in format and range from specific annotations identifying regions of the

genome bound by transcription factors (TFs) to broad epigenetic marks such as histone modifications and long-range chromatin interactions. Given that the majority of trait-associated variants, 88 % according to a recent survey (Hindorff et al. 2009), do not map to protein-coding loci, the availability of these data provides a promising opportunity to interpret the large numbers of non-genetic variants. It is not, however, currently clear to what extent genetic variation in many of the regions identified in these projects might have phenotypic effects.

Perhaps the most readily interpretable regulatory annotations are TF binding sites. Many TFs bind specific sequence motifs in the genome, and so variants that result in changes in these motifs, particularly at high-information content positions within the motif, might have a direct effect on the binding affinity of the relevant proteins. However, Maurano et al. (2012) find that variants at high-information content, conserved residues of the CTCF TF motifs aligned under regions with experimental evidence of CTCF binding, had no effect on binding intensity, implying there is substantial contextual buffering of variants in TF motifs, and it appears our understanding of the importance of specific sequence variants in these regions is still limited.

As with transcript splicing signals, TF motifs are typically represented as position weight matrices, and so the effect of a variant allele on an aligned motif can be calculated straightforwardly as the difference in alignment score between the two alleles. However, TF motifs are typically short—on the order of 10–20 nucleotides in length—and are found in numerous locations throughout the genome, and so most instances of motifs are unlikely to be functionally important (Pique-Regi et al. 2011). It is therefore important to consider further contextual evidence, such as protein–DNA interaction data for the TF of interest in order to increase prediction accuracy. ChIP-seq data for over 100 TFs in dozens of cell lines and tissues is available from ENCODE and Roadmap Epigenomics projects. The JASPAR database provides the largest open access database of TF motifs, and software such as MOODS (Korhonen et al. 2009) and the MEME suite (Bailey et al. 2009) can be used to align these motifs to sequence of interest and to check the effect of sequence variants. The Ensembl VEP identifies variants that overlap TF motifs lying in matched ChIP-seq peaks and identifies if the variant allele increases or decreases the match to the motif consensus sequence and if the variant lies in a high-information position within the motif.

Active regulatory regions are often recognisable by an accessible chromatin environment, and so assays which identify regions of open chromatin, such as DNase1 hypersensitivity and FAIRE (formaldehyde-assisted identification of regulatory elements), can help identify regulatory elements. DNase1 footprinting (Neph et al. 2012b) can identify specific genomic regions that are likely bound by proteins even when the specific factor cannot be identified and so provide a more specific prediction of a functionally important region. Data from both assays are again available in a wide range of tissues and cell lines. The potential role of variants in establishing accessible chromatin is still not well understood, but Degner et al. (2012) find thousands of variants with significant association with differential chromatin accessibility and argue that variants in these regions may make an important contribution to phenotypic variation.

Other available data include epigenetic marks such as DNA methylation and various histone modifications that mark actively transcribed or repressed genomic

regions and which are associated with regulatory elements such as enhancers and promoters. Two recent software packages, ChromHMM (Ernst and Kellis 2012) and Segway (Hoffman et al. 2012), integrate open chromatin and histone modification data to segment the entire genome into distinct functional regions. They find that these methods identify biologically important regions such as transcription start sites and enhancers. Annotations from these tools may be used to identify the likely functional context of non-coding variants, though we have relatively little understanding of the effect of sequence variation on the elements discovered, and because these tools do not take the sequence into account, it is not possible to compare different predictions for different alleles.

Data from the various techniques discussed here are typically made available in BED (or similar) format (see the Appendix for a description of this file format), and so variants can be annotated as overlapping or lying near these elements as described earlier. There are also Web resources available to identify occupied annotations given variant identifiers or coordinates. RegulomeDB (Boyle et al. 2012) finds overlaps with a wide range of data from the ENCODE project and TF motif alignments and then assigns a rule-based score based on the consistency and specificity of available annotations. HaploReg (Ward and Kellis 2012b) similarly finds overlaps with non-coding annotations but also provides information about linked variants and their associated annotations.

Conservation and Constraint

Genomic regions conserved by natural selection over evolutionary time are likely to be functionally important. By comparing the human sequence to that of other primate and mammalian genomes, we can identify regions and even specific nucleotides that appear to be under constraint. Conservation metrics derived from these sequence alignments provide a powerful means to identify potentially functional sequence features even in the absence of further evidence and can be used to identify and prioritise potentially important variant loci, even within annotation categories. Indeed, several of the quantitative approaches we discussed above make extensive use of conservation information, either at the DNA or protein sequence levels, to derive their scores.

There are several methods that can provide nucleotide resolution conservation scores (important for annotating SNVs), including GERP (Davydov et al. 2010) and phyloP (Siepel et al. 2006), which are based on different algorithmic approaches, but which both use multiple sequence alignments to identify genomic regions with less variation than would be expected under some background model. Nucleotide level conservation scores can also be used to identify runs of especially constrained sequence, which may correspond to functional elements, and these regions can also be used as an informative regional annotation.

Conservation has proven to be an important signal in coding regions, but many regulatory elements appear to have a much faster evolutionary rate, and there is frequently little detectable evolutionary conservation, for example, Schmidt et al. (2010)

find that most binding events for the two transcription factors they study are species specific even among vertebrates. The recent availability of allele frequency data across the genome from projects such as the 1000 Genomes Project (Consortium, The 1000 Genomes Project 2012) offers an alternative approach to estimating constraint on sequence features at potentially shorter timescales than possible using interspecies comparison. Ward and Kellis (2012a) use several metrics of sequence diversity such as variant density, heterozygosity and derived allele frequency computed from the 1000 Genomes Project data to demonstrate that a wide range of non-coding elements demonstrate detectable levels of constraint in human populations. These measures can potentially be used to prioritise variants according to the constraint of overlapping annotations.

Integrative Approaches

Recently, two complementary techniques have been released that integrate a wide variety of the classes of data discussed above with the aim of prioritising candidate functional variants. GWAVA (Ritchie et al. 2014) is a method aimed to identify likely functional regulatory variants and consists of a classifier trained to discriminate between annotated regulatory variants involved in human disease from the HGMD from several different sets of control variants from the 1000 Genomes Project. Features used to differentiate between these classes of variants include genetic context, regulatory annotations, conservation and measures of variation in human populations. The authors demonstrate that the method can identify likely functional variants in a number of contexts relevant to human genetics studies. CADD (Kircher et al. 2014) is also an integrative approach that includes several of the same annotations used in GWAVA, but is also applicable to variants in coding regions as it incorporates transcript-level annotations from the Ensembl VEP and predictions from SIFT and PolyPhen (described earlier). Instead of training on known disease-implicated variants, CADD is trained to discriminate between variants that have become fixed in the human lineage, which presumably represent tolerable variation, from simulated variants unobserved in human populations. This approach is appealing as it can assign a single score to variants falling in any class of genomic element and supports a systematic approach to ranking and prioritising variants across the genome.

Overlap with Known Variants and Associated Loci

While the majority of variants discovered so far in the human genome have not been characterised, an obvious aid to the interpretation of some candidate variant is to check for co-located or nearby variants with some established phenotypic association. These data may take a range of forms, from statistical association with a complex

phenotype such as a GWAS signal to empirical evidence that the variant results in increased expression of some particular gene. Locus-level phenotypic annotation, such as the effect of a gene knockout in a model organism, can also provide useful insight into the possible functional role of a genetic or regulatory variant.

There are a number of useful databases that can be consulted to find known phenotype associations; these can typically be queried either by the variant locus or phenotype of interest. The HGMD (Stenson et al. 2009) aims to collect variants that are "responsible for human inherited disease" and contains thousands of variants curated from the literature that have been implicated in a wide range of human diseases, though with a bias towards monogenic disorders. The Online Mendelian Inheritance in Man (OMIM) resource also includes detailed characterisation of human genes and associated phenotypes and includes some related genetic variants. The NHGRI GWAS catalogue (Hindorff et al. 2009) collects information from GWAS studies and identifies both specific variants and nearby loci associated with the relevant phenotypes.

Even in the absence of any phenotypic data, it is useful to establish if a candidate variant is novel or has been discovered before to find allele frequency information in different populations. A rare variant in one population may be common elsewhere in the world, and as discussed above, allele frequency can be informative about functional constraint. Data from large variant discovery studies such as the HapMap, 1000 Genomes and NHLBI Exome Sequencing Projects can be used to find allele frequencies for several populations around the world. These data are also collated centrally in the Ensembl and dbSNP databases, among other resources.

Summary

Next-generation association studies using sequencing technologies are already exploring the phenotypic consequences of novel variants at lower allele frequencies than previously feasible, and we expect to find variants with direct effects on phenotypic variation. The various resources we have reviewed here can of course be used after an association analysis has been performed to identify candidate functional variants among those linked to the association signals and to inform hypotheses for experimental validation. However, by identifying variants a priori more likely to play a functional role in the trait of interest, annotations may also be used to increase power to discover loci in the first place. This might be especially fruitful for rare variant studies where the sample sizes needed to reliably detect associations using single locus tests are still prohibitive. In a recent study, Schork et al. (2013) find that trait-associated variants are substantially enriched in various functional categories and that annotations can help identify associations that are more likely to replicate in independent samples. We anticipate that careful incorporation of annotation resources into future association studies will yield substantial insights into the contribution of rare variants to human phenotypes.

Appendix

Relevant variant and annotation file formats:

- GFF (General Feature Format): A line-oriented, tab-delimited text file format for describing the location of genomic features. GFF was originally designed to represent gene models but is now used for a wide range of genomic features. The format requires the following eight columns on each line: sequence name, feature source, feature name, start coordinate, end coordinate, score, strand and frame. The ninth column can contain any number of attributes represented as tag-value pairs separated by semicolons.
  - http://www.sequenceontology.org/gff3.shtml
- BED (Browser Extensible Data): BED is also a general format for describing genomic features and again is a line-oriented text file which uses whitespace to delimit data columns. Only three columns are required for a valid BED file: the chromosome (or scaffold) name, the start coordinate and the end coordinate. There are nine further optional fields to include further information such as the name of the feature, associated scores and various display configurations that define how the data is represented in a genome browser. Large BED files can be converted to an efficient binary format known as bigBed.
  - https://genome.ucsc.edu/FAQ/FAQformat.html
- GTF (General Transfer Format): Originally a version of GFF specialised for representing gene models, GTF is now identical to GFF version 2.
- VCF (Variant Call Format): A text file format designed to represent sequence variants (SNVs, indels and structural variants) called against a reference sequence, with a line representing each individual variant. Required tab-delimited columns define the position and alleles of the variant, and further columns can include genotypes, quality scores and QC filters. VCF also supports the inclusion of arbitrary metadata, such as functional annotations for variants, in the INFO column (often identified with a "CSQ" tag).
  - http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
- GVF (Genome Variation Format): A version of GFF (version 3) specialised for representing genomic variants. The same columns as required for GFF are also required, but there are also a number of required attributes in the ninth column to include variant identifiers and allele sequences, etc. Optional attributes are also available which can represent functional annotations such as genetic consequences.
  - http://www.sequenceontology.org/resources/gvf.html
- SAM (Sequence Alignment/Map Format): A tab-delimited text format for representing sequence reads aligned against some reference sequence (typically a reference genome assembly). Each line represents the alignment of a single read

and has 11 mandatory fields that include details of the alignment sequence, position, quality and a compact representation of the alignment itself in CIGAR format. There is also an efficient binary version of SAM known as BAM. The SAMtools package can be used to convert between SAM and BAM formats.

  – http://samtools.sourceforge.net/

- WIG (Wiggle Track Format): WIG format is used to represent quantitative data across a reference sequence such as conservation scores, GC percentage, etc. It is again a line-oriented format with the value corresponding to each reference position represented on a separate line. Data can be represented with either fixed or variable steps between each data point. Large WIG files can be converted to an efficient indexed binary format called bigWig.

  – https://genome.ucsc.edu/FAQ/FAQformat.html

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P et al (2010) A method and server for predicting damaging missense mutations. Nature 7(4):248–249. doi:10.1038/nmeth0410-248

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server Issue):W202–W208. doi:10.1093/nar/gkp335

Betel D, Wilson M, Gabow A, Marks DS, Sander C (2007) The microRNA.org resource: targets and expression. Nucleic Acids Res 36(Database):D149–D153. doi:10.1093/nar/gkm995

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22(9):1790–1797. doi:10.1101/gr.137323.112

Consortium, The 1000 Genomes Project (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Consortium, The ENCODE Project (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74. doi:10.1038/nature11247

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(12), e1001025. doi:10.1371/journal.pcbi.1001025

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK et al (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482(7385):390–394. doi:10.1038/nature10808

Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37(9), e67. doi:10.1093/nar/gkp215

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6(5):R44. doi:10.1186/gb-2005-6-5-r44

Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. Nature Publishing Group 9(3):215–216. doi:10.1038/nmeth.1906

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S et al (2012) Ensembl 2013. Nucleic Acids Res. doi:10.1093/nar/gks1236

González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440–449. doi:10.1016/j.ajhg.2011.03.004

Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A et al (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28(17):2267–2269. doi:10.1093/bioinformatics/bts368

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F et al (2012) GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res 22(9):1760–1774. doi:10.1101/gr.135350.111

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367. doi:10.1073/pnas.0903103106

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31(13): 3429–3431

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature 9(5):473–476. doi:10.1038/nmeth.1937

Hu J, Ng PC (2012) Predicting the effects of frameshifting indels. Genome Biol 13(2):R9. doi:10.1186/gb-2012-13-2-r9

Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6(10), e1001154. doi:10.1371/journal.pgen.1001154

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A et al (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 2012(Database Issue):D306–D312

Isken O, Maquat LE (2007) Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. Genes Dev 21(15):1833–1856. doi:10.1101/gad.1566807

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. PLoS Biol 2(11), e363. doi:10.1371/journal.pbio.0020363

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315. doi:10.1038/ng.2892

Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E (2009) MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics 25(23):3181–3182. doi:10.1093/bioinformatics/btp554

Liu XX, Jian XX, Boerwinkle EE (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32(8):894–899. doi:10.1002/humu.21517

Lopes MC, Joyce C, Ritchie GRS, John SL, Cunningham F, Asimit J, Zeggini E (2012) A combined functional annotation score for non-synonymous variants. Hum Hered 73(1):47–51. doi:10.1159/000334984

MacArthur DG, Balasubramanian S, Frankish A et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828. doi:10.1126/science.1215040

Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA (2012) Widespread site-dependent buffering of human regulatory polymorphism. PLoS Genet 8(3), e1002599

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. Bioinformatics 26(16):2069–2070. doi:10.1093/bioinformatics/btq330

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M et al (2013) The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 41(Database Issue):D64–D69. doi:10.1093/nar/gks1048

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK et al (2012a) BEDOPS: high-performance genomic feature operations. Bioinformatics 28(14):1919–1920. doi:10.1093/bioinformatics/bts277

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B et al (2012b) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489(7414):83–90. doi:10.1038/nature11212

Ng P, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11(5): 863–874. doi:10.1101/gr.176601

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21(3):447–455. doi:10.1101/gr.112623.110

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C et al (2011) The Pfam protein families database. Nucleic Acids Res 40(D1):D290–D301. doi:10.1093/nar/gkr1065

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. doi:10.1093/bioinformatics/btq033

Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nature Methods 11(3):294–296. doi:10.1038/nmeth.2832

Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J (2013) The RNAsnp web server: predicting SNP effects on local RNA secondary structure. Nucleic Acids Res. doi:10.1093/nar/gkt291

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A et al (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328(5981):1036–1040. doi:10.1126/science.1186176

Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF et al (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet 9(4), e1003449. doi:10.1371/journal.pgen.1003449

Siepel A, Pollard KS, Haussler D (2006) New methods for detecting lineage-specific selection. Presented at the Proceedings of the 10th International Conference on Research in Computational Molecular Biology, RECOMB 2006: April 2–5, 2006, Venice Lido, Italy, pp 190–205

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN (2009) The Human Gene Mutation Database: 2008 update. Genome Med 1(1):13. doi:10.1186/gm13

Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32(4):358–368. doi:10.1002/humu.21445

UniProt Consortium (2011) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40(Database Issue):D71–D75. doi:10.1093/nar/gkr981

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164. doi:10.1093/nar/gkq603

Ward LD, Kellis M (2012a) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. doi:10.1126/science.1225057

Ward LD, Kellis M (2012b) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40(Database Issue):D930–D934. doi:10.1093/nar/gkr917

The 1000 Genomes Project

Adam Auton and Tovah Salcedo

Introduction

Following the publication of the draft human genome sequence in 2001 (IHGSC 2001; Venter et al. 2001), human geneticists embarked on efforts to categorize genomic differences between individuals in a systematic fashion. Multiple studies were initiated with the aim of investigating human genetic variation (Pennisi 2007), most prominently the Human Genome Diversity Project (Jakobsson et al. 2008; Li et al. 2008; Cavalli-Sforza 2005) and the International HapMap Project (2005, 2007). Many of these studies used DNA microarrays to genotype common polymorphisms across individuals from a number of populations from around the world. However, since the conclusion of these projects, direct sequencing of many human genomes has become practical and cost effective through technological improvements in massively parallel short-read sequencing methods (Metzker 2010; Lander 2011). Whole-genome sequencing allows for the discovery of previously unknown polymorphisms, including de novo, rare, or length variations, as well as genotyping of known polymorphisms.

The need to develop a more complete picture of genomic variation has become particularly relevant in recent years, as results from genome-wide association studies (GWAS) have suggested rare genetic variants as a primary candidate for explaining the heritability of many genetic disorders (Visscher et al. 2012). Such rare variants, with frequencies below 1 % in the population, are much more numerous than common variants but are generally not included on the DNA microarrays used for many GWAS. Characterizing the distribution of rare variation is therefore important for understanding the genetic structure of the human population and subsequently

A. Auton (✉) • T. Salcedo
Albert Einstein College of Medicine, 1301 Morris Park Ave, Price Center, Room 353B, Bronx, NY 10461, USA
e-mail: Adam.auton@einstein.yu.edu; tovah.salcedo@einstein.yu.edu

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_6

investigating the basis of genetic disease. The 1000 Genomes Project (hereafter abbreviated by 1000G) is an ongoing international collaborative effort to develop a deep catalog of human genetic variation, specifically with the goal of cataloging as much rare variation from the global human population as possible. Using a sample of populations from around the world, the 1000G explored a variety of sequencing and variant calling methods to maximize cost efficiency for discovery of rare variants.

The 1000G enhances our ability to perform GWAS in two ways. First, having a catalog of rare variants will allow future GWAS to consider rare, but potentially important, variants without necessitating genome sequencing of novel samples (Nielsen 2010). Second, the project will improve our understanding of how rare variants are stratified between populations. As rare variants are generally believed to be the result of recent mutations, it is expected that rare variation will be highly specific to local populations. Understanding such fine-scale stratification is critical, as unrecognized stratification can cause false disease associations to be identified (Mathieson and McVean 2012).

Project Aims

The primary goal of the 1000G was to build a comprehensive catalog of human variation, with a particular focus on rare variants. Specifically, the first stated project goal was to confidently identify at least 95 % of all variants—single nucleotide changes, small insertion or deletion events (indels), and large structural variants—which segregate at a minor allele frequency (MAF) of at least 1 % in the sampled populations across the genome (Meeting Report 2007). The second major goal of the 1000G was to identify at least 95 % of variants with a MAF of 0.1 % within coding regions of the genome (Meeting Report 2007). The project samples were derived from 27 geographical populations (Table 1), with populations selected within five continental clusters. Importantly, all data (from raw sequencing data to processed variant calls) generated by the project were made publicly available almost immediately upon completion, allowing all researchers to access the project findings prior to publication.

Project Implementation

Human samples were collected from multiple global populations with the aim of reflecting large population groups. A clustered sampling strategy was used to maximize the power to detect rare variants that were shared between local populations (Fig. 1, Table 1). Approximately 100 individuals were sampled from each of the 27 populations representing five continental regions: West Africa, the Americas, South Asia, East Asia, and Europe. Cell lines for all sampled individuals were deposited with (and are currently available through) the Coriell Institute for Medical Research (Camden, New Jersey).

Table 1 Populations sampled by the 1000G

Population	Population abbreviation	Ancestry group	Target sample size	Sequencing technologies
Yoruba in Ibadan, Nigeria	YRI	West Africa	100	Illumina, SOLiD, 454
Luhya in Webuye, Kenya	LWK	West Africa	100	Illumina, SOLiD, 454
Gambian in Western Division, The Gambia	GWD	West Africa	100	Illumina, SOLiD
Mende in Sierra Leone	MSL	West Africa	100	Illumina, SOLiD
Esan in Nigeria	ESN	West Africa	100	Illumina, SOLiD
		West Africa	*500*	
Utah residents (CEPH) with Northern and Western European ancestry	CEU	Europe	100	Illumina, SOLiD, 454
Toscani in Italia	TSI	Europe	100	Illumina, SOLiD
British from England and Scotland	GBR	Europe	100	Illumina, SOLiD
Finnish from Finland	FIN	Europe	100	Illumina, SOLiD
Iberian populations in Spain	IBS	Europe	100	Illumina, SOLiD
		Europe	*500*	
Han Chinese in Beijing, China	CHB	East Asia	100	Illumina, SOLiD, 454
Japanese in Tokyo, Japan	JPT	East Asia	100	Illumina, SOLiD, 454
Han Chinese South	CHS	East Asia	100	Illumina, SOLiD
Chinese Dai in Xishuangbanna	CDX	East Asia	100	Illumina, SOLiD
Kinh in Ho Chi Minh City, Vietnam	KHV	East Asia	100	Illumina, SOLiD
		East Asia	*500*	
African ancestry in the Southwest USA	ASW	Americas	62	Illumina, SOLiD
African Caribbean in Barbados	ACB	Americas	100	Illumina, SOLiD
Mexican ancestry in Los Angeles, CA	MXL	Americas	70	Illumina, SOLiD
Puerto Rican in Puerto Rico	PUR	Americas	90	Illumina, SOLiD
Colombian in Medellín, Colombia	CLM	Americas	89	Illumina, SOLiD
Peruvian in Lima, Peru	PEL	Americas	89	Illumina, SOLiD
		Americas	*500*	
Gujarati Indian in Houston, TX	GIH	South Asia	100	Illumina
Punjabi in Lahore, Pakistan	PJL	South Asia	100	Illumina
Bengali in Bangladesh	BEB	South Asia	100	Illumina
Sri Lankan Tamil in the UK	STU	South Asia	100	Illumina
Indian Telugu in the UK	ITU	South Asia	100	Illumina
		South Asia	*500*	
Total		**Global**	**2,500**	

The populations sampled for the 1000G and the ancestry groups with which they were clustered Population names and abbreviations were developed by NHGRI

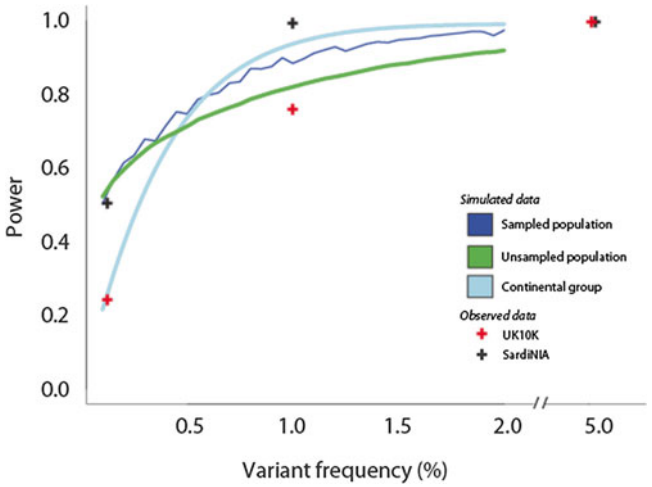

Fig. 1 Simulated data illustrating power to detect variants as a function of allele frequency using a clustered sampling scheme. The *dark blue line* represents power in the sampled populations; *green* represents power in unsampled (but closely related) populations, and *light blue* represents power across the whole continent. Also shown are estimates from real data (*plus* signs) showing the observed power of the 1000G to detect SNPs discovered by the UK10K (*black*) and SardiNIA (*red*) projects

The project aimed to describe at least three types of human genetic variants: single nucleotide polymorphisms (SNPs), small indels (<50 bp), and large-scale (≥50 bp) segmental variants (SVs). As the data generated by the project were freely accessible to everyone, research groups inside and outside of the consortium were free to use the data to test their algorithms and compare the results with others, facilitated by the data coordination cluster (Clarke et al. 2012). Data are currently available via the consortium website (http://www.1000genomes.org/), which serves as a portal for all data and analysis files and project updates.

Sequence data were generated at nine independent sequencing centers. The vast majority of the data consisted of low-coverage genome sequence augmented with targeted exome sequencing. For the low-coverage sequencing, individuals were sequenced to an average depth of coverage of 5×. Approximately three quarters of these data were generated using the Illumina platform, with the remainder being composed of SOLiD data and a small amount of 454 sequence (Table 1). In addition, the same DNA samples were sequenced using targeted exome methods, achieving a mean coverage of 85× over ~15,000 genes. As with the low-coverage sequencing, most of this data was comprised of a mix of Illumina and SOLiD. In concert with these sequencing methods, 2.4 million SNPs were also genotyped using the Illumina HumanOmni2.5-Quad SNP array.

Once the raw sequence data were collected, the 1000G used several approaches to identify variants (Fig. 2; 1000GC 2012), integrating the independent call sets produced by multiple research groups. Initial assessment of these variant call sets indicated that each had unique properties in terms of sensitivity and specificity

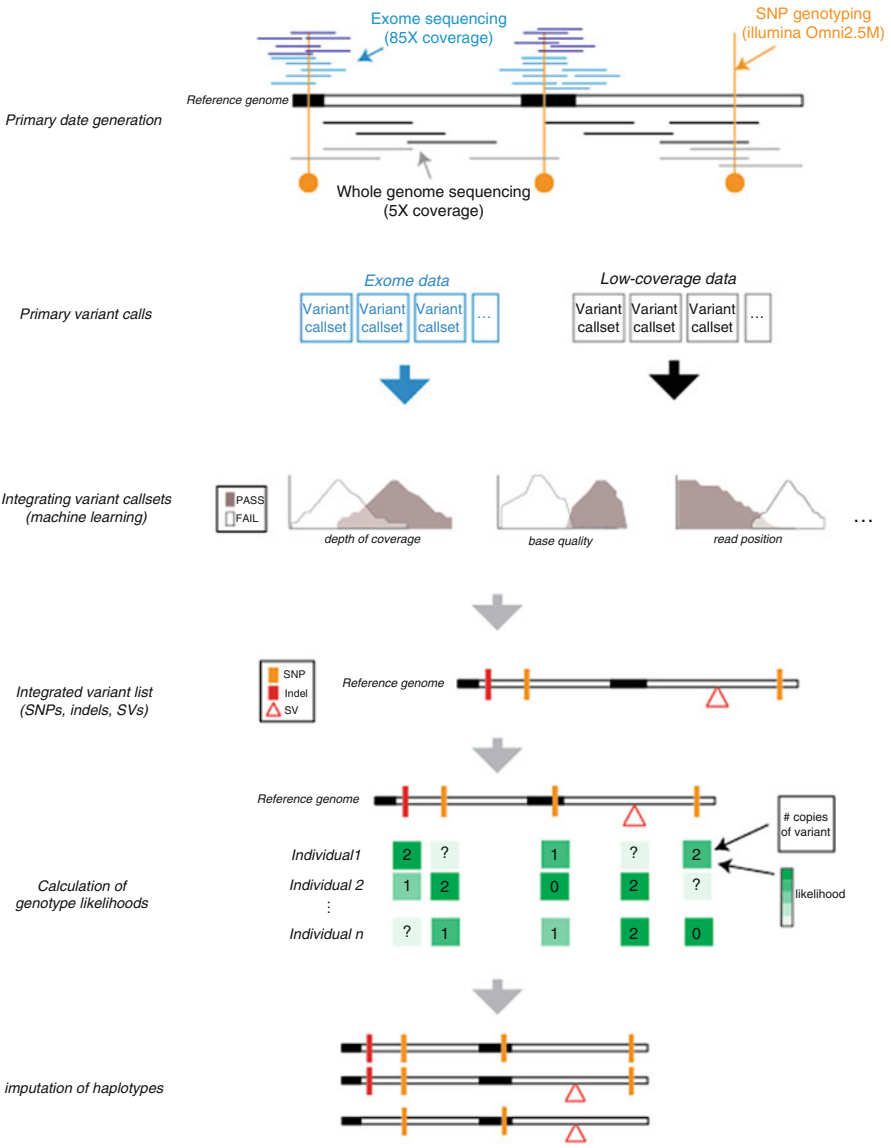

Fig. 2 Workflow schematic of 1000G, from data generation to individual genotypes. Starting with the primary short-read data, variant sites were identified using a number of differing bioinformatic algorithms. The properties of the resulting calls from the competing algorithms were passed to machine learning algorithms to identify a set of high-quality sites via comparison to known high-quality sites. Once an integrated set of variable sites had been identified, genotypes were called by using imputation to exploit the LD structure of the genome

and that a combined call set would provide a higher-quality dataset than any of the individual call sets alone. In order to select high-quality variants from the union of the call sets, multiple statistics were collected for each variant, including informa-

tion about the uniqueness of the sequence surrounding the variant, the quality of reads supporting the variant, and the distribution of variant calls in the population. Using this information, machine learning approaches were trained using Omni or HapMap genotypes to identify high-quality variants and separate them from low-quality variants.

A key feature of the low-coverage 1000G design is that the power to correctly identify a polymorphic site strongly depends on the allele count within the population sample. For variant alleles occurring at least ten times in the entire 1000G sample, the low-coverage whole-genome sequencing had similar power to detect variants as the high-coverage exome sequencing (Fig. 3a). As the size of the dataset increases, the number of variant alleles required for detection remains largely constant and hence represents increasingly lower frequency and thus increasingly rare variants. In the final 2,500-individual 1000G cohort, which represents 5,000 haploid genomes, an allele present in ten copies will represent a variant segregating at 0.2 % in the sample. As such, the 1000G is able to use the low-coverage design to efficiently discover variation in the population down to frequencies of less than 0.5 %, despite there being only limited power to identify variation in any given genome sequenced to low coverage.

The outcome of the variant calling procedure was a list of loci where variation had been detected in the 1000G panel, which then had to be integrated across individuals in order to score genotypes (Fig. 2). As low-coverage sequencing provides only an incomplete representation of a given genome, directly using such data for genotype calling would be expected to include a large number of genotypes that are either missing or misidentified. To see this, consider a heterozygote variant. If we had 8× coverage, we would have a very high probability of having observed both alleles if we had sequenced that site to a coverage of 10× or greater (Fig. 3b). However, outside of exons, the 1000G only sequenced to an average of 5×. As such, the genotype at a given site is often ambiguous. For example, ~8.7 % of genotypes in the 1000G dataset were covered by less than two reads, making it impossible to determine the genotype without prior information.

In order to address this problem, the 1000G adopted statistical imputation methods that take advantage of the correlation in genotypes between nearby sites, known as linkage disequilibrium (LD). By using such methods, high genotype accuracies could be achieved even for the low-coverage data. For example, for variants with a frequency of at least 1 %, the genotypes calling from low-sequencing data incorporating LD was nearly as accurate as calls generated from high-coverage exome sequencing (Fig. 3c; 1000GC 2012). Imputation methods therefore provided a suitable means by which accurate genotypes could be obtained. However, there remained a trend of higher genotype accuracy for samples with higher sequencing depths, although the discordance between the Illumina Omni array and sequence-based calls was mostly below 0.5 % after 8× coverage is reached (1000GC 2010).

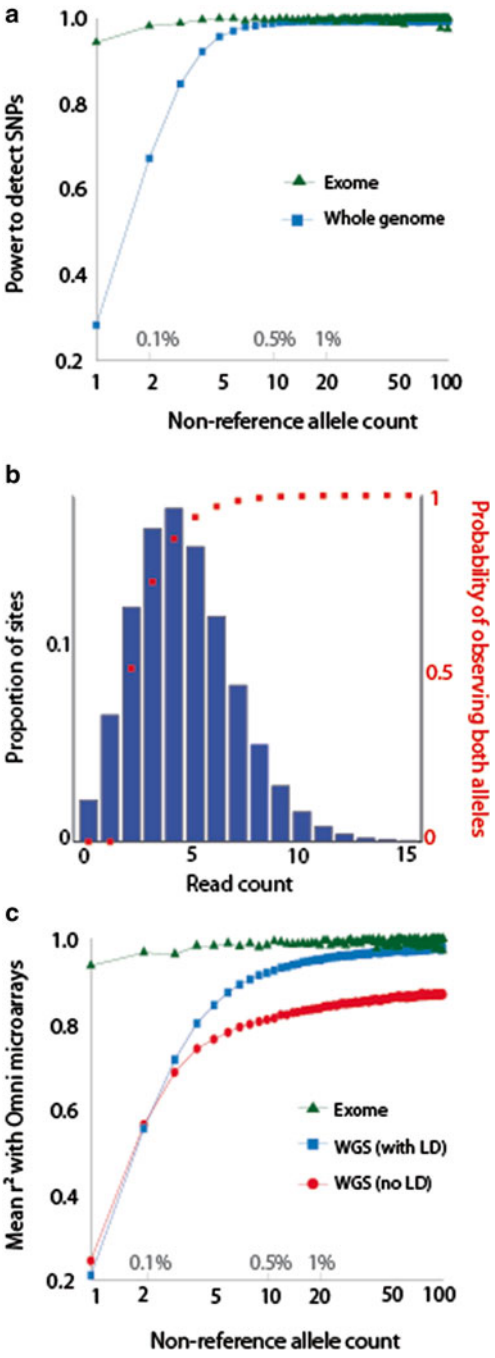

Fig. 3 Power of 1000G to detect and genotype variation. (**a**) Power to detect a variant in low-coverage WGS versus exome as a function of non-reference allele count in the sample. Reproduced from 1000GC (2012). (**b**) Probability of detecting both alleles in a heterozygous individual. Numbers reflect exemplar individual (HG00096) for a randomly selected region (chromosome 20). *Blue bars* = proportion of sites covered by a given number of reads. *Red circles* = probability of observing both alleles if a given site is heterozygous for a variant. (**c**) Genotype accuracy of low-coverage WGS and exome calls, as compared to OMNI genotypes. For the low-coverage genotypes, accuracy is shown both before and after incorporating information from linkage disequilibrium. Reproduced from 1000GC (2012)

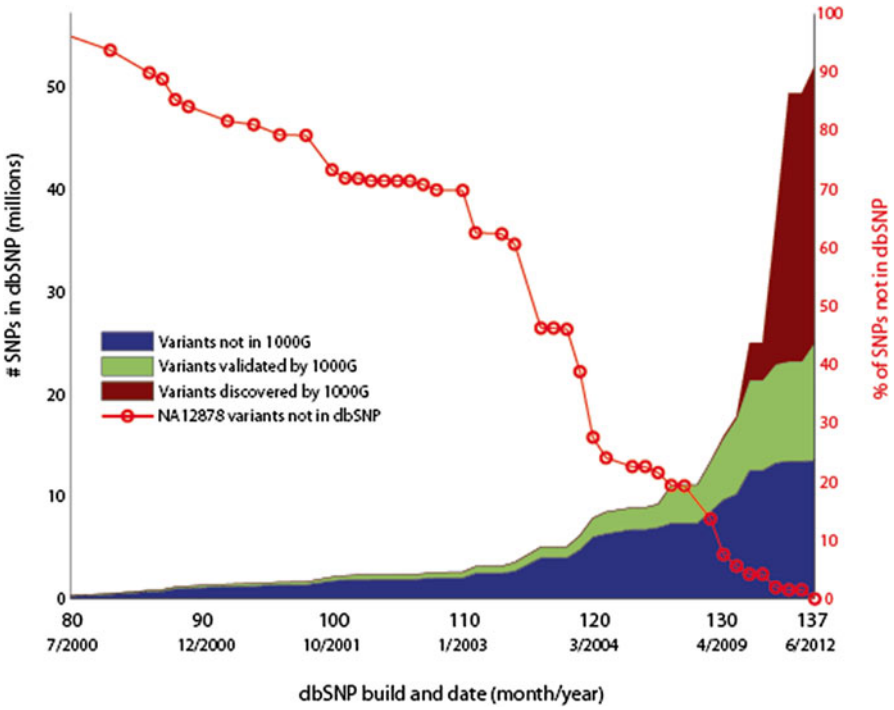

Fig. 4 Numbers of SNPs discovered over time, partitioned into variants present in dbSNP that have not been seen in 1000G (*blue*), SNPs present in dbSNP prior to the start of the 1000G that have been subsequently confirmed by the 1000G (*green*), and SNPs uniquely discovered by 1000G (*dark red*). The *red line* shows the fraction of variants in an exemplar individual (NA12878) that would have been classed "novel" relative to dbSNP as a function of dbSNP build

Results

At the time of writing, the 1000G has successfully generated genome sequence from 1,092 individuals representing 14 populations across the globe and is expanding to a set of ~2,500 individuals by the conclusion of the project. A total of 38 million SNPs have been reported from across the genome, of which 58 % on the autosomes and 77 % on the X chromosome were novel or unknown prior to the start of the project. In addition, 1.38 million autosomal indels and 14,000 SVs were discovered, 62 % and 54 % of which were novel, respectively. The impact of the 1000G can be seen by considering its contribution to the database of known human genetic variation, dbSNP. Of the ~50 million variants that have been discovered to date, roughly 50 % were identified for the first time by the 1000G (Fig. 4). In addition, ~15 % of the variants that have been discovered by other projects have been verified by the 1000G.

The 1000G estimated that 94 % of the genome was "accessible"—that is, for 94 % of the human reference assembly, the 1000G could use short-read sequencing methods to detect variation. The remaining 6 % of the human reference genome is largely comprised of highly repetitive regions that make mapping of short reads challenging, and hence the false-positive rate for calling variants is expected to be high. In addition, the 1000G also defined a smaller fraction of the genome that increased the stringency used to define the accessible genome. In this case, 72.2 % of the reference genome was retained, which may be used for analyses that require high specificity.

Major Biological Findings

Genetic Variation and Its Functional Consequences in Humans

The 1000G provides insight into the patterns of genetic variation that are expected with a typical human genome. A typical individual is expected to harbor between 3.7 and 4.7 million variants (depending on ancestry). While 3.6–3.9 million variants within a given genome are expected to be common (>5 % frequency), the remainder are expected to be less common within the population, and up to 150,000 variants may be present at less than 0.5 % frequency. The vast majority of these polymorphisms are expected to have no functional consequences, as they fall outside of coding or other functional regions of the genome. However, the dataset also provides an estimate of polymorphisms with a greater probability of having functional consequences. The number of exonic SNPs that differ between an average individual and the reference human genome included ~2,600 to 4,000 non-synonymous variants (those that change the identity of the amino acid encoded by a nucleotide triplet) and ~1,400 to 1,900 synonymous changes (those that cause changes to the triplet but not amino acid identity; 1000GC 2012). In addition, the typical individual was estimated to harbor ~72 to 91 indels in exons that caused frameshifts and ~78 to 97 in-frame indels (1000GC 2012).

In coding regions, it is clear that natural selection has operated to affect patterns of diversity in the genome. For example, the observation that 85 % of non-synonymous coding SNPs are found at MAF $\leq$ 0.5 %, while only 65 % of synonymous coding SNPs are that rare, may be explained by purifying selection preventing many non-synonymous mutations from reaching appreciable frequency in the population. Likewise, while identifying patterns associated with natural selection around noncoding regions is more difficult, there are some patterns observed by 1000G that are indicative of functional constraint. For example, DNA motifs found by the ENCODE project (ENCODE Project Consortium 2012) to be bound by the CTCF transcription factor show reduced levels of diversity compared to identical motifs elsewhere in the genome that are not bound by CTCF.

Despite the influence of natural selection, there are still a number of variants expected to have strong deleterious consequences segregating within the human

population. Among those are so-called loss-of-function (LOF) variants, which are expected to abolish gene function by creating premature stop codons, disrupting splice sites, disrupting the transcriptional reading frame, or deleting functionally important coding regions (MacArthur et al. 2012). On average, individuals were estimated to harbor ~150 LOF variants. A separate category of variants also expected to have major impacts on gene function are those rated as "damaging" according to the Human Gene Mutation Database (Stenson et al. 2009). The 1000G estimated that the average person carries 20–40 such variants. Approximately 10–20 of these LOF and 2–5 of the damaging variants were rare (MAF < 0.5 %). In addition, a typical individual appears to harbor 700–900 losses of transcription factor binding motifs, most of which correspond to common variants, and ~200 motif gains.

Distribution of Variants Across Populations

It has long been understood that modern humans had their geographical origin in Africa, and thus African populations harbor more genetic diversity than other human populations (Cann et al. 1987). Concordant with this, the greatest numbers of SNPs identified by 1000G were found in African populations, and African populations harbor the bulk of variants that are found only in a single ancestry group (Fig. 5a). Conversely, common polymorphisms with MAF ≥ 10 % tend to be represented in all populations.

As rare variants tend to be the result of recent mutations, there generally has not been sufficient time for these variants to have spread throughout the global human population from their localities of origin (Nelson et al. 2012; Maher et al. 2013). It is expected that rare variants will often be stratified by population or found very locally.

Rare variant stratification was investigated by the 1000G by considering those variants found exactly twice in the 1000G dataset. For the same reason that we expect very rare variants to be found in the same population, we predict that both copies of rare variants captured only twice overall will be found within the same population more often than in different populations, and this was observed by the 1000G (Fig. 5b). Interestingly, due to having described these captured-twice variants, the 1000G had the ability to describe relatively fine-scale cases in which variants were shared across populations from different global regions when the two populations involved have known historical connections. For example, increased sharing was identified between American populations (CLM, MXL, and PUR) and Spanish (IBS) individuals, as distinct from other European populations (Fig. 5), consistent with the known historical gene flow between these groups.

While sharing of rare variants likely reflects demographic factors, large frequency differences between populations likely reflect the action of local adaptation. Interestingly, the 1000G identified relatively few common variants that differentiate between populations within ancestry groups (1000GC 2010, 2012). Of the three major geographic clusters analyzed for locally high-frequency variants, 722

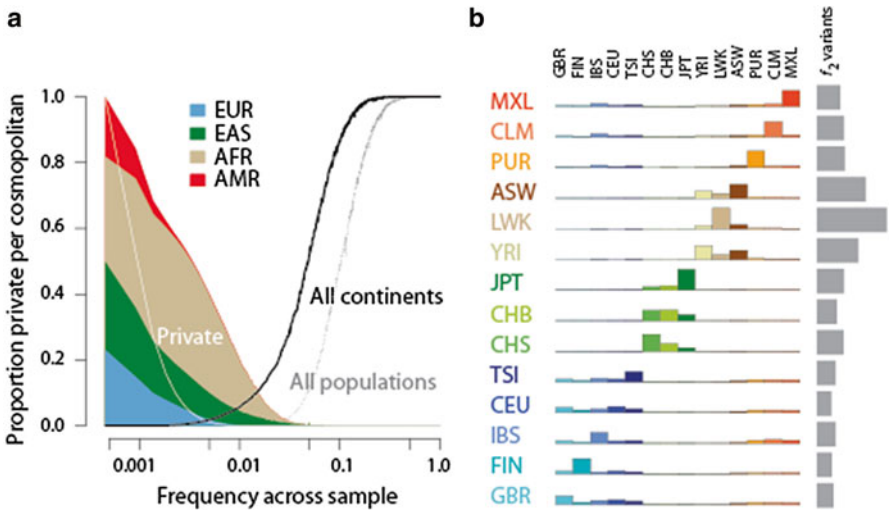

Fig. 5 Distribution of variation across geographical regions and populations. (**a**) Distribution of private variants. *Solid colors*: private variants shown by ancestry groups. *AFR* African, *AMR* Americas, *EAS* East Asian, *EUR* European. *Lines*: private variants shown across samples; *white line*—found in a single population, *solid black line*—found in all ancestry groups, *dotted black line*—found in all populations. (**b**) Population distribution of alleles observed exactly twice in 1000G data (known as "f2" variants). *Left*: fraction of f2 variants identified in a given population (*x*-axis) relative to every other population (*y*-axis). Abbreviations as in Table 1. *Right*: number of f2 variants found in a randomly sampled individual genome from each population. Reproduced from 1000GC (2012)

differentiated SNPs were observed within Africa, 530 within Asia, and 915 within Europe. However, care must be taken when interpreting such signals, as there was evidence that sequencing technology contributed to stratification among population samples, particularly within continental groups. Such artifacts highlight the care that needs to be taken when using sequencing data for population genetic analysis, although the problem may be lessened as read lengths and accuracy improve across sequencing platforms.

Identifying Insertion and Deletion Variants

Because of alignment ambiguities, insertion and deletion variants (indels) are harder to characterize than SNPs and thus remain relatively less well categorized. Correctly genotyping small indels has long been recognized as a difficult task when using short-read data (e.g., Lunter et al. 2008; Shao et al. 2013). Unfortunately, the processes that lead to local misalignment of short reads are qualitatively similar to the processes that lead to identifying true indel variation. As such, local misalignment of reads can result in patterns that are mistaken for legitimate indel variation.

Nonetheless, indel variation represents a significant fraction of human genetic variation, and the 1000G therefore aimed to incorporate indel variation into its call set. Multiple research groups contributed primary indel call sets, which were subjected to validation experiments using three orthogonal technologies in order to provide an independent assessment of variant call accuracy. Estimates provided by these validation experiments highlighted the difficulties in obtaining accurate indel calls. Specifically, the validation experiments suggested that a full 36 % of the initial indel calls were likely to be false positives. For this reason, the 1000G adopted extensive filters for the indel call set in order to select a much reduced subset expected to have a lower FDR. Subsequently, the implied FDR of indels in the integrated call set was ~5.4 %, although this estimate was based on extrapolation from the earlier validation experiments.

In addition to short indels, the project also aimed to characterize larger structural variants (SVs), which were defined as variants >50 bp in length. Given the size of these variants, novel analysis methodologies were required to enable their detection, which made use of short-read assembly features such as variation in read depth, distances between read pairs, split reads, etc. (Handsaker et al. 2011; 1000GC 2012). However, even when adopting multiple approaches, accurately calling and genotyping SVs from low-coverage data remains highly challenging. For this reason, the project focused efforts on calling a specific subset of SVs that could be called with reasonable confidence, namely, biallelic deletions (1000GC 2012). As was done for SNPs and short indels, multiple initial call sets were generated using a variety of approaches, which were subsequently combined into a single call set after a first round of validation. The resulting dataset contained 14,422 large biallelic deletions and had a relatively low FDR (2.1 %). However, it should be remembered that this call set represents only a fraction of the total amount of structural variation segregating in humans, and future work will need to focus on improving methods for detection of these more complex types of variation.

Looking Forward: Implications for Other Projects

Together, the published 1000G data accomplished the primary project goal of describing most rare variants (MAF ≥ 1–5 %) among humans, having identified an estimated 98 % of SNPs with MAF ≥ 1 %. The 1000G is perhaps the largest effort to date aiming to characterize rare human variation on a genome-wide basis and provides technical and biological insights into how to organize future studies of human genetics and genomics; indeed, the project itself has led to developments in methods and standards for analyzing short-read data. Given that hundreds of terabytes of data were generated by 1000G, both storing of data and assuring access to it by all consortium members were nontrivial. Managing such large datasets necessitated some technical innovations (Clarke et al. 2012). Two major new data file formats were developed by members of the 1000G consortium, both of which have become de facto standards within the genomics community: SAM/BAM,

which summarizes aligned sequence information (Li et al. 2009), and VCF, which summarizes variation among individuals (Danecek et al. 2011). Additionally, the 1000G maintains social media accounts and hosts a website FAQ to facilitate use of project data by investigators outside the consortium.

A major application of the 1000G is its potential use as a reference cohort for imputation (Nielsen 2010). The large scale of the 1000G dataset improves the ability to impute genotypes, particularly for rare variants, allowing researchers to improve the power and precision of GWAS. Using the 1000G, genotype data imputed for SNPs not present in GWAS arrays suggested an accuracy rate of 90–95 % for both African and non-African populations (1000GC 2012). While accuracy rates were somewhat lower for rarer variants (60–90 % for MAF 1–5 %), adding rare variants to imputed haplotypes can increase the number of identified variants in LD with GWAS hits, and accounting for haplotype structure across populations may help with identifying candidate variants for follow-up studies. Beyond imputation from microarrays, the 1000G also has implications for future GWAS designs, with simulation studies suggesting that, combined with imputation off the 1000G, very low-coverage (0.1×) whole-genome sequencing may be sufficient to conduct GWAS (Pasaniuc et al. 2012).

Despite the great advances made by 1000G, there is more to learn about rare human variation and how it is distributed across populations, particularly small and/or isolated populations. When the 1000G compared project SNPs to those described by UK10K (described elsewhere in this book) and the SardiNIA Study (Pilia et al. 2006), there was a good match for common SNPs (MAF $\geq$ 5 %): 99.7 % and 99.3 %, respectively. However, for rarer variants (MAF $\leq$ 1 %), while the 1000G catalog matched 98 % of rare variants in UK10K, only 76 % of the same rare variants were identified in SardiNIA (Fig. 1). Notably, the Sardinian population has been isolated for many generations and is the subject of intense study because of its unusual history (Pilia et al. 2006). This suggests that additional populations that are not very closely related to those sampled by 1000G may show departures from the described patterns, which could be important for future studies, particularly as very rare variants will be highly specific to local populations. It has already been documented that, even for common genetic variants, disease symptoms (e.g., lung function; Kumar et al. 2010) and medical treatment plans (e.g., drug regimen in hepatitis C viral infections; Ge et al. 2009) based on associations in one population may not be applicable in other populations, and this effect may be exacerbated when rare variants are involved (Bustamante et al. 2011; Need and Goldstein 2009).

Going forward, the 1000G provides some empirical estimates of patterns of variation that we should expect to observe in newly sequenced genomes. Following the 1000G, the vast majority of variants that are >5 % MAF are now known (Fig. 4). However, a newly sequenced genome is still likely to contain thousands of variants that have not been previously observed. Likewise, the stratification of rare variation means that sequencing efforts specific to local populations will still discover large amounts of rare variation, particularly for those not closely related to the 1000G samples. This begs the question of how surprised we should be if we do not observe a variant in the 1000G, and what information (if any) this conveys about clinical

relevance. Finally, as sequencing technologies improve, the fraction of the genome accessible to short reads will increase, allowing for previously unaccessible variants to be characterized.

References

Bustamante CD, González Burchard E, De La Vega FM (2011) Genomics for the world. Nature 475:163–165

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cavalli-Sforza LL (2005) The Human Genome Diversity Project: past, present, future. Nat Rev Genet 6:333–340

Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C et al (2012) The 1000 Genomes Project: data management and community access. Nat Methods 9:1–4

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Ge D, Fellay J, Thompson AJ, Simon JS, Shianna JV et al (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461:399–401

Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269–276

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

International Human Genome Sequencing Consortium, Lander E, Linton LM, Birren B, Nusbaum C et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, Szpeich ZA, Degnan JH, Wang K, Gurreiro R et al (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Kumar R, Seibold MA, Aldrich MC, Williams LK, Reiner AP et al (2010) Genetic ancestry in lung-function predictions. N Engl J Med 363:321–330

Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470:187–197

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 319:1100–1104

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J (2008) Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. Genome Res 18:298–309

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828

Maher MC, Uricchio LH, Torgerson DG, Hernandez RD (2013) Population genetics of rare variants and complex diseases. Hum Hered 74(3-4):118–128

Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44:243–246
Meeting Report: A workshop to plan a deep catalog of human genetic variation. September 17–18, 2007, Cambridge
Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46
Need AC, Goldstein DB (2009) Next generation disparities in human genomics: concerns and remedies. Trends Genet 25:489–494
Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100–104
Nielsen R (2010) In search of rare human variants. Nature 467:1050–1051
Pasaniuc B, Rohland N, McClaren PJ, Garimella K, Zaitlan N et al (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44:631–635
Pennisi E (2007) Human genetic variation. Science 318:1842–1843
Pilia G, Chen W-M, Scuteri A, Orrú M, Albai G et al (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2:1207–1223
Shao H, Bellos E, Yan H, Liu X, Zou J, Li Y, Wang J, Coin LJM (2013) A population model for genotyping indels from next-generation sequence data. Nucleic Acids Res 41, e46
Stenson PD, Mort M, Ball EV, Howells K, Phillips AD et al (2009) The Human Gene Mutation Database: 2008 update. Genome Med 1:13
Venter JC, Adams MD, Myers EW, Li PW, Mural MJ et al (2001) The Sequence of the Human Genome. Science 291:1304–1351
Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discoveries. Am J Hum Genet 90:7–24

The UK10K Project: 10,000 UK Genome Sequences—Accessing the Role of Rare Genetic Variants in Health and Disease

Dawn Muddyman

What Is UK10K?

From 2010 to 2013, UK10K was Britain's largest genomic sequencing consortium, awarded £10.5 million by the Wellcome Trust to investigate how low-frequency and rare genetic variants contribute to human disease (www.uk10k.org). This collaborative project brought together researchers working on obesity, autism, schizophrenia, and a number of rare conditions (familial hypercholesterolemia, thyroid disorders, learning disabilities, ciliopathies, congenital heart disease, coloboma, neuromuscular disorders, and severe insulin resistance) to generate whole genome and exome sequence data for almost 10,000 highly phenotyped individuals. The data generated by UK10K not only enabled the discovery of novel disease-causing genes by the consortium, but was also made available to the research community during the life of the project as a managed access data resource; providing access to data an order of magnitude deeper than was previously possible, and empowering future research into human genetics.

Motivation Behind the Project

Although many hundreds of genes involved in disease processes have already been discovered, the picture is far from complete. For most traits, only a small fraction of the genetic contribution has been explained, suggesting that many more disease loci remain unknown. Whilst highly valuable, linkage analyses and genome-wide

D. Muddyman (✉)
Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
e-mail: dm11@sanger.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_7

association studies (GWAS) are restricted to the identification of those genes whose variants either have strong and distinctive effects, or those that have weaker effects but are more common (minor allele frequency [MAF] ≥ 5 %). Candidate gene re-sequencing studies have demonstrated that some mutations can, however, have an effect on disease phenotypes whilst existing at a rare or low allele frequency (MAF < 5 %). Taking advantage of new technology-sequencing platforms and falling sequencing costs, the UK10K project set out to detect variants with allele frequencies as low as 0.1 %.

It was anticipated that the project's outcomes would have a far-reaching impact across the scientific, research, and medical community. The unprecedented scale and quality of the data generated has already been recognized as an excellent resource for further research into human genetics, whilst the data processing pipeline and the statistical analyses developed during the project provided examples of current best recommended practice. It is hoped that the discovery of novel, rare disease-causing variants identified by UK10K will lead to further insight into disease processes, and improvements in disease diagnoses and the development of new therapies.

UK10K Project Design

The project consisted of five key stages:

Genome-Wide Sequencing of 4,000 Cohort Samples

To maximize the amount of variation detected, whole genome sequencing at 6× depth was performed on the DNA of 4,000 highly phenotyped individuals of UK origin. It was anticipated that this coverage would provide enough power to detect all accessible SNVs, indels, and structural variants down to a 0.1 % allele frequency, and improve the accuracy of genotype calls on sequenced individuals. This 'Cohort' group as it was referred to within the project was composed equally of subjects recruited from two well-established studies: the Avon Longitudinal Study of Parents and Children (ALSPAC, www.bristol.ac.uk/alspac) and the TwinsUK study (www.twinsuk.ac.uk).

The Cohorts

TwinsUK is Britain's largest adult twin registry. Composed of more than 12,000 identical and non-identical twins, TwinsUK is an invaluable resource for studying the genetic and environmental aetiology of age-related complex traits and diseases.

The 2,000 samples selected for sequencing (one per twin pair) were taken from unrelated females from all over the UK, approximately three-fifths of which were dizygotic and two-fifths monozygotic. Where possible twins who were already part of the MuTHER (multiple tissue human expression resource), and/or HATS study (healthy ageing twin study) were preferentially selected for inclusion in Cohorts.

A core set of 63 UK10K phenotypes were selected to ensure as much overlap as possible between TwinsUK and ALSPAC phenotypic data (derived from physical examinations and questionnaires for both groups), and was made available alongside sequence data in the European genome-phenome archive (EGA). The Cohorts phenotypes included measurements for liver, kidney and lung function, cardiovascular function and hypertension, and anthropometric data such as waist and hip size, leg length, and head circumference.

ALSPAC is a longitudinal, population-based birth cohort study that recruited over 13,000 pregnant women in the Avon area, collecting data from the eighth gestational week onwards. DNA was collected from approximately 9,000 children who continued to supply phenotypic data up until the age of 18 years after which many participants re-consented their participation in the study, providing data into adulthood. In contrast to the TwinsUK samples, the 2,000 samples supplied by ALSPAC were from teenage individuals, based in and around a single region of the UK (Avon).

By collaborating with established longitudinal studies such as TwinsUK and ALSPAC, UK10K was able to investigate the contribution of genetic variants to phenotypic variation over time. Modelling correlated and longitudinal phenotypic measurements in association tests reduced phenotypic variance and increased the power of analyses. Further gains in power were achieved by imputing low-frequency variants into non-sequenced individuals with existing genome-wide association scan (GWAS) data. By preferentially including samples that overlapped with existing studies for which DNA methylation, gene expression, and metabolic profiling data were available, it was possible to explore genetic associations in functionally relevant variation, and to develop new analytical methods for incorporating functional annotation into association testing. It was anticipated that by including individuals from a continuum of trait, age, and geographical distribution, the resulting data would be widely applicable and used internationally.

Whole Genome Sequencing

A ‘production pipeline’ was developed specifically for the project, managing the flow of samples from the point of arrival through to DNA quality control (which included Picogreen quantification and Sequenom Genotyping), multiplexed sequencing and data generation, and lane QC (confirming sample identity by genotype matching, checking base quality, even-GC representation, and library insert size), prior to read pair mapping (BAM format) and variant calling (VCF format).

For whole genome sequencing, 1–3 μg DNA was sheared to 100–1,000 bp then subjected to Illumina paired-end DNA library preparation. Following size selection (300–500 bp insert size, sufficient to span Alu repeats), DNA libraries were multiplexed in a single pull-down experiment (with indexing barcodes attached prior to pull-down, enabling the sample of origin to be determined for each read) and sequenced using the Illumina HiSeq platform as paired-end 100 base reads (according to the manufacturer's protocol).

Realignment was made around known indels from the 1000 Genomes Project Pilot (1000 Genomes Project Consortium 2010) to improve raw BAM alignment for SNP calling, and then base quality scores were recalibrated using GATK (DePristo et al. 2011). BAQ tags were added using SAMtools, the BAMs were merged and then any duplicates removed. SNP and indel variants were called on the data using both SAMtools (Li 2011) mpileup and GATK UnifiedGenotyper, then merged and annotated with allele frequencies from 1000 Genomes, dbSNP entry date, and rsIDs. Functional annotation was added using the Ensembl Variant Effect Predictor against Ensembl 64, and finally BAM and VCF files were deposited in the EGA. Cumulative single-sample and multiple-sample releases were made throughout the duration of the project, enabling the scientific community to benefit from access to the data long before the end of UK10K.

Direct Association of Traits in the Sequenced Individuals to the Variants Found in Section 'Genome-Wide Sequencing of 4,000 Cohort Samples'

The next stage of the project involved directly associating newly discovered variants with quantitative traits, and identifying those variations linked with disease.

Primary association analyses focused specifically on testing the associations of intermediate and rare sequence variants with selected quantitative traits (including cardiometabolic traits, blood pressure, and body mass index), using single variant association tests and also collapsing together multiple low-frequency/rare variants in order to detect association to a gene or region. With almost 4,000 whole genome samples sequenced in total, imputed into a further 10,000 samples (approximately) using 1000 Genomes panels and IMPUTE2, power was sufficient to detect single variant associations contributing 0.1 % variance.

Secondary analyses were directed at determining the impact of novel loci on longitudinal analyses and age effects (using linear and logistic regression methods), maternal and parent of origin effects (for ALSPAC; exploiting maternal genome-wide SNP data and phenotypic records, and matching of the surrounding haplotype), and the analysis of correlated traits and pleiotropy (analysing correlated intermediate traits to assess potential pleiotropic effects at novel QTLs).

Sequencing and Association Analysis of 6,000 Exomes from Samples with Extreme Phenotypes

The UK10K project was constrained in terms of having a fixed duration and budget, and as a result limited itself to investigating three key areas of disease for which there were already some recognized rare causal variants: obesity, neurodevelopmental disorders (autism and schizophrenia), and a selection of rare conditions (including familial hypercholesterolemia, thyroid disorders, learning disabilities, ciliopathies, congenital heart disease, coloboma, neuromuscular disorders, and severe insulin resistance). To identify novel and rare variants associated with these diseases, close to 6,000 DNA samples from subjects with extreme disease phenotypes were whole exome sequenced to an average depth of 72×. High depth sequencing was necessary to enable the precise calling of rare variants, and the increased costs associated with higher depth sequencing were met by compromising on exome, rather than whole genome sequencing. Sequencing exomes (protein-coding exons and flanking conserved sequence) greatly reduced the overall sequencing costs for this stage of the project, though it was acknowledged that exome sequencing would not capture variants outside of these regions.

The objectives of this stage were to identify novel variants and genes involved in these conditions, first by association and then by determining as far as possible causal variants and mode of action. Selecting for extreme traits of interest substantially increased the power of analyses (using the Cohorts data as a common control set), and exome sequencing enabled more than 90 % of the target region to be sequenced to a sufficient enough depth to accurately call heterozygous sites (and singletons), from 4Gb total sequence per sample. The resulting publications, describing the contribution of novel variants to the genetic variation underlying these disease phenotypes, can be found on the UK10K website (All UK10K publications are available at: www.uk10k.org/publications_and_posters.html).

Whole Exome Sequencing

For whole exome sequencing 1–3 μg DNA was sheared to 100–400 bp, then subjected to Illumina paired-end DNA library preparation and enriched for target sequences according to the manufacturer's recommendations (Agilent Technologies; SureSelectXT Automated Target Enrichment for Illumina paired-end Multiplexed Sequencing). Enriched libraries were multiplexed (see section 'Whole Genome Sequencing') and sequenced using the Illumina HiSeq platform as paired-end 75 base reads (according to the manufacturer's protocol). Algorithms were developed to call base substitutions, indels, and CNVs from the exome data, using a read-depth approach.

Statistical Methods

The statistical methods applied in this stage of the project were similar to those employed in earlier stages, and broadly included:

- Imputation (both internally and into other GWAS cohorts).
- Family-based method development (for TwinsUK, Neurodevelopmental and Rare data).
- Meta-analysis of rare variants. Multivariate methods (such as Fisher's combined probability test, Stouffer's *z*-score method, SKATmeta (http://cran.r-project.org/web/packages/skatMeta/vignettes/skatMeta.pdf) and metaSKAT (Lee et al. 2013)) were used as an alternative to single-point analyses, which would have be underpowered given the size of sample sets used.
- Identification and evaluation of appropriate controls to alleviate bias in case–control analyses. Controls were selected from the Cohorts data, as well as from other exomes of non-related phenotypes both from within UK10K and from other sources such as dbGAP, the NHLBI exome-sequencing project, and the 1000 Genomes Project.
- Development of robust pipelines for quantitative trait and case–control analyses.
- Correcting for population stratification.
- Defining a genome-wide significance threshold for testing.

Great care was taken when matching cases to controls to consider the effects of population structure.

The Exome Collections

The full lists of studies sequenced as part of UK10K are described on the project website (http://www.uk10k.org/studies/) as well as in the EGA; however, a summary is presented below (and in more detail in section 'UK10K Sample Sets'):

Neurodevelopmental Disorders Group

It is estimated that neurodevelopmental traits such as autism spectrum disorders (ASDs) and schizophrenia affect up to 2 % of the world's population. Autism and schizophrenia are complex conditions involving multiple susceptibility genes and environmental factors, and often overlap in terms of characteristic clinical features. It has been proposed that these traits are part of a continuum of genetic and molecular events in the nervous system, and that rare variants in multiple genes may account for much of the unexplained susceptibility observed for these particular neurodevelopmental traits. Thus, further characterization of underlying genes and pathways as part of UK10K could significantly improve diagnostic classification for these conditions.

As both autism and early onset schizophrenia are uncommon and evidence suggests that rare, relatively penetrant alleles might be involved—it was decided at the outset that including families and individuals from isolated populations would be beneficial to enrich for genetic effects. The project selected 3,000 well-characterized cases of autism or schizoaffective disorders from founder populations and from collections of families demonstrating a clustering of cases to enrich for genetic effects and allow validation by segregation. Subjects were predominantly of UK origin (four Finnish studies were also used), and all represented genetically enriched cases—coming from families with multiple affected members (ASD and schizophrenia), representing early onset cases (schizophrenia), or being part of special interest populations (such as the Kuusamo schizophrenia study). Some cases presenting with intellectual disability as well as schizophrenia were also included, as variants associated with this more severe phenotype might have been more penetrant. Analyses to identify variants for these conditions fell into three categories:

- Family sample analyses—for families with a high loading of autism or schizophrenia, where one or a few highly penetrant variants were likely to contribute to the observed phenotype.
- Singleton analyses—where inheritance patterns could be dominant, recessive, or oligogenic.
- De novo analysis—where trios of two unaffected parents and one affected child were sequenced to identify de novo variants present in the child, but neither parents.
- Population analyses—performing single point tests and tests for aggregation of variants in genes and pathways, separately for autism and schizophrenia data.

Variants identified in these 'extreme' phenotype populations were assessed for relationships with 'normal' cognitive and behavioural traits as observed in controls.

Rare Diseases Group

Although linkage and homozygosity mapping have identified many of the causal variants underlying many mendelian diseases, the basis for many rare genetic diseases (where significant locus heterogeneity can attenuate the power of linkage studies) is much less clear. To further our understanding of such conditions, exomes were sequenced from 125 cases of each of the following conditions spanning a broad range of extreme phenotypes, some of which have the potential to respond well to therapeutic intervention:

- Severe insulin resistance
- Thyroid disorder
- Learning disabilities
- Ciliopathies
- Familial hypercholesterolaemia
- Neuromuscular disease
- Coloboma
- Congenital heart disease

Limiting the study to eight rare diseases maximized power in the presence of probable locus heterogeneity. In total 1,000 'rare disease' samples were submitted by collaborating PIs from existing collections, and sequenced. This number of samples was sufficiently powered to detect genes with causal mutations in 10 % of patients with any false 'discoveries' removed during segregation analyses and follow-up re-sequencing of candidate genes. Power was further increased as the number of causal variants per exome reduced, due to improved specificity of variant detection algorithms and better variant sampling in control datasets. Wherever possible samples from families with multiple affected members were used to enrich for genetic aetiology and enable segregation analyses. As for the neurodevelopmental disorders, there were three tiers of analyses used to discover candidate genes in this group:

- Within-family analyses—identifying candidate variants shared by affected individuals within the same family, examining trios to identifying candidate de novo variants, and examining single affecteds.
- Across-family analyses—identifying candidate genes shared by affected individuals in different families.
- Association analyses—looking at single gene and gene-set enrichment of functional variants.

Whole genome amplification and re-sequencing of candidate genes in additional patients (provided by each of the collaborating disease groups), and functional analyses in model systems were also used to determine causality for candidate genes.

Obesity Group

Obesity (defined in Caucasians as having a body mass index (BMI) > 30 kg/m^2) is a widely recognized and growing public health problem associated with type 2 diabetes, cardiovascular disease, and some cancers. Once considered a problem exclusive to high-income countries, obesity is becoming more prevalent in middle- and low-income countries.

Over the last decade, much progress has been made in the detection of monogenic causes of obesity; however, the variants found to be associated with high BMI in cohort studies are estimated to account for less than 1 % of the variance of BMI in European adults, and little is known about the causal genes underlying early onset obesity. By including both clinically extreme (obese children) and population extreme obesity (BMI > 40 kg/m^2) as a phenotype in UK10K, it was hoped to gain a better understanding of rare variants associated with this condition. A total of 1,500 samples from obese individuals were submitted from three separate studies: the Severe Childhood Onset Obesity Project (or 'SCOOP' study), the Generation Scotland: Scottish Family Health Study (http://www.genetics.med.ed.ac.uk/generation-scotland), and obese individuals from the TwinsUK cohort. Generation Scotland is a multi-institution population-based resource, aiming to identify the genetic basis of common complex diseases. The SCOOP cohort is a subset of Caucasian patients with severe early onset obesity in whom all monogenic causes

of obesity have been excluded, derived from the larger 'Genetics of Obesity (GOOS) Study' (http://www.goos.org.uk/) that consists of children with an age-adjusted BMI greater than 3 standard deviations above the mean, and obesity onset at less than 10 years old. Prior to inclusion in the study, SCOOP subjects were sequenced for *MC4R*, which contains the highest proportion of variants that cause obesity, and their leptin levels were measured.

The majority of samples were supplied by SCOOP (1,000 samples in total), with just over 400 samples from Generation Scotland, and 69 provided by the TwinsUK registry.

To uncover variants underlying both monogenic and polygenic forms of obesity, tiered filtering analyses were directed towards identifying 33 known human obesity genes and 88 functional candidates using single affected and cross-family analyses (monogenic disease-causing candidates) and case–control analyses employing regression and collapsing methods (complex obesity-associated variants). Exome-wide single variant and gene-region-based association tests were used to identify associations with obesity, and trio and family-based analyses were used (within the Generation Scotland dataset) to search for causal de novo mutations or segregating variants.

Imputation into Additional GWAS Samples

Association analyses were extended by imputing low-frequency variants into non-sequenced individuals with existing GWAS data. Genotype imputation makes predictions at un-genotyped markers in GWAS samples based upon the correlation between markers in reference panels with known sequence/dense genotypes, and less dense genotypes in GWAS data. Imputing into additional TwinsUK and ALSPAC samples, and other case–control and cohort studies with GWAS data in this way increased the power of UK10K analyses, and the potential for discovering novel variant candidates. Using a reference panel of 4,000 whole genome Cohort samples sequenced at 6×, it was possible to impute down to below 0.1 % allele frequency.

Providing a Sequence Variation Resource for Use in Further Studies

The final aim of UK10K was to provide a genotype/phenotype resource that would support future research into human genetics, making controls available for sequence-based studies and enabling imputation into other GWAS and exome-sequencing studies. To this end, whole genome and exome sequence data from the project (including basic phenotype and quantitative trait information, as well as allele

frequency summary data for the Cohorts datasets) was deposited in the EGA. Although any researcher may apply to use UK10K datasets, applicant approval and the subsequent granting of access to UK10K data is strictly managed by an independent Data Access Committee.

Managed Data Access

A major challenge for the project was creating a structure that would enable a highly diverse collection of studies to function collectively under the common goals of UK10K, without violating any of the individual studies' terms of use. To ensure absolute clarity regarding project participation, an ethical governance framework (http://www.uk10k.org/ethics.html) was devised that clearly defined UK10K policy on ethical and regulatory approvals, informed consent, data access, and withdrawal.

Implementing a mechanism for managed data access was crucial to assuring sample providers that the terms of data use would be respected for all UK10K studies (Muddyman et al. 2013). In order to access UK10K data in the EGA all prospective data users must first complete a data access application (downloaded from the UK10K website, http://www.uk10k.org/data_access.html), outlining a research proposal for specific, named datasets. The application must then be submitted to an independent Data Access Committee for review and approval, prior to data access being granted. Broadly speaking, the access agreement requires that users will respect the confidentiality and security of the data, agree that the data will only be used for research purposes and will not be redistributed, and that no attempts will be made to identify participants. It also clearly states the specific constraints imposed by Research Ethics Committees (RECs) for each of the individual studies (for example, some exome datasets may not be used for control purposes), and how specific datasets should be acknowledged. Whilst UK10K is committed to making its data as available and widely used as possible, it is equally committed to ensuring that applicants, once approved, respect the terms of data usage. Failure to abide by these terms would result in current and future access to the data being immediately withdrawn, and journal editors being alerted to the breach in use of UK10K data.

Publications

UK10K Publications fell into three categories: a flagship consortium paper describing primary analyses using the Cohorts data (under review), papers from the consortium's exome and statistical analysis groups on specific phenotypes and analytical methods, and manuscripts produced by researchers outside of the consortium. A one-year publication moratorium was imposed on all non-consortium Data Users, to protect first publication rights of the data generators. The moratorium expired on 2 July 2013 for the Cohorts datasets and 2 January 2014 for the exome datasets. Keen to ensure that output from the project was made publicly available as

soon as possible, all consortium manuscripts were sponsored to ensure immediate open access, and uploaded to the project website (http://www.uk10k.org/publications_and_posters.html).

UK10K Sample Sets

For more information on these sample sets and constraints of use, please refer to the UK10K Data Access Agreement. The following datasets *may not* be used for control purposes:

UK10K_NEURO_ASD_SKUSE
UK10K_NEURO_ASD_TAMPERE
UK10K_RARE_FIND
UK10K_NEURO_ASD_BIONED
UK10K_NEURO_ASD_MGAS
UK10K_RARE_CHD
UK10K_NEURO_ASD_TEDS
UK10K_NEURO_FSZNK
UK10K_RARE_CILIOPATHIES
UK10K_NEURO_FSZ
UK10K_NEURO_ASD_FI
UK10K_NEURO_UKSCZ
UK10K_NEURO_IMGSAC
UK10K_OBESITY_SCOOP
UK10K_RARE_COLOBOMA

The Cohorts Group

There are no constraints attached to the use of these datasets, which may be used for control purposes.

UK10K_COHORT_ALSPAC

(EGA study ID: EGAS00001000090)

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a two-generation prospective study. Pregnant women living in one of three health districts in the former county of Avon with an expected delivery date between April 1991 and December 1992 were eligible to be enrolled in the study, and this formed the initial point of contact for the development of a large, family-based resource. Information was collected on children and mothers through retrieval of biological

materials (e.g. antenatal blood samples, placentas), biological sampling (e.g. collection of cord blood, umbilical cord, milk teeth, hair, toenails, blood, and urine), self-administered questionnaires, data extraction from medical notes, linkage to routine information systems and at repeat research clinics.

UK10K_COHORT_TWINSUK

(EGA study ID: EGAS00001000108)

The TwinsUK resource is the UK's largest adult twin registry of 12,000 identical and non-identical twins, used to study the genetic and environmental aetiology of age-related complex traits and diseases. The register is predominantly female, with a mean age of mid-50. Only female twins were used to provide samples for UK10K.

The Neurodevelopmental Disorders Group

There are no constraints attached to the use of the Muir, Edinburgh, Collier, Aberdeen, Gallagher, and Gurling datasets which may be used for control purposes in analyses.

UK10K_NEURO_MUIR

(EGA study ID: EGAS00001000122)

This sample set consists of subjects with schizophrenia, autism, or other psychoses all with mental retardation (learning disability). These subjects represent the intersection of severe forms of neurodevelopmental disorders, appear to have a higher rate of familiality of schizophrenia than typical, and are likely to have more serious and penetrant forms of mutations.

UK10K_NEURO_EDINBURGH

(EGA study ID: EGAS00001000117)

This sample set comprises subjects with schizophrenia, recruited from psychiatric in- and out-patient facilities in Scotland. All diagnoses are based on standard research procedures and family histories are available. Patients have IQ > 70 and the cohort includes the following groups: 100 cases with detailed clinical, cognitive, and structural and functional neuroimaging phenotypes; 138 familial cases who are the probands of families where DNA has been collected from other affected members; 162 unrelated individuals. In most cases, patients and their families may be re-contacted to take part in further studies.

UK10K_NEURO_ASD_SKUSE

(EGA study ID: EGAS00001000114)

This sample set of UK origin consists of clinically identified subjects with Autism Spectrum Disorders, mostly without intellectual disability (i.e. verbal IQs >70). The subjects represent children and adults with autism, asperger syndrome or atypical autism, identified according to standardized research criteria (ADI-algorithm, ADOS). A minority have identified comorbid neurodevelopmental disorders (e.g. ADHD). Family histories are available, with measures of broader phenotype in first-degree relatives.

UK10K_NEURO_ASD_TAMPERE

(EGA study ID: EGAS00001000115)

This sample set consists of Finnish subjects with autism spectrum disorders (ASD) with IQs >70 recruited from a clinical centre for the diagnosis and treatment of children with ASD.

UK10K_NEURO_ASD_BIONED

(EGA study ID: EGAS00001000111)

The BioNED (Biomarkers for childhood onset neuropsychiatric disorders) study has been carrying out detailed phenotypic assessments evaluating children with an autism spectrum disorder. These assessments included ADI-R, ADOS, neuropsychology, EEG, etc.

UK10K_NEURO_ASD_MGAS

(EGA study ID: EGAS00001000113)

The MGAS (Molecular Genetics of Autism Study) samples are derived from clinical samples seen by specialists at the Maudsley hospital, and have had detailed phenotypic assessments with ADI-R and ADOS.

UK10K_NEURO_FSZ and A.2.8 UK10K_NEURO_FSZNK

(EGA study ID: EGAS00001000118 [FSZ] and EGAS00001000119 [FSZNK])

These Finnish schizophrenia samples (FSZ: Kuusamo and FSZNK: non-Kuusamo) were collected from a population cohort using national registers. The entire resource collected by the Finnish National Institute for Health and

Welfare (THL) consists of 2,756 individuals from 458 families—of whom 931 were diagnosed with schizophrenia spectrum disorder, each family having at least two affected siblings.

Samples were supplied from families originating from an internal isolate (Kuusamo) with a three-fold lifetime risk for the trait. The genealogy of the internal isolate is well documented and the individuals form a 'megapedigree' reaching back to the seventeenth century.

Samples were also supplied from families outside of Kuusamo, all of which had at least two affected siblings. All diagnoses are based on DSM-IV and for a large fraction of cases there is cognitive data.

UK10K_NEURO_ASD_FI

(EGA study ID: EGAS00001000110)

These samples are a subset of a nationwide collection of Finnish autism spectrum disorder (ASD) samples. The samples were collected from Central Hospitals across Finland in collaboration with the University of Helsinki and consisted of individuals with a diagnosis of autistic disorder or Asperger syndrome from families with at least two affected individuals. All diagnoses were based on ICD-10 and DSM-IV diagnostic criteria for ASDs.

UK10K_NEURO_IOP_COLLIER

(EGA study ID: EGAS00001000121)

This set was made up of samples taken from three different studies (all of UK origin).

The Genetics and Psychosis (*GAP*) samples, taken from subjects with schizophrenia ascertained as a new-onset case.

The Maudsley twin series consisting of probands ascertained from the Maudsley Twin Register, and defined as patients of multiple births who had suffered psychotic symptoms.

The Maudsley family study (*MFS*) consisting of over 250 families with a history of schizophrenia or bipolar disorder.

UK10K_NEURO_UKSCZ

(EGA study ID: EGAS00001000123)

These samples were collected from throughout the UK and Ireland, and fell into two main categories: cases with a positive family history of schizophrenia, either collected as sib-pairs or from multiplex kindred's—and samples that were systematically collected within South Wales, and in addition to a full diagnostic work up also underwent detailed cognitive testing. All samples obtained a DSM IV diagnosis of schizophrenia or schizoaffective disorder.

UK10K_NEURO_IMGSAC

(EGA study ID: EGAS00001000120)

Samples of UK origin were supplied from the IMGSAC cohort; an international collection of families containing children ascertained for autism spectrum disorders. Affected individuals were phenotyped using ADI-R and ADOS. Individuals with a past or current medical disorder of probable etiological significance or TSC were excluded. Where possible, the IMGSAC study performed karyotyping on one affected individual per family to exclude Fragile X syndrome.

UK10K_NEURO_ASD_GALLAGHER

(EGA study ID: EGAS00001000112)

Individuals in this Irish sample set were diagnosed with ADI/ADOS, measures of cognition/adaptive function, and approximately 50 % also presented with comorbid intellectual disability. This group represented a more severe, narrowly defined cohort of ASD subjects for the UK10K project.

UK10K_NEURO_GURLING

(EGA study ID: EGAS00001000225)

This sample set consisted of DNA from multiply affected schizophrenia families, diagnosed using the SADS-L and DSMIIIR criteria. All families were collected to ensure uni-lineal transmission of schizophrenia (i.e. families only had one affected parent with schizophrenia, or a relative of only one transmitting/obligate carrier parent with schizophrenia). Families with bi-lineal transmission of schizophrenia (i.e. with both parents being affected) were not sampled for this study. All families had multiple cases of schizophrenia and related disorders, and were selected to ensure an absence of cases of bipolar disorder both within the family and in any relatives on either side of the family.

UK10K_NEURO_ABERDEEN

(EGA study ID: EGAS00001000109)

This sample set comprises cases of schizophrenia with additional cognitive measurements, collected in Aberdeen, Scotland.

The Rare Diseases Group

There are no constraints attached to the use of the SIR, Neuromuscular, Thyroid, and Familial Hypercholesterolemia datasets, which may be used for control purposes in analyses.

UK10K_RARE_SIR

(EGA study ID: EGAS00001000130)

The Severe Insulin Resistance (SIR) sample set was supplied by the Cambridge Severe Insulin Resistance Study Cohort.

UK10K_RARE_NEUROMUSCULAR

(EGA study ID: EGAS00001000101)

These samples were taken from the Molecular Genetics of Neuromuscular Disorders Study, and fell into the following groups:

1. Congenital muscular dystrophies and congenital myopathies.
2. Neurogenic conditions.
3. Mitochondrial disorders.
4. Periodic paralysis.

UK10K_RARE_COLOBOMA

(EGA study ID: EGAS00001000127)

Ocular coloboma is the most common significant developmental eye defect with an incidence of approximately 1 in every 5,000 live births, resulting from the failure of optic fissure closure during embryogenesis. The samples used in UK10K mostly comprised isolated coloboma cases without systemic involvement (aka 'non-syndromal coloboma'). There is strong evidence from family studies that coloboma has a major genetic component with autosomal dominance being the most common pattern of inheritance. However, many cases are isolated or show complex patterns of familial clustering. The genes responsible for isolated coloboma are largely unknown, but in a small number of families mutations in SHH, CHX10, and PAX6 have been identified indicating marked genetic heterogeneity. Thus, in addition to the clinical benefits of achieving a molecular diagnosis there are also major scientific advantages to identifying coloboma genes, as these are likely to provide insights into the complex process of optic fissure closure, that is critical to normal eye development. In the longer term, understanding the molecular basis of the disease may provide clues to therapeutic strategies.

UK10K_RARE_CHD

(EGA study ID: EGAS00001000125)

The Congenital Heart Disease (CHD) samples used for UK10K were supplied from the Genetic Origins of Congenital Heart Disease Study (GOCHD Study).

UK10K_RARE_CILIOPATHIES

(EGA study ID: EGAS00001000126)

The ciliopathies are an emerging group of disorders that arise from dysfunction of cilia (both motile or immotile forms). It is predicted that over 100 known conditions are likely to fall under this category, but only a handful have thus far been studied in any depth. Most individual ciliopathies are rare with just a small number of cases having been reported, thereby presenting researchers with often insurmountable difficulties for causative gene identification. Samples were supplied from the Cilia in Disease and Development study (CINDAD).

UK10K_RARE_FIND

(EGA study ID: EGAS00001000128)

Familial INtellectual Disability (FIND) is a cohort of families with intellectual impairment. Affected family members are at the extreme end of the spectrum with the majority having moderate to severe mental retardation where the recurrence risks suggests most are likely to have monogenic causes. A subset of the cohort underwent detailed analysis of the X chromosome by Sanger sequence analysis of exomes and more recently by detailed high resolution aCGH of the X chromosome. Samples from the first study where no causal variant could be identified were selected for inclusion in UK10K. The sample set comprised mostly non-syndromic cases, selected for bias towards families with male sib-pairs to enrich for non-X linked disease genes.

UK10K_RARE_THYROID

(EGA study ID: EGAS00001000131)

Samples were supplied from two different cohorts of subjects: ‘Individuals with Congenital Hypothyroidism (CH)’ due either to dysgenesis or dyshormonogenesis; and patients with ‘Resistance to Thyroid hormone (RTH)’, a disorder characterized by elevated thyroid hormones and variable tissue refractoriness to hormone action. The CH group was enriched for genetic aetiologies by recruiting cases that were familial, on a consanguineous background or syndromic. The RTH cohort consisted of cases in which candidate gene analyses were negative.

UK10K_RARE_HYPERCHOL

(EGA study ID: EGAS00001000129)

Familial Hypercholesterolemia is a condition where the affected person has a consistently high level of LDL, which can lead to early clogging of the coronary arteries.

All patients selected for this study did not to carry the common APOB and PCSK9 mutations, and had no detectable LDLR mutations (tested for screening for 18 common mutations and SSCP, HRM, and MLPA screening for gross deletions/insertions.

The Obesity Group

There are no constraints attached to the use of the Generation Scotland and obese TwinsUK datasets, which may be used for control purposes in analyses.

UK10K_OBESITY_SCOOP

(EGA study ID: EGAS00001000124)

The Severe Childhood Onset Obesity Project (SCOOP) cohort is composed of Caucasian patients of UK origin with severe early onset obesity (all patients have a BMI Standard Deviation Score > 3 and obesity onset before the age of 10 years), in whom all known monogenic causes of obesity have been excluded.

UK10K_OBESITY_GS

(EGA study ID: EGAS00001000242)

The Generation Scotland: Scottish Family Health Study (GS:SFHS) is a family-based genetic study with more than 24,000 volunteers across Scotland, consisting of DNA, clinical, and socio-demographic data. This sample set consists of individuals from families with extreme obese subjects, including trios of extreme obese subjects with non-obese patients and multiple obese subjects within the same family.

UK10K_OBESITY_TWINSUK

(EGA study ID: EGAS00001000306)

This sample set consisted of extremely obese individuals from the TwinsUK study, with a BMI > 40.

References

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

Genomes Project Consortium (2010) A map of human genome variation from population-scale. Nature 467(7319):1061–1073

Lee S, Teslovich TM, Boehnke M, Lin X (2013) General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 93:42–53

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993

Muddyman D et al (2013) Implementing a successful data-management framework: the UK10K managed access model. Genome Med 5:100

Population Isolates

Ilenia Zara

Introduction

Population isolates have been of large interest for decades in human genetics. They were studied to successfully map highly penetrant mutations responsible for rare recessive diseases, and recently to assess complex traits and common diseases, with particular emphasis on detecting founder causative variants. The existence of large data sets, well-ascertained pedigrees, and detailed clinical records are only a subset of the features that make conducting a genetic study on population isolates convenient. In addition, the homogeneous environment and homogeneous genetic background help in minimizing noise in association tests, and the reduced genetic complexity allows highly accurate genotype imputation when using a population-specific reference panel. Furthermore, variants rare in the general population can have drifted to higher frequencies in the isolate, boosting power to detect association at these variants. However, not all isolates are alike. Here, we briefly describe the differences among isolates in terms of size, time since foundation, and early demographic history, and we discuss how these differences affect strategies in genetic studies on those populations. We also present several examples of successful and ongoing studies of complex traits on population isolates, focusing on the strategy used and on consequent results.

Population isolates are, by definition, populations resulting from a founder effect. As they start with a limited set of founders, only a subset of the genetic variability present in the original population is available at the settlement, and their genotypic

I. Zara (✉)
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,
Hinxton, Cambridgeshire CB10 1HH, UK

CRS4 (Centre for Advanced Studies, Research and Development in Sardinia), Pula, ltaly
e-mail: ilenia.zara@gmail.com

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_8

makeup can change over time under the effect of several evolutionary mechanisms, like population bottlenecks, a marked reduction in population size followed by the expansion of a small random sample of the original population, and genetic drift, the phenomenon whereby chance or random events modify the allele frequencies in a population. Population bottlenecks can originate from wars, infectious disease epidemics, or natural disruptions. The consequent reduction of the population size leads to higher levels of inbreeding, increasing the amount of linkage disequilibrium (LD), and consequently modifying the haplotype patterns. Over subsequent generations, recombination tends to break LD while inbreeding and genetic drift create it. The longer the population recovery takes after a bottleneck, the greater the effect of genetic drift is expected to be. During this process, common variants are rarely lost from an isolate, whereas rare variants may be lost or drift to higher frequencies than in the original population. Other evolutionary mechanisms, including mutation and natural selection, contribute to shape the population genetic structure, but they act in a much slower timescale than genetic drift and their effects are more significant in old isolates (Peltonen et al. 2000).

Use of Population Isolates in Genetic Studies

Taking into account the unique characteristics of the study population is extremely important, as those can influence advantages and disadvantages in genetic studies, especially for complex traits.

Population isolates vary in terms of:

- Size: i.e., macro-isolates, for instance Finnish or Sardinians with roughly 5.4 and 1.6 million inhabitants, respectively (KUNTIEN ASUKASLUVUT AAKKOSJÄRJESTYKSESSÄ 2012; http://www.sardegnastatistiche.it/documenti/12_117_20120516113258.pdf), and micro-isolates, for example, small religious communities like Old Order Amish (Arcos-Burgos 2002) and the Pomaks, who generally live in districts of 1,000–2,000 individuals, or subpopulations living in a village or clusters of villages, like the subisolates living in Ogliastra, a secluded area of Sardinia (Pistis et al. 2009) (Fig. 1), and the Mylopotamos villages in Crete.
- Time since foundation: i.e., young isolates like Kuusamo—a subisolated population founded roughly 350 years ago in northeastern Finland (Fig. 2) (Varilo et al. 2003)—relatively recent isolates like the Finnish general population—approximately 2,000 years old (Jakkula et al. 2008)—and old isolates like Sardinians, more than 10,000 years old (Contu et al. 2008; Francalacci et al. 2013).
- Early demographic history: i.e., isolates originated by a main founder event, like for example Icelanders (Helgason et al. 2001), show a substantially homogeneous gene pool (Helgason et al. 2003, 2005), whereas a significant substructure needs to be accounted for in genetic studies on isolates that experienced different waves of internal migration with multiple bottlenecks and multiple founder events, like Finnish (Jakkula et al. 2008) (Fig. 2).

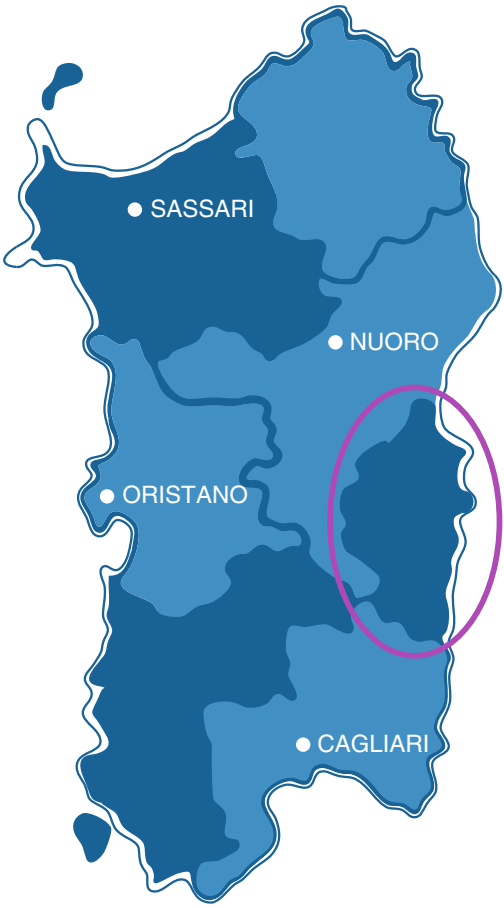

Fig. 1 The island of Sardinia and the secluded region of Ogliastra under the *circle*

Other factors, such as the number of founders and the population growth rate, contribute to determine the amount of variability present at the settlement and the role of evolutionary mechanisms in modifying it. For example, the Kuusamo population was settled by 34 families in the 1680s and reached the present-day population size of more than 16,000 individuals in less than 350 years, without experiencing significant immigration (Varilo et al. 2003). On the other hand, a large pre-Neolithic settlement has been suggested in Sardinia. The island was inhabited by ~300,000 individuals during the Bronze Age, and the population size did not significantly increase until around 300 years ago (Contu et al. 2008). So while the Kuusamo population is characterized by a high level of genetic drift, and a drastically reduced haplotype diversity (Varilo et al. 2003), the Sardinian population shows higher inter-individual variability while maintaining a substantial genetic homogeneity (Contu et al. 2008; Francalacci et al. 2013) and, as further described below, it shows evident effects of selection and carries very old mutations (Keller et al. 2012).

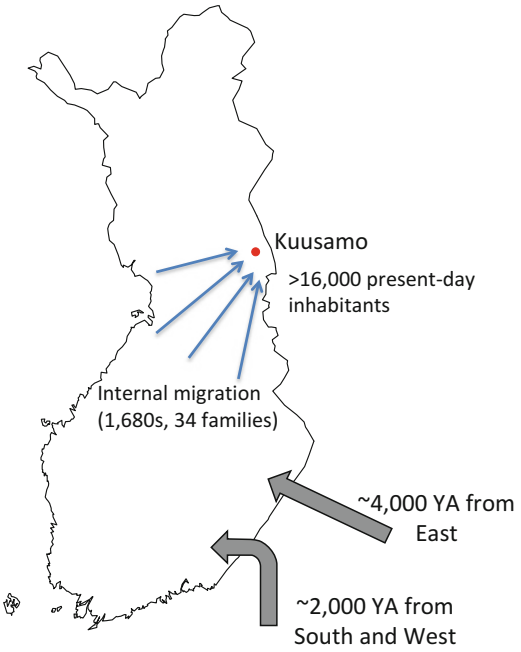

Fig. 2 Different migration waves in Finland, in particular to Kuusamo

Table 1 Advantages and disadvantages of population isolates in genetic studies for complex traits

Advantages	
More uniform environment	More uniform genetic background
Good genealogical and clinical records	Easier to standardize phenotype definitions
Reduced genetic complexity	Increased levels of LD
Enrichment in some phenotypes/diseases	Increased frequency for some disease variants
Can carry ancient variants	
Disadvantages	
Lower number of affected people	Less opportunity for replication
Lower number of variants overall	Genes less polymorphic
Association at population-specific variants cannot be replicated in other population	

All types of populations mentioned above have advantages and disadvantages that are summarized in Table 1. While outbred populations allow genetic studies to be performed on very large cohorts, the geographically restricted area in which population isolates usually live, sharing lifestyle, sanitary conditions, and exposure to pathogens, helps in minimizing the environmental contribution to complex trait variation, increasing power to detect genetic effects. The logistic advantage is particularly evident in micro-isolates, as for example the SardiNIA cohort (Pilia et al. 2006), in which volunteers living in four close towns have been measured for more than 300

quantitative traits every three years since 2001. Furthermore, diagnostic criteria and phenotypic definition are more easily standardized across a relatively restricted area, as for example in Finland, where a few medical schools with shared academic traditions train all the clinicians in the country (Peltonen et al. 2000).

In small and young isolates, the higher level of inbreeding results in an increased level of LD and in a small set of extended haplotypes (Varilo et al. 2003; Jakkula et al. 2008; Kristiansson et al. 2008). The reduced haplotype diversity allows genome scans to be performed on a significantly lower number of individuals (Shifman and Darvas 2001), and the increased level of inbreeding has inspired new methods for Identity-by-descent (IBD) detection and haplotype phasing, such as the long-range phasing (LRP) method (Kong et al. 2008). Furthermore, most strategies for association detection still use an indirect approach, i.e., the power to detect association is proportional to the extent of LD between the tested variant and the causative variant (Fig. 3) (Kruglyak 2008), so increased levels of LD can boost power.

In macro-isolates, the mean levels of LD were suggested to be only slightly higher than in more outbred populations (Eaves et al. 2001). However, this kind of isolate usually offers the possibility to collect large data sets characterized by significant inter-individual variability, while maintaining genetic homogeneity (Jakkula et al. 2008; Contu et al. 2008). This can help in better matching of cases and controls in disease studies, thus reducing the risk of detecting false positive associations. Indeed, most protein-coding variants are expected to have a geographically restricted segregation pattern, and minimizing differences in ancestry is extremely important to detect true positive associations (Do et al. 2012).

Extended and well-ascertained pedigrees are frequently available in studies on isolates, giving greater opportunity to observe the same rare variant in more chromosomes segregating through families than in a study on unrelated individuals or small families, typical of outbred cohorts. In addition, some variants that are rare or absent in the general population may have drifted to higher frequency, or may exist only in the isolate. Although associations with population-specific rare variants are hard to generalize to other populations, they can be useful to explain part of the

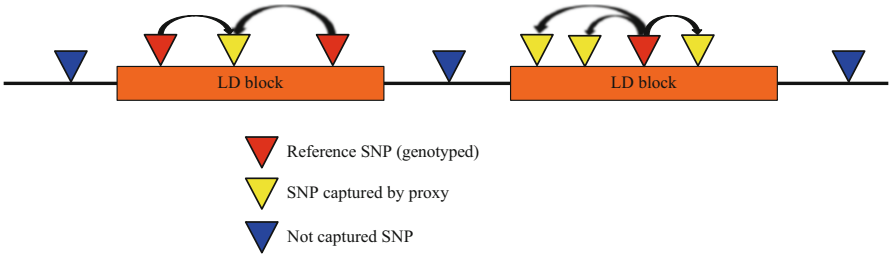

Fig. 3 Schematic representation of a genomic region to be tested for association with a phenotype. Genotyped SNPs (in *red*) are tested directly. Other associations are captured through linkage disequilibrium (by proxy) with the reference SNPs. The three SNPs indicated by *blue triangles* are neither genotyped nor in linkage disequilibrium with the reference SNPs; phenotypic association at one of these SNPs would be missed

missing heritability of complex traits (Manolio et al. 2009), as well as to better understand the underlying biological mechanisms or the etiology of a disease. For example, in a study of five LDL-cholesterol (LDL-C) associated loci in the SardiNIA cohort (Sanna et al. 2011), additional variants independently associated with LDL-C within those loci were discovered through imputation from 256 sequenced Sardinians with extreme LDL-C values. This set of variants includes a novel and rare missense variant within the *LDLR* gene that seems to be Sardinian specific. The overall findings of this study increased estimates of the heritability of LDL-C in Sardinians accounted for by these genes from 3.1 to 6.5 % (Sanna et al. 2011).

Reduced genetic complexity, resulting in a smaller amount of variants overall, may seem a disadvantage, if for example disease causing variants are very rare or absent in the study population. For instance, the C282Y mutation in the *HFE* gene, identified as the main genetic basis of hereditary hemochromatosis, is very rare in Sardinians, but it is common in northern Europeans (Candore et al. 2002). However, disease-causing genes are also expected to be less heterogeneous in isolates, and this can significantly increase the genotypic relative risk (GRR), and hence the ability to identify associated variants (Shifman and Darvas 2001). An example is the significant reduction in the number of mutations found in specific related disease genes, like *BRCA1* and *BRCA2* in Ashkenazi Jews (Roa et al. 2006). The reduced heterogeneity at complex disease-associated loci, and the relative increasing of the GRR, can also result in an enrichment of relatively common multifactorial diseases. Examples are the high frequency of autoimmune diseases in Finland and Sardinia, in particular of type 1 diabetes (T1D) and multiple sclerosis (MS) (The Diamond Project Group 2006; Pugliatti et al. 2006), and the high prevalence of MS in the Orkneys (http://www.orcades.ed.ac.uk/multiplesclerosis.html).

Finally, as mentioned above, old isolates can carry ancient mutations, and thus can be useful to reconstruct parts of human genetic history, linking old variants to archeological findings (Contu et al. 2008; Francalacci et al. 2013). For example, Sardinians have been found to be the most closely related modern European population to Ötzi, the Iceman discovered in 1991 on an Alpine glacier near the Italian-Austrian border. Ötzi is one of the oldest natural human mummies ever found, dated to ~5,300 years ago, and his complete genome has been recently sequenced (Keller et al. 2012). Analysis of the structure of common ancestry between the Iceman and present-day inhabitants of Sardinia suggested that Sardinian-related components were more widespread in Neolithic Europe, and that Ötzi was not a recent migrant (Sikora et al. 2012).

Successful and Ongoing Studies on Population Isolates

Genome-wide association studies (GWAS) have been successful in identifying common variants associated to complex traits, but a substantial portion of heritability remains unexplained (Manolio et al. 2009). In recent years, attention has shifted to low frequency and rare variants, which are hypothesized to have larger effects

(Asimit and Zeggini 2010), and high-throughput sequencing technologies are currently used to overcome the limitation of tag SNP-based genotyping. This approach is particularly useful in studies on population isolates, where the reduced genetic complexity supports high-quality imputation in large homogeneous sample sets.

Different strategies can be employed for refining genetic maps at loci of interest or over the whole genome:

- Genotyping with fine mapping or custom arrays like Illumina Immunochip, Metabochip, or Exome Chip (Cortes 2011; Voight et al. 2012; http://genome.sph.umich.edu/wiki/Exome_Chip_Design).
- Using the 1,000 Genomes Project (1KGP) resource (The 1,000 Genome Project Consortium 2012) as a reference for imputation on a scaffold of genotyped samples.
- Generate a reference panel for imputation from:
  - Low-pass whole-genome sequencing of many samples from the study population.
  - Deep sequencing or whole exome sequencing of key samples from the study population.

Power to detect rare variants associated with complex diseases or complex traits depends on several factors, such as depth and the number of sequenced samples (Li et al. 2011; Le and Durbin 2011). Costs and benefits must be carefully evaluated before choosing the most effective strategy to achieve the study goals.

Here, we briefly outline several successful and ongoing next-generation association studies, using one of these strategies or combining several of them (Fig. 4) (adapted from Zeggini et al., 2011).

The First Next-Generation GWAS: deCODE and Collaborators

The deCODE company (http://www.decode.com/) provides one of the most impressive examples of the systematic use of an extensive genealogical database, including anonymous patient records from the national health-care system, large pedigrees, and high-throughput genotyping, and sequencing data.

In 2011, Hilma Holm, Kari Stefansson and colleagues applied a next-generation association study design (Fig. 4) (Zeggini 2011), combining whole-genome sequence and GWAS data from Icelandic individuals, and detected a susceptibility locus for sick sinus syndrome (SSS) at *MYH6*, a previously unidentified susceptibility locus for the disease (Holm et al. 2011). A GWAS of 7.2 million SNPs, either directly genotyped or imputed from one or more of four sources, with 792 SSS cases and 37,592 controls, identified an association between SSS and a synonymous variant on chromosome 14q11. To refine this association, 7 SSS cases, four of which carrying the risk allele at the detected variant, and 80 controls were whole-genome sequenced at 10× depth, on average, and ~11 million detected variants were imputed into the

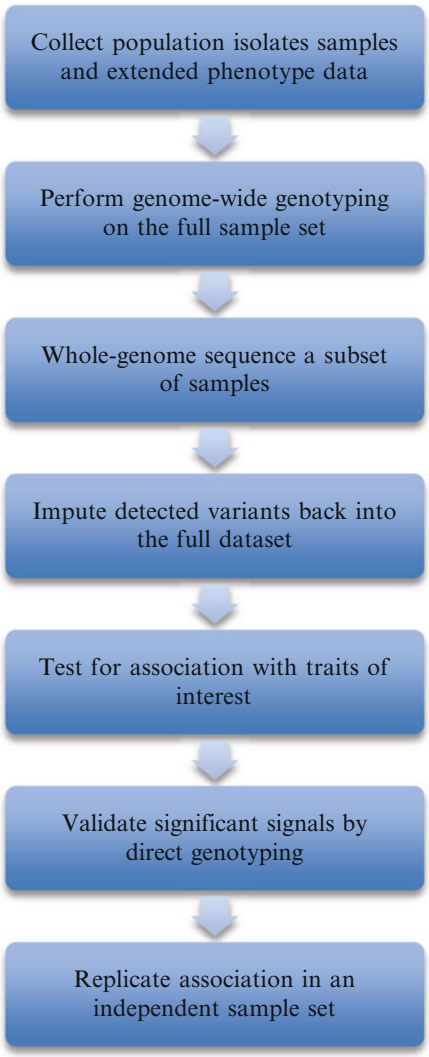

Fig. 4 Overview of the steps involved in a next-generation complex trait association study

full GWAS data set using the LRP approach (Kong et al. 2008) for phasing chip-typed samples and the IMPUTE (Marchini et al. 2007) model for imputation. Strong association was found between SSS and the c.2161C>T missense variant in exon 18 of the *MYH6* gene, encoding the alpha heavy chain subunit of cardiac myosin. No significant association remained within the 14q11 region after accounting for association with c.2161C>T, nor was found outside the 14q11 region. The c.2161C>T variant was validated through direct genotyping in 874 Icelanders and genotyping data were combined with the 87 whole-genome sequenced samples to create a new reference panel for imputation. After this imputation run, the association between SSS and c.2161C>T was stronger–$p = 1.5 \times 10^{-29}$, OR = 12.53 (95 % CI 8.08–19.44),

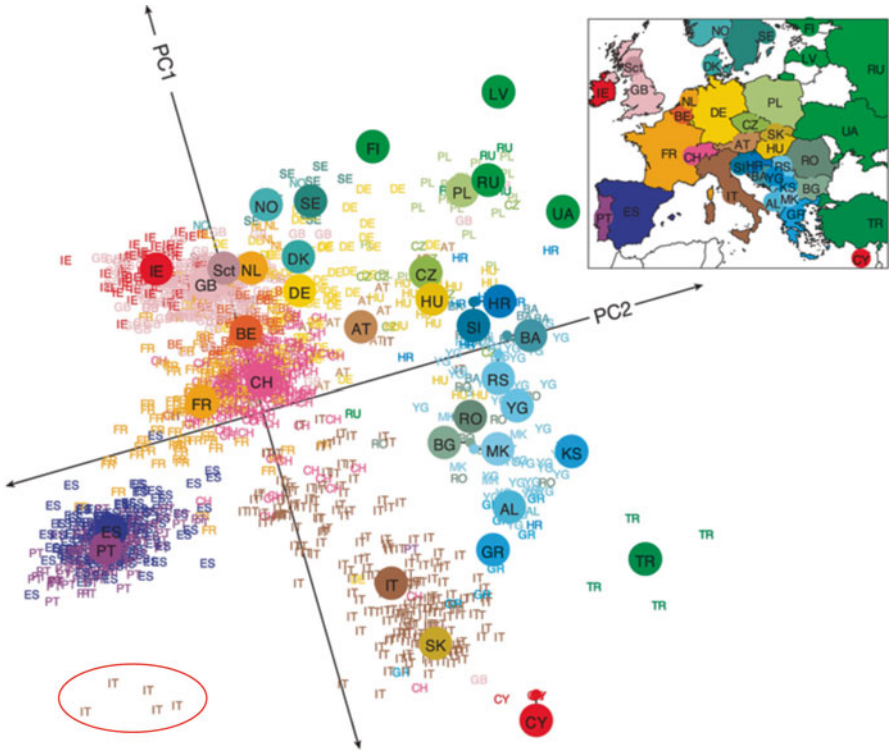

Fig. 5 Population structure within Europe (adapted from Novembre et al. 2008)

estimated risk allele frequency (RAF) in Iceland = 0.38 % — and it was replicated in 469 Icelandic cases and 1,185 controls — $p = 3.8 \times 10^{-5}$, OR = 12.95 (95 % CI 3.83–43.80), RAF = 0.21 % (Holm et al. 2011).

The lifetime risk of being diagnosed of SSS is ~50 % for c.2161C>T carriers, and ~6 % for noncarriers, and the c.2161C>T sibling recurrence risk ratio is 1.52, considerably higher than most common risk variants for complex diseases. Holm and colleagues also showed that in patients who have not been diagnosed with SSS, this variant has a substantial effect on heart rate, and that other common variants in the *MYH6* gene modulate cardiac conduction, affecting both heart rate and the PR interval, the portion between the beginning of the P wave (atrial depolarization) and the QRS complex (ventricular depolarization) of an electrocardiogram (Holm et al. 2011).

This variant was neither present in HapMap (http://hapmap.ncbi.nlm.nih.gov/) nor 1,000 Genomes Project data sets and was not identified in additional 1,776 European non-Icelander controls and 135 US cases. Consequently, it is likely to be Icelandic specific, its age is estimated to be ~870 years (or 29 generations), and it is a good example of the type of variants that can be discovered through next-generation GWAS approaches on well-characterized isolated populations (Holm et al. 2011).

Although the contribution of this particular variant may not generalize to populations outside Iceland, these results suggest that it is worth looking for other mutations in the same gene. This also provides valuable information for further analyses of the protein structure, aimed to better understand the biology of the disease (Holm et al. 2011).

HEllenic Isolate Cohorts

The HEllenic Isolate Cohorts (HELIC) project (http://www.helic.org/) is an ongoing cohort study aiming to investigate the effects of low frequency and rare variants on complex traits of medical relevance in two isolated populations, employing a next-generation GWAS approach.

Individuals enrolled in the HELIC study are from:

- The MANOLIS substudy (Minoan Isolates, the work name is in honor of Manolis Giannakakis, 1978–2010) that focuses on a set of mountainous villages (Mylopotamos villages) on the island of Crete, Greece.
- The Pomak villages, a set of religiously isolated mountainous villages in the North of Greece.

The MANOLIS population has size 4,000 and is characterized by high longevity, whereas the Pomak villages have population size of 11,000, and are characterized by a high incidence of metabolic-related cardiovascular diseases. The cohort collection, including biological samples and extensive phenotype data, started in 2009, and ~3,000 individuals were recruited and characterized for a wide array of anthropometric, cardiometabolic, biochemical, hematological, and diet-related traits.

Both cohorts were defined as genetic isolates based on genome-wide IBS statistics, which assess the degree of relatedness compared to the general Greek population, and by calculating the proportion of individuals with at least one "surrogate parent" as a means for accurate long-range haplotype phasing and imputation (Dedoussis et al. 2012; Kong et al. 2008). Indeed, 80–82 % of subjects have been found to have at least one surrogate parent in the isolates, compared to ~1 % in the outbred Greek population. Furthermore, GWAS results for glycemic traits and meta-analyses for fasting glucose confirmed 14 out of 18 previously associated loci for glycemic traits, and one of two previously associated loci for fasting glucose, providing validation of the HELIC-Pomak and MANOLIS cohorts for use in complex trait association mapping (Zeggini et al. 2012).

Recently, 250 individuals from the MANOLIS study have been whole-genome sequenced at 6× depth to enable imputation and subsequent association testing. Analysis of those whole-genome sequences is currently ongoing at the Wellcome Trust Sanger Institute, and imputation of variants detected in those samples into the full analysis cohorts will enable assessment of low frequency and rare variant associations with quantitative traits of cardiometabolic relevance (Zeggini et al. 2012).

The Orkney Complex Disease Study

The Orkney Complex Disease Study (ORCADES) is a genetic epidemiology study on inhabitants of the Orkney Islands, an archipelago in northern Scotland with Viking and pre-Anglo-Saxon British heritage (Wilson et al. 2001; http://www.orcades.ed.ac.uk/). Orkney was inhabited by the Picts, a little understood pre-Anglo-Saxon population, ~5,000 years ago. Norsemen invaded the region about AD 800, making Orkney a colony until 1468, when the islands were transferred to Scotland, and an increasing number of Scottish settlers arrived from Britain. Results of Y chromosome haplogroup analyses validated the hypothesis of an origin by admixture between Celtic and Norwegian populations. It also showed that surnames in Orkney conserve the subdivision between indigenous names, typical of the islands, and those brought to the islands with Scottish settlers (Wilson et al. 2001).

ORCADES is led by Jim Wilson, Harry Campbell, and Sarah Wild at the University of Edinburgh, and Alan Wright at the Medical Research Council, Human Genetics Unit. The study aims to discover the genetic variants influencing the risk of common, complex diseases, such as diabetes, osteoporosis, stroke, heart disease, myopia, glaucoma, and chronic kidney and lung disease in the isolated population of Orkney through analysis of next-generation genotyping and sequencing data (http://www.orcades.ed.ac.uk/). Approximately 2,200 individuals with at least two Orcadian grandparents were recruited from 2005 to 2011. Subjects were phenotyped for cardiovascular traits and some of them were further characterized for parameters related to bone and eyes clinical status. Genotypes generated for the epidemiology study are also used for population genetics projects, designed to better explore the level of homozygosity, the population structure, and the genetic history of Orkney.

Another ongoing study on Orkney is the Multiple Sclerosis in the Northern Isles of Scotland (NIMS) project. It aims to investigate the genetic and nongenetic factors contributing to the increased risk of developing the disease in Orkney and Shetland. Indeed, Orkney and Shetland are believed to have the highest prevalence of MS in the world, with ~402 cases per 100,000 in Orkney and ~295 per 100,000 in Shetland (http://www.orcades.ed.ac.uk/multiplesclerosis.html).

ORCADES contributed to the discovery of over 800 new gene associations for complex traits in collaboration with several international consortia (http://www.orcades.ed.ac.uk/).

The SardiNIA Project and the Case–Control Study of Type 1 Diabetes and Multiple Sclerosis in Sardinia

Sardinia is an island in the center of the Mediterranean Sea, whose isolated population is characterized by high inter-individual variability among the coastal regions, and strong isolation and lack of migration in the central-eastern region (Contu et al. 2008; Angius et al. 2001; Francalacci et al. 2013). Its considerable population size

allows large sample collections from both the general population and the internal isolates.

Two main projects are ongoing in Sardinia:

- SardiNIA (Pilia et al. 2006), a longitudinal study on aging and metabolic related traits, focused on ~6,100 individuals in ~800 families living in four small towns in the central-eastern region of Sardinia named Ogliastra (Fig. 1). It started on 2001, and volunteers have been characterized for more than 300 quantitative traits; measurements are repeated every three years. A subset of the SardiNIA sample set was recently characterized for more than 272 immune traits, allowing the finding of new immune cell trait-associated SNPs through next-generation GWAS (Orrù et al, 2013).
- A case–control study of MS and T1D (Sanna et al. 2010) focused on ~10,000 individuals from the general population, of which ~2,000 MS unrelated patients and ~1,000 trios, ~2,000 T1D unrelated patients, and ~3,000 unrelated controls with at least three Sardinian grandparents.

Both these studies are led by Francesco Cucca and Serena Sanna at the Istituto di Ricerca Genetica e Biomedica (IRGB-CNR), and David Schlessinger at the Laboratory of Genetics, National Institute on Aging (NIA), Baltimore, Maryland, USA, in collaboration with Gonçalo Abecasis at the Center for Statistical Genetics, University of Michigan, USA, the Center of Advanced Research and Development in Sardinia (CRS4), local Universities, and clinical centers.

Sardinians are genetically distinct from other European populations (Fig. 5) (Novembre et al. 2008). To better explore the contribution of rare and population-specific variants, an ambitious sequencing project has been undertaken (partially included in the SardiNIA Medical Sequencing Discovery Project, dbGaP Study Accession: phs000313.v1.p1) (Sanna et al. 2012). Roughly 2,000 samples from the SardiNIA cohort, and 1,500 samples from the case–control study, were sequenced at 4× depth, on average, and a reference panel for imputation is being generated from their sequence data. While waiting for the full sample set to be sequenced, several panels from subset of samples enabled preliminary imputation runs. Results on 17.6 million SNPs, 5.3 million of which not in dbSNP137, discovered in 2,120 sequences and imputed on both the SardiNIA and case–control studies were recently presented (Sanna et al. 2012; Sidore and et al. 2100; Zara et al. 2012).

Samples from the SardiNIA study were genotyped with the Illumina Metabochip and the Affymetrix 6.0 array (Nishida et al. 2008), and imputation was performed, using those arrays as baseline scaffold, on both the Sardinian and the 1,000 Genome Project (1KGP)-based reference panels. Beyond the better imputation quality and accuracy, an additional example of the advantages offered by population-specific reference panels is the association detected between the Q40X mutation in the HBB gene and a variety of blood phenotypes in the SardiNIA cohort. The Q40X mutation is responsible for the β0-thalassemia in homozygotes and protects against malaria in heterozygotes. The variant is common in Sardinia, due to balancing selection against malaria (MAF ~5 %), but very rare elsewhere. For example, on the 1KGP panel, the variant was seen only on two chromosomes, and was imputed with such

low accuracy that no association was detected (for total hemoglobin, $p=0.34$ with estimated MAF=0.02 %, whereas $p=1.7\times10^{-265}$ after imputation on the Sardinian reference panel a (Sanna et al. 2012)).

The high prevalence of MS and T1D in Sardinia is well known (The Diamond Project Group 2006; Pugliatti et al. 2006). While the prevalence of both diseases shows a North–South gradient in Europe, with a higher prevalence in the North and a lower prevalence in the South, Sardinia represents an exception to this trend. Moreover, the major risk allele for MS in Europeans, HLA-DRB1*1501 (The International Multiple Sclerosis Genetics Consortium and The Wellcome Trust Case Control Consortium 2011) has low frequency, and is only weakly associated in Sardinians (Sanna et al. 2012; Marrosu et al. 2001), suggesting that other factors contribute to increase the risk of developing MS there (Marrosu et al. 2004). Samples were genotyped with the Illumina Immunochip and the Affymetrix 6.0 array. Unrelated individuals were selected to perform two case–control studies on variants imputed from the Sardinian and the 1KGP reference panels. For MS, the major risk haplotype in Sardinians was found to be HLA-DRB1*03:01-DQB1*02:01 ($p=6.35\times10^{-45}$, OR=1.74, frequency 0.21 in controls and 0.33 in cases), while the HLA-DRB1*1501 allele was found to have frequency 1 % in controls and 2 % in cases and a p-value of 9.59×10^{-8} (Zara et al. 2012). For T1D, the imputation on the Sardinian reference set boosted association at known susceptibility loci, for example, the INS locus, where the −23HphI variant, a functional SNP previously described (Barrat et al. 2004), was associated with $p=5\times10^{-15}$ after imputation on the Sardinian reference panel, and $p=1\times10^{-7}$ after imputation on the 1KGP reference panel. Novel-associated variants, at both known and novel loci, were found for MS, and further analyses are ongoing to better understand these findings (Zara et al. 2012).

Conclusions

We have discussed how features of population isolates can influence advantages and disadvantages of using this kind of population in genetic studies. We also described several examples of genetic studies of complex traits on population isolates, focusing on the strategy used, and on consequent results.

Multifactorial traits are the result of the complex interplay between genetic and environmental risk factors, and very little is known about the environmental exposures influencing the variation of complex traits or the risk of developing a disease. Genetic studies on population isolates can help in minimizing those environmental effects and boost the power to detect association at rare variants.

High-throughput sequencing technologies are particularly useful in studies on population isolates, whose specific genetic variation is often not described, even in extensive international resources like the 1KGP and the HapMap project. For example, a key goal of the 1KGP was to identify more than 95 % of SNPs at 1 % frequency in a broad set of populations. The current resource includes 98 % of the

SNPs with frequencies of 1.0 % in 2,500 UK sampled genomes (the Wellcome Trust-funded UK10K project), but it includes only 76.9 % of the SNPs with frequencies of 1.0 % in 2,000 genomes sequenced in the SardiNIA study (The 1,000 Genome Project Consortium 2012).

The most effective strategy for a next-generation association study on an isolated population depends on the population features, and on the study goal, and costs and benefits must be carefully evaluated. The 1KGP and HapMap resources offer a valuable reference for imputation for only computational cost, but integration of in-site next-generation sequencing and GWAS data enable better exploration of the population-specific genetic variation.

This sequencing-based approach will refine GWAS results, increasing the spectrum of variants assessed, and helping to better understand biological aspects underlying the variation of complex traits and the etiology of diseases. It also confirms the value of population isolates in genetic studies for mapping complex traits.

Acknowledgments I thank Serena Sanna, Francesco Cucca, and Richard Durbin for critical revisions of the manuscript.

I also thank Eleftheria Zeggini and Andrew Morris for critical revisions and for the opportunity to contribute to this work.

References

KUNTIEN ASUKASLUVUT AAKKOSJÄRJESTYKSESSÄ. Population Register Centre. 31 August 2012. Retrieved 16 September 2012.

Angius A et al (2001) Archivial, demographic and genetic studies define a Sardinian sub-isolate as a suitable model for mapping complex traits. Hum Genet 109(2):198–209

Arcos-Burgos M, Muenke M (2002) Genetics of population isolates. Clin Genet 61:233–247

Asimit J, Zeggini E (2010) Annu Rev Genet 44:293–308

Barrat et al. "Remapping the insulin gene/IDDM2 locus in type 1 diabetes." Diabetes. 2004 Jul;53(7):1884–9.

Candore G et al (2002) Frequency of the HFE gene mutations in five Italian populations. Blood Cells Mol Dis 29(3):267–273

Contu D et al (2008) Y-chromosome bases evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS One 3(1), e1430

Cortes A, Matthew AB (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13:101

Dedoussis G et al. An evaluation of genetic characteristics of two population isolates from Greece: the HELIC-Pomak and MANOLIS studies, ASHG 2012—San Francisco, CA

Do R et al (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21(R1):R1–R9

Eaves IA et al (2001) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet 25(3): 320–323

Francalacci P et al (2013) Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341(6145):565–569

Le SQ, Durbin R (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res 21(6):952–960

Helgason A et al (2001) MtDNA and the islands of the north Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 68:723–737
Helgason A et al (2003) A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci fro relative homogeneity caused by genetic drift. Ann J Hum Genet 67:281–297
Helgason A et al (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37(1):90–95
Holm H et al (2011) A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 43(4):316–320
Jakkula E et al (2008) The genome-wide patterns of variation expose significant substructure in a founder population. Ann J Hum Genet 83:1–8
Keller A et al (2012) New insight into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun 3:698
Kong A et al (2008) Detection of sharing by descent, long-range phasing and haplotype impotation. Nat Genet 40(9):1068–1075
Kristiansson K et al (2008) Isolated population and complex disease gene identification. Genome Biol 9(8):109
Kruglyak L (2008) The road to genome-wide association studies. Nat Rev Genet 16(5):275–284
Li Y et al (2011) Low-coverage sequencing: Implications for design of complex trait association studies. Gen Res 21(6):940–951
Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nat Rev 461(7265):747–753
Marchini J et al (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913
Marrosu MG et al (2001) Dissection of the HLA association with multiple sclerosis in the founder population of Sardinia. Hum Mol Genet 10(25):2907–2916
Marrosu MG et al (2004) The co-inheritance of type 1 diabetes and multiple sclerosis in Sardinia cannot be explained by genotype variation in the HLA region alone. Hum Mol Genet 13(23): 2919–2924
Nishida N et al (2008) Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. BMC Genomics 9:431
Novembre J et al (2008) Genes mirror geography in Europe. Nature 456(7218):98–101
Orrù V et al (2013) Genetic variants regulating immune cell levels in health and disease. Cell 155:242–256
Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. Nat Rev Genet 1:182–190
Pilia G et al (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet 2(8), e132
Pistis G et al (2009) High differentiation among eight villages in a secluded area of Sardinia revealed by genome-wide high density SNPs analysis. PLoS One 4(2), e4654
Pugliatti M et al (2006) The epidemiology of multiple sclerosis in Europe. Eur J Neurol 13(7): 700–722
Roa BB et al (2006) Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nat Genet 14:185–187
Shifman S, Darvas A (2001) The value of isolated populations. Nat Genet 28:309–310. doi:10.1038/91060
Sanna S et al (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat Genet 42(6):495–497
Sanna S et al (2011) Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet 7(7), e1002198
Sanna S et al. Using low-pass whole-genome sequencing to create a reference panel for genome imputation in an isolated population, ASHG meeting 2012, San Francisco, CA
Sidore C et al. Whole Genome Sequencing of 2100 Individuals in the founder Sardinian Population, ASHG meeting 2012, San Francisco, CA

Sikora M et al. On the Sardinian ancestry of the Tyrolean Iceman—oral communication ASHG 2012—San Francisco, CA

The 1,000 Genome Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

The Diamond Project Group (2006) Incidence and trends of childhood Type 1 diabetes worldwide 1990-1999. Diabet Med 23(8):857–866

The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476(7359):214–219

Varilo T et al (2003) The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum Mol Genet 12(1):51–59

Voight BF et al (2012) The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8(8):e1002793

Wilson JF et al (2001) Genetic evidence for different male and female roles during cultural transitions in the British Isles. Proc Natl Acad Sci U S A 98(9):5078–5083, Epub 2001 Apr 3

Zara I. et al. Sequencing-based and multiplatform Genome-Wide Association study for multiple sclerosis and Type 1 DIabetes in Sardinians, ASHG Meeting 2012, San Francisco, CA

Zeggini E (2011) Next-generation association studies for complex traits. Nat Genet 43(4): 287–288

Zeggini E. et al. Validation of the HELIC population isolate collections as cohorts for complex trait association mapping. ASHG 2012—San Francisco, CA; 2012

Natural Selection at Rare Variants

Yali Xue and Chris Tyler-Smith

Introduction

'Selection' refers to the non-random increase or decrease in allele frequency in a population over a number of generations. While it can potentially act on any variant whatever its frequency, rare variants raise particular issues because their low frequency usually reflects a recent origin, and thus few generations for past selection to have acted and influenced their spread. Thus, the consequences of any selection may be difficult to detect, and several of the approaches used for detecting selection acting on common variants may not be useful for rare variants. Hence, there is a need for a chapter focussing specifically on selection at rare variants.

In several of the following sections, it will be useful to begin by considering new mutations, which are the most extreme forms of rare variant. Every individual carries ~60 new mutations in the parts of the genome that are accessible to current sequencing technologies (e.g. Kong et al. 2012). While most of the 3–4 million variants in any individual genome are shared with others and may be common in the population, the rare variants, including the new mutations, are less shared. Thus, as more and more individuals from a population are sequenced, the number of common variants discovered saturates, but the number of rare variants continues to increase. Consequently, most of the variants discovered by sequencing a large population sample are rare (The 1000 Genomes Project Consortium 2012) and the properties of this group are of considerable importance for many aspects of human genetics, as illustrated by the other chapters in this book. In this chapter, we consider how natural selection acts on rare variants, how such selection can be detected

Y. Xue (✉) • C. Tyler-Smith (✉)
The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
e-mail: ylx@sanger.ac.uk; cts@sanger.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_9

from the datasets that are available now or potentially in the future, its consequences for the patterns of rare variation in populations, and how it is possible to make use of these patterns to inform the functional interpretation of rare variants.

The Characteristics and Fate of New Mutations in a Population

New mutations occur at approximately random positions in the genome, and their characteristics in humans can be evaluated either by simulation or from the accumulating observations of such mutations. Their functional properties are most readily evaluated in protein-coding regions, where functional annotation is most highly developed. Coding mutations can be classified according to whether or not they change an amino acid, and if so, how deleterious the change may be to the protein (Adzhubei et al. 2010; Kumar et al. 2009). Such analyses suggest that over half of new mutations may be moderately or strongly deleterious (Boyko et al. 2008), as shown in the first column of Fig. 1. In contrast, negligible proportions of common variants with frequency >10 % fall into these deleterious categories (Fig. 1, columns 5–8). Corresponding functional characteristics of new mutations in non-coding regions have not been evaluated as thoroughly, but are also expected to be more deleterious than common non-coding variants.

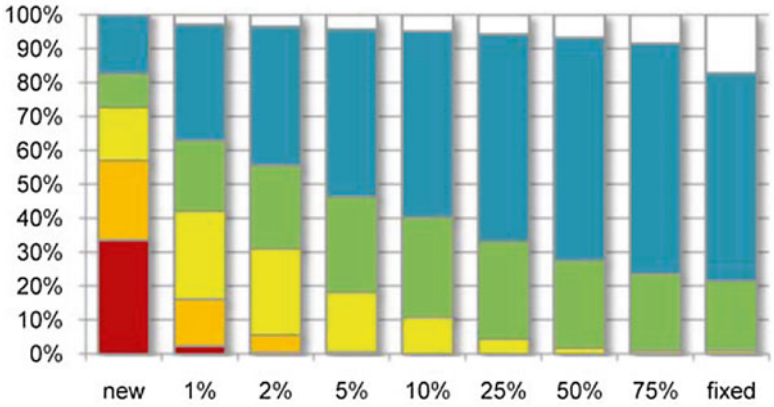

Fig. 1 Boyko et al.'s estimates of the deleterious properties of nonsynonymous variants in an African American sample stratified according to allele frequency. *Left*: new mutations; *middle*: SNPs at different derived allele frequencies, as indicated on the horizontal axis; *right*: human-chimpanzee fixed differences. Colours indicate strongly deleterious (*red*), moderately deleterious (*orange*), weakly deleterious (*yellow*), nearly neutral (*green*), neutral (*blue*), or positively selected (*white*) variants. Note the high proportions of strongly and moderately deleterious variants among new mutations and rare SNPs, and their near-absence from common SNPs and fixed differences. From (Boyko et al. 2008)

New mutations will often, just by chance, not be transmitted to offspring: their carrier may not have children, or if for example they have just one or two children, 50 % or 25 %, respectively, of their new mutations will on average not be passed on. Over many generations, the usual fate of new mutations and rare variants more generally is thus for them to be lost from the population. Some, however, will increase in frequency. This is the neutral process of genetic drift. In this chapter, we are interested in the rare variants that are observed in a population, and most of these have been transmitted, at least for a few generations.

Types of Selection

The forms of selection relevant here are positive (Darwinian) selection, in which an advantageous variant increases in frequency in the population and may eventually be fixed, and purifying (negative) selection, in which a disadvantageous variant decreases in frequency and may be eliminated. Although positive selection has undoubtedly influenced allele frequencies in human populations, there is debate about its prevalence, particularly in the form of classic selective sweeps (Hernandez et al. 2011; Colonna et al. 2014); furthermore, current methods for detecting positive selection have very little power when the selected variant is rare (Jobling et al. 2014). In contrast, purifying selection is widely accepted as ubiquitous. For these reasons, we concentrate here on purifying selection. The strength of purifying selection ranges along a continuum from, at one extreme, lethality, to, at the other extreme, negligible strength indistinguishable from neutrality. Consequently, moderately deleterious variants can survive in the population and are enriched among rare variants for two reasons. First, as described in the previous section, over half of new mutations (at least in protein-coding regions) may be deleterious, and second because a deleterious variant will increase in frequency more slowly than a neutral one, and thus remain rare for longer. Indeed, many currently rare variants could never become common because of the long-term effect of purifying selection. Some methods to detect purifying selection would also have low power for rare variants, but a number of approaches are available for detecting purifying selection at rare variants, and are considered in the next section.

Detecting Purifying Selection at Rare Variants

Detection by Variant Deficit

Variants that have a *dominant* lethal effect before birth may arise as new mutations, but would not be observed in the population, or, since penetrance of the phenotypic characteristics associated with a variant is often variable (Cooper et al. 2013), might occasionally be compatible with live birth but would be extremely rare.

They might also be enriched among spontaneous miscarriages and thus detected by a genomic comparison of miscarriages with healthy births. Alternatively, if a sufficiently large sample of new mutations in healthy individuals were examined, dominant lethal mutations would appear as positions with zero or reduced numbers of mutations. With approximately 60 new mutations per individual (Kong et al. 2012), the current population of about 7×10^9 people carries over 120 new mutations per nucleotide. Although beyond the scope of current sequencing technology and analysis, a deficit of new mutations at particular positions may become detectable by this approach in the future.

More commonly, a lethal effect may be *recessive*. Such variants can persist in the population in unaffected heterozygous carriers and can be detectable from a lack or decrease in the number of offspring homozygous for the variant. This could manifest as increased spontaneous miscarriages, or early mortality, as observed in a number of Mendelian conditions (Fig. 2). This figure illustrates the inheritance of Lethal Contractural Syndrome Type 3, a recessive condition characterized by severe multiple joint contractures affecting the limbs, together with severe muscle wasting and atrophy, caused in this family by inheriting two copies of a rare nonsynonymous variant Asp253Asn in *PIP5K1C* (Narkis et al. 2007). Affected individuals die of respiratory insufficiency minutes to hours after birth. An estimate of the number of autosomal recessive lethals in the well-characterized Hutterite population suggests about 0.6 per individual (Gao et al. 2014).

There is a second class of lethal variant to consider: those that are lethal in an evolutionary sense, in that they are not transmitted to offspring, but have less overt effects on the carrier. The most obvious members of this class are variants leading to infertility, which has complex genetic and non-genetic causes and overall affects around 10 % of couples. Schemes for population sampling, for example, that used by The 1000 Genomes Project, may require informed consent from non-vulnerable adults, but when these schemes sample random unrelated individuals (as contrasted with mother-father-child trios or larger families), some infertile individuals are expected to be included. Thus, such sequence datasets

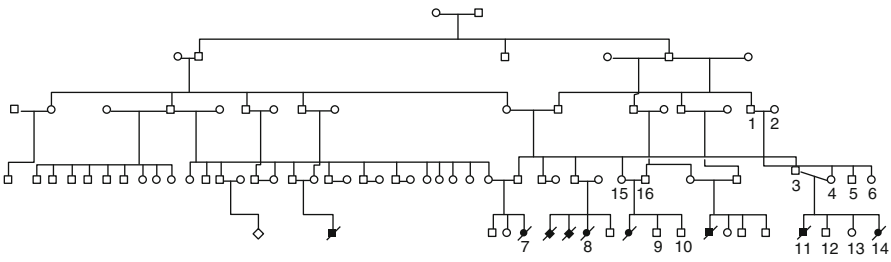

Fig. 2 Pedigree showing segregation of Lethal Contractural Syndrome Type 3 (LCCS3), caused by a mutation in *PIP5K1C*. *Open symbols*: unaffected individuals; *closed symbols*: affected individuals. Affected individuals (*circles*: females; *squares*: males) died within hours of birth, or (*diamonds*) were terminated after diagnosis of the condition during pregnancy. From (Narkis et al. 2007)

potentially include both heterozygous evolutionarily lethal dominant rare variants and homozygous evolutionarily lethal recessive rare variants.

At a population level, purifying selection acting on a recessive variant would lead to a departure from Hardy–Weinberg equilibrium at that variant, driven by a deficit in the number of homozygotes observed compared with the number predicted from the heterozygote frequency. In practice, such a departure would be difficult to detect, for two reasons. First, Hardy–Weinberg equilibrium is often used as a genotyping quality-control metric, and variants showing departures filtered out of datasets because most are due to technical errors in genotype calling. Second, the number of individuals required to detect a departure for a rare variant would be large, for example, 40,000 for a 1 % frequency variant and over four million for a 0.1 % variant. Nevertheless, this approach may become applicable in the future when large numbers of high-quality genome sequences are available.

Detection by Functional Annotation

Databases of known disease-causing variants, for example, the Human Gene Mutation Database (Stenson et al. 2014) which included 156,932 entries in mid-2014 (http://www.hgmd.org/) or ClinVar (Landrum et al. 2014) with 111,294 variants in mid-2014 (http://www.ncbi.nlm.nih.gov/clinvar/), provide one example of how *functional annotation* can identify rare variants that are likely to be under strong purifying selection. A survey in 2012 suggested that each apparently healthy individual in the general population carried around two severe disease alleles, defined those that were listed (after filtering) in the Human Gene Mutation Database at that time (Xue et al. 2012). Although such databases are steadily increasing in size, the majority of potential severe disease-causing variants have not yet come to the attention of medical geneticists, and thus these databases do not yet provide a comprehensive catalogue of severely deleterious variants.

Functional predictions, especially in protein-coding regions as described above, provide another approach to functional annotation (see also Chap. 5). The most obviously deleterious variants are those that lead to loss of function (LoF) of the protein. These include SNPs that introduce a stop codon or alter an essential splice site, indels that change the reading frame, and large deletions that remove much or all of a coding region. LoF variants are not all deleterious: some are common in the population and probably neutral (MacArthur et al. 2012) and a few are even advantageous and positively selected (e.g. Xue et al. 2006). Nevertheless, they provide a functional class enriched for variants that are likely to be acted on by purifying selection, and we will refer to applications of this principle below. Applying similar reasoning suggests that nonsynonymous and synonymous annotations identify two classes of variant that, on average, are subject to milder forms of purifying selection and are sometimes contrasted with intergenic variants. Among non-protein-coding variants, approaches to identifying those most likely to experience purifying selection are starting to be developed (e.g. Khurana et al. 2013).

Detection from Haplotype Structure

Haplotype structure can provide an additional source of information about likely purifying selection. If we consider rare variants with equal frequency in the population, for example, those observed twice in a particular sample as studied in Phase 1 of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012), the haplotype structures on which they lie will differ, on average, according to the action of purifying selection on them. Neutral rare variants will typically share the same recent origin and lie on a long shared haplotype (Fig. 3a). In contrast, evolutionarily lethal rare variants must represent independent new mutations and are therefore likely to lie on different haplotypes (Fig. 3c). The consequences of purifying selection on mildly deleterious variants are considered further in the next section; here, we note that they tend to be more recent in origin than neutral variants of the same frequency, so will lie on a longer shared haplotype which has accumulated fewer additional mutations (Fig. 3b).

In summary, strong purifying selection can be detected at individual highly deleterious rare variants. Weak purifying selection on mildly deleterious variants cannot be detected at the individual variant level, but can be detected on classes of variant.

Consequences of Purifying Selection at Rare Variants

Consequences for the Age of Rare Variants

The consequences of purifying selection on rare variants have been investigated using both population-genetic theory and experimental datasets. In 1974, Maruyama modelled the spread of a mutation using a diffusion approximation of a branching process, and predicted that a deleterious variant would tend to be younger

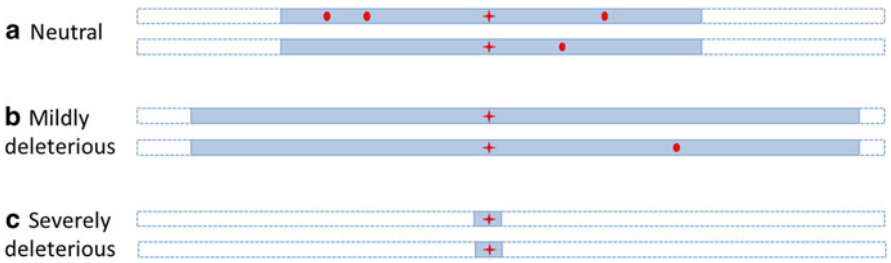

Fig. 3 Characteristics of haplotypes shared by two copies of a (**a**) neutral, (**b**) mildly deleterious or (**c**) severely deleterious rare variant. The rare variant is represented by the *red cross*, the shared haplotype by the *solid blue bar* and the non-shared haplotype by the *dotted open bar. Red ovals* represent additional mutations that have arisen on the shared haplotype after the origin of the rare variant. Compared with the neutral variant, the mildly deleterious variant has a more recent origin and consequently the shared haplotype is longer and carries fewer subsequent mutations. The severely deleterious variant has arisen independently by recurrent mutation and so lies on two different haplotypes

than a neutral variant of the same frequency (Maruyama 1974). In one sense, this conclusion is counter-intuitive, since purifying selection reduces the spread of a deleterious variant and might be expected to increase the time it takes to reach a certain frequency. The more appropriate intuition is that although a deleterious allele is less likely to reach this certain frequency than a neutral one, if it does so, it is likely that it took fewer steps.

This prediction has been tested in pilot data from the Genome of the Netherlands (GoNL) Project, which sequenced the genomes of 47 mother-father-child trios at 12× coverage and generated high-quality haplotype data via family-based phasing (Kiezun et al. 2013). The authors devised a Neighbourhood-based Clock (NC) statistic to capture information about the age of rare alleles. This statistic incorporates information about the physical distance to the nearest completely linked lower-frequency variant or nearest detectable recombination event, and a higher value corresponds to a younger age. NC values were indeed higher, and thus ages younger, for nonsynonymous rare variants and probably damaging nonsynonymous rare variants than for synonymous variants of the same frequency (Fig. 4a). In a more direct approach to estimating variant age, Mathieson and McVean analysed data from the 1000 Genomes Project Phase 1, which sequenced 1,092 individuals from 14 populations at 4–5× coverage (The 1000 Genomes Project Consortium 2012), focussing on variants called exactly twice (f_2 variants) (Mathieson and McVean 2014). The 1000 Genomes haplotypes were not phased as accurately as the GoNL samples, and the authors made a maximum likelihood estimate of the age based on an upper bound to the f_2 shared haplotype length and the number of singleton mutations that have arisen on these haplotypes. Median ages for f_2 variants shared within a population were 170–320 generations within Africa, 50–160 generations within Europe or Asia, contrasted with 320–670 generations for variants shared between Europe and Asia, and 1,000–2,300 generations for variants shared between Africa and Europe or Asia. Most relevant here is that median age differed between functional annotation classes for f_2 variants shared within a population: 58, 83, 112, and 125 generations for LoF, coding, functional noncoding, and unannotated variants, respectively, with all of these differences being highly significant (Fig. 4b) (Mathieson and McVean 2014). No such differences were seen for f_2 variants shared between continents. The authors suggest that purifying selection acts on these mildly deleterious variants (and most strongly on the LoF class) to gradually eliminate them from the population, but that those that have survived for long enough to be shared between continents represent the subset of each annotation class that is effectively neutral.

Consequences for the Geographical Distribution of Rare Variants

The consequence of purifying selection on mildly deleterious variants for their geographical distribution, as suggested above, is that they are more likely to be found in the same population than neutral variants of the same frequency. Purifying selection reduces their survival time and thus spread. For evolutionarily lethal variants,

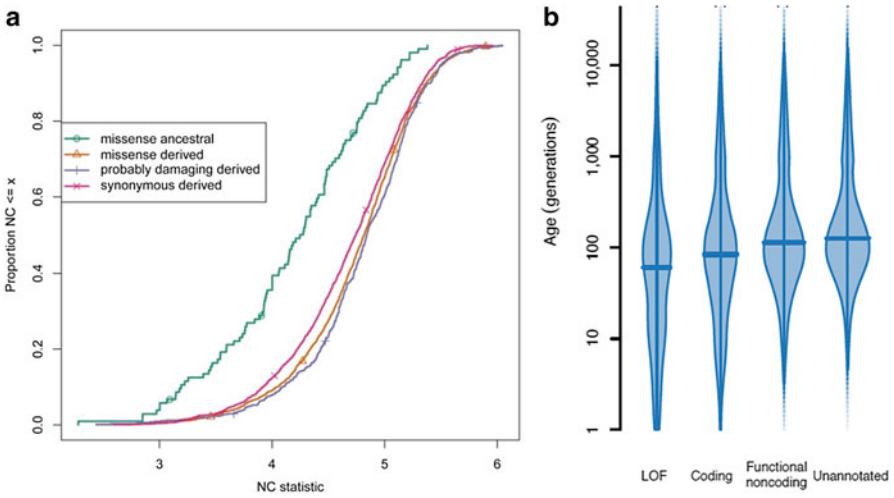

Fig. 4 Ages of damaging variants are younger than neutral variants of the same frequency. (**a**) Cumulative distribution of the NC statistic (higher value ≡ younger age) for alleles with a minor allele count of three in the Genome of the Netherlands pilot study. Taking synonymous variants (*pink*) as approximately neutral, nonsynonymous (*missense*) variants as a whole (*orange*) are shifted towards higher values and probably damaging nonsynonymous variants (*purple*) even more so. Thus, these damaging classes of variant are younger. From (Kiezun et al. 2013). (**b**) Ages (estimated in generations) of different classes of rare variant in the 1000 Genomes Project Phase 1 data. Horizontal bars show the median value for each class. Compared with unannotated variants, variants of the same frequency annotated as functional but non-coding, coding, or LoF are each significantly younger. From (Mathieson and McVean 2014)

the expectation is different. They do not spread, so each represents an independent mutation. In the same way that they are likely to occur on different haplotypes (Fig. 3), they are also likely to occur in different populations (Fig. 5).

Using Selection at Rare Variants to Inform Functional Interpretation of Rare Variants

In the previous sections, we have summarized insights into expectations and observations relevant to the action of selection on rare variants. In this final section, we ask whether these insights can be used to inform interpretation of rare variants whose function is unknown, but of interest. For example, in a sequencing study searching for variants with large effect leading to a rare severe genetic disease, many candidate causal variants are usually found, and it can be difficult to identify the true causal variant (MacArthur et al. 2014). Functional testing in a cell line or model organism currently provides a gold standard for establishing causality, but is impossible to apply on a large scale and not applicable to all phenotypes, so prioritization of the candidates is necessary. In this scenario, the causal variant can be assumed to

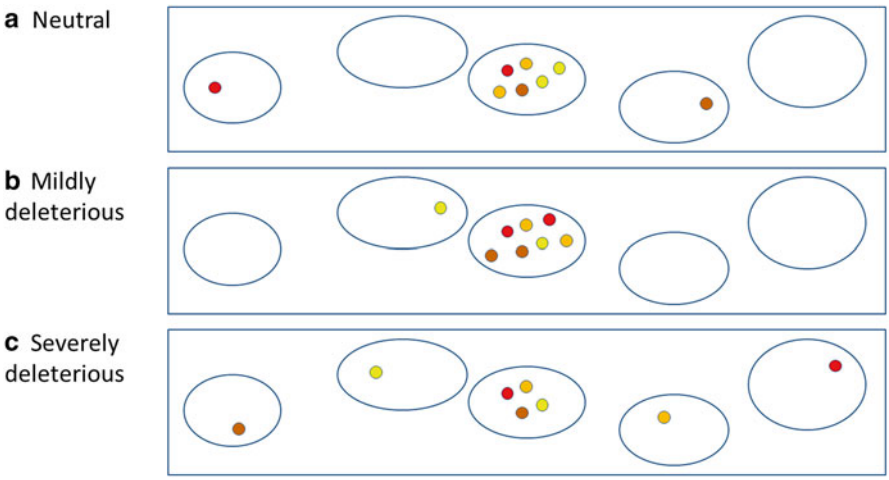

Fig. 5 Geographical distribution of two copies of a (**a**) neutral, (**b**) mildly deleterious, or (**c**) severely deleterious rare variant. In each panel, the two copies of the rare variant are shown as *two dots* with the same colour, and different populations by *ovals*. Four rare variants in the central population are considered. Compared with the distribution of neutral variants (**a**), the second copy of a mildly deleterious variant (**b**) is more likely to be found in the same population as the first, while the second copy of a severely deleterious variant (**c**) arises by an independent mutation and is therefore more likely to be found in a different population

be rare and have deleterious consequences, and a shortlist of rare variants with deleterious functional annotations can be drawn up. In prioritizing the entries on this list, prior information about the condition and its genetic basis is used if available. But if after taking into account all available information the shortlist still contains many candidates, the expectation that the causal variant will be subject to strong purifying selection, while many of the other candidates will not, may be considered. If several individuals with similar phenotypes have been sequenced:

- Assuming recessive inheritance, the causal variant should lie on a longer, less variable and more geographically focussed haplotype than other variants.
- Assuming dominant inheritance and evolutionary lethality, the causal variant should lie on independent haplotypes from independent geographical locations.

While the additional information provided is indirect and only applicable when the same causal variant is discovered in more than one individual, identifying the true causal variant in rare diseases can be so challenging that any additional insights that can be extracted from the data available are valuable. The consideration of the consequences of purifying selection offers additional insights.

Acknowledgement Our work is supported by The Wellcome Trust, grant 098051.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. doi:10.1038/nmeth0410-248

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4(5), e1000083. doi:10.1371/journal.pgen.1000083

Colonna V, Ayub Q, Chen Y, Pagani L, Luisi P, Pybus M, Garrison E, Xue Y, Tyler-Smith C (2014) Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. Genome Biol 15(6):R88. doi:10.1186/gb-2014-15-6-r88

Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132(10):1077–1130. doi:10.1007/s00439-013-1331-2

Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M. An estimate of the average number of recessive lethal mutations carried by humans. arXiv. 2014;14077518.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Genomes P, Sella G, Przeworski M (2011) Classic selective sweeps were rare in recent human evolution. Science 331(6019):920–924. doi:10.1126/science.1198878

Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C (2014) Human evolutionary genetics, 2nd edn. Garland Science, New York

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, The 1000 Genomes Project Consortium, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342(6154):1235587. doi:10.1126/science.1235587

Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, Boomsma DI, van Duijn CM, Slagboom PE, van Ommen GJ, Wijmenga C, Genome of the Netherlands Consortium, de Bakke PI, Sunyaev SR (2013) Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet 9(2), e1003301. doi:10.1371/journal.pgen.1003301

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. Nature 488(7412):471–475. doi:10.1038/nature11396

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(7):1073–1081. doi:10.1038/nprot.2009.86

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980–D985. doi:10.1093/nar/gkt1113

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Gallego Romero I, The 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C (2012) A

systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828. doi:10.1126/science.1215040

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG, Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508(7497):469–476. doi:10.1038/nature13127

Maruyama T (1974) The age of a rare mutant gene in a large population. Am J Hum Genet 26(6): 669–673

Mathieson I, McVean G (2014) Demography and the age of rare variants. PLoS Genet 10(8), e1004528. doi:10.1371/journal.pgen.1004528

Narkis G, Ofir R, Landau D, Manor E, Volokita M, Hershkowitz R, Elbedour K, Birk OS (2007) Lethal contractural syndrome type 3 (LCCS3) is caused by a mutation in PIP5K1C, which encodes PIPKI gamma of the phophatidylinsitol pathway. Am J Hum Genet 81(3):530–539. doi:10.1086/520771

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1–9. doi:10.1007/s00439-013-1358-4

The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E, Burton J, Leonard S, Rogers J, Tyler-Smith C (2006) Spread of an inactive form of Caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78(4):659–670. doi:10.1086/503116

Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, Mort M, Phillips AD, Shaw K, Stenson PD, Cooper DN, Tyler-Smith C, The 1000 Genomes Project Consortium (2012) Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am J Hum Genet 91(6):1022–1032. doi:10.1016/j.ajhg.2012.10.015

Collapsing Approaches for the Association Analysis of Rare Variants

Jennifer L. Asimit and Andrew Morris

In testing for associations with rare variants, alternative methods to those used for common SNPs are required due to their lack of power at lower frequency. Collapsing approaches overcome this power loss by testing for an association with an aggregate of rare variants. These tests pool information across the rare variants such that a single test is performed on the summary statistic, and are powerful tools, provided that certain conditions are satisfied. In this chapter, the general framework of collapsing methods is explored, including optimal conditions for attaining high power. Comparisons are made between specific collapsing methods, as well as data-adaptive versions that have been developed to recover much of the power loss from nonideal settings.

Introduction

An abundance of powerful single variant association tests have been developed and employed in genome-wide association studies to successfully identify disease-associated SNPs. Although many SNPs have been identified as having an effect on disease susceptibility, they only account for a small proportion of heritability. This has been part of the motivation in searching for disease associations with variants of lower minor allele frequency (MAF). Variants with MAF < 0.01 are often referred to as rare variants, while low-frequency variants are those with $0.01 <$ MAF < 0.05.

J.L. Asimit (✉)
Wellcome Trust Sanger Institute, Hinxton, UK
e-mail: ja11@sanger.ac.uk

A. Morris
Department of Biostatistics, University of Liverpool, Liverpool, UK
e-mail: A.P.Morris@liverpool.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_10

Due to the few observations of the rare minor allele at a specific variant, there are few causal rare variants that are present in many individuals. Consequently, these tests experience a dramatic reduction in power to test for an association with a single low-frequency/rare variant. This has been demonstrated in simulation studies by Li and Leal (2008). Single variant tests experience a further reduction in power in the presence of allelic heterogeneity, since different individuals then contribute to an association signal at different variants in the locus. A locus-based approach, where SNPs in the locus are collectively tested for an association, is an alternative to single-variant tests, and is often employed when allelic heterogeneity is an issue.

Multi-marker tests combine information across the variants in a locus, and simultaneously test the multiple variants for an association, using multivariate methods. This approach has higher power than single-variant tests when there are multiple SNPs of moderate effect within the locus, but has the caveat of requiring multiple degrees of freedom, which lowers the power of the test. In addition, in simulation studies, it has been demonstrated that the power is further reduced as the MAF decreases and as the number of rare causal variants increases (Li and Leal 2008). On the contrary, there are numerous rare variants collectively, and a powerful alternative is to test for an association with their aggregate. That is, the information on rare variants within the locus is pooled into a single summary statistic for a "super locus," which is used in a univariate test for an association of rare minor alleles with a trait. This combining of information across multiple rare variant sites results in fewer degrees of freedom, and under certain conditions, increased power.

The low MAFs of rare/low-frequency variants make them challenging to access, and until recently, the majority of genotype arrays had been designed with a focus on common variants. Lower MAF variants may now be accessed via high-density genotyping arrays (e.g., Illumina Omni 2.5 M), but they remain underpowered to detect rare variant effects. Sequencing is the gold standard for accessing rare variants, being the most accurate method for obtaining genotypes. However, at the moment, sequencing has a high cost, making it not easily available for large samples. An approach to overcome these issues is to make use of a high-density reference panel, such as the 1000 Genomes Project data (The 1000 Genomes Project Consortium 2010), which is composed of sequence data from 1,092 individuals of various ancestries and enables access to variants with MAF as low as 0.01 across the genome, as well as MAFs of 0.001–0.005 in gene regions. Many more rare/low-frequency variants are expected to be identified within the data of the ongoing UK10K Project (http://www.uk10k.org/), for which 10,000 individuals (primarily from the UK) have been sequenced; the exomes of 6,000 cases for various disorders are high-depth sequenced with the aim of identifying associated rare/low-frequency variants, while 4,000 individuals from population-based cohorts with deep phenotype data are genome-wide sequenced at low-depth (average 6×), which may be useful as a reference panel for other studies. Such high-density reference panels may be used to select variants for genotyping, as many of the variants within them would not have been included on genotyping arrays due to their MAF. Alternatively, a cost-effective use of such reference panels is to employ them in imputation. Genotypes that are not directly typed may be estimated by taking advantage of the

genetic correlations within the reference panel and extrapolating to the study sample. Several review papers on rare variants and collapsing methods are available, namely Asimit and Zeggini (2010), Bansal et al. (2010), and Dering et al. (2011), and in the next section, we discuss such methods in a general framework, including choices of MAF and/or functional annotation filtering. Extensions of collapsing methods to imputed data are discussed in the subsequent session. In the final section, the considerations required prior to implementing any of the burden tests are discussed, as well as the use of family-based designs for detecting rare variant associations.

Methods

In the implementation of collapsing methods, an MAF threshold is required, such that all variants with MAF below the threshold are aggregated for association testing. For many of these approaches, a fixed-allele frequency threshold is imposed, where the threshold is selected based on the specifications of the study. The outcome of the test is highly dependent on the choice of threshold, and some caution is required in the selection of a threshold, as a balance needs to be achieved between retaining casual rare/low-frequency variants and reducing the number of non-causal variants within the aggregate. A relatively high MAF threshold allows the inclusion of more candidate causal variants, but this lenient threshold has the cost of permitting a larger number of non-causal variants in the aggregation. In turn, non-causal variants contribute neutral effects and may have a detrimental effect on the power to detect an association with the "super locus." In simulation studies (Li and Leal 2008), it has been demonstrated that the power of collapsing methods decreases as the MAF of the non-causal variant increases (from 0.02 to 0.05), and further power reductions accrue with the inclusion of an additional non-causal variant of the same MAF. Rather than restricting inclusion according to a particular MAF threshold, variants may be weighted according to an inverse relationship with allele frequency (Madsen and Browning 2009).

In addition to selection of an MAF threshold, further filters may be applied to minimize the incorporation of nonfunctional variants in the analysis. Variants may be classified according to their degree of predicted functionality such that those predicted to be neutral are excluded, or a weighting scheme may be implemented such that variants with low functionality prediction scores are down-weighted when collapsing the variants. Bioinformatics tools such as Polymorphism Phenotyping-2 (PolyPhen-2) (Adzhubei et al. 2010), which is a development of PolyPhen (Ramensky et al. 2002), and Sorting Intolerant From Tolerant (SIFT) (Ng and Henikoff 2003) may be employed to predict the potential functionality of non-synonymous-coding variants. Information for these two tools has been combined into a functional annotation score, Combined Annotation scoRing toOL (CAROL) (Lopes et al. 2012), which has higher predictive power and accuracy than either of the individual tools alone. However, despite coding variants being more likely to be functional, only a small proportion of genome variation is attributed to coding

variants, and there is increasing evidence that non-coding variants are associated with complex traits. In particular, non-coding variants have been verified to be associated with disorders such as Hirschsprung disease (Emison et al. 2005), asthma (Haller et al. 2009), cleft palette, and ankyloglossia (Pauws et al. 2009). In order to overcome the coding variant limitation of the previously mentioned bioinformatics tools, Genome-Wide Annotation VAriant (GWAVA) was developed (Ritchie et al. 2014). GWAVA is a tool that predicts functionality for both coding and non-coding variants and is based on a classifier trained to differentiate between variants known to be disease implicated and those that are known to not have a role in disease etiology.

Various forms of collapsing approaches have been proposed, and there are several properties that are shared among these methods. Firstly, all such burden tests may be expressed within a general framework as follows

$$G\left(y_i\right) = \alpha + \beta \sum_j w_j x_{ij}, \tag{1}$$

where x_{ij} is the coded genotype of individual i at SNP j, w_j is the weight given to SNP j, and the summation runs through all variants j within the region. In having a single shared effect estimate β for all of the rare variants, rather than individual effect estimates β_j for each variant, there are fewer degrees of freedom and hence, a power gain compared to single-variant analysis. The coding of the genotypes depends on the specific method, while the weights depend on both the method and filtering of the variants according to MAF threshold and functional annotation. The weighted sum of coded genotypes may be regarded as a genetic score for individual i (Madsen and Browning 2009). Genotype coding typically takes on values 0/1 for the absence/presence of the minor allele at the variant or is the count of minor alleles carried by the variant, $x_{ij} \in \{0,1,2\}$, which implicitly assumes a common direction of effect among the variants. It is thus apparent that a combination of deleterious and protective variants will result in a dilution effect upon aggregation, as both risk-increasing and risk-decreasing variants share a common effect estimate. Various simulation studies involving collapsing methods have demonstrated this limitation of burden tests (Han and Pan 2010; Asimit et al. 2012; Ladouceur et al. 2012). In this section, particular burden tests will be discussed, including how they fit into the general modelling framework described above.

The regression framework burden tests of Morris and Zeggini (2010) consider two forms of collapsing across the variants below a predetermined MAF threshold. Although, their RVT1 and RVT2 methods are presented in a linear regression form for quantitative traits, they are extended with ease to case–control studies by considering a logistic regression. In the RVT1 model, the locus information for each individual is collapsed down into the proportion of rare/low-frequency variants at which there is at least one minor allele. This coincides with model (1), where the coded genotype x_{ij} takes on the value 1 when individual i carries at least one minor allele at variant j and variant j is below the MAF threshold, and 0, otherwise. The weight w_j of each x_{ij} is $1/n_i$, where n_i is the number of successfully genotyped rare variants

for individual i. Thus, for model RVT1, β is the expected increase of the phenotype for an individual with a minor allele at each rare variant in comparison to one with none. Analysis of deviance is used to compare the maximized likelihoods of the null ($\beta = 0$) and unconstrained β models in the construction of likelihood ratio tests of disease association with an accumulation of rare variants.

The RVT2 model differs from RVT1 by assigning unit weight to the first coded genotype that takes the value 1 at a rare variant, and weight 0 otherwise; effectively the phenotype is modelled as a function of the indicator that at least one rare minor allele is carried by an individual. Morris and Zeggini (2010) demonstrate, via simulation studies, that the test based on the presence/absence of any rare minor allele (RVT2) is less robust to the presence of minor alleles at non-causal rare variants than the test based on proportions (RVT1). The proportions-based method has the potential caveat of being adversely affected by the presence of linkage disequilibrium (LD) among the aggregated variants. However, low-frequency variants are rarely found to be in strong LD with each other (Pritchard 2001). As both models are in a regression framework, a vector of covariate measurements z_i for individuals is easily incorporated to account for nongenetic risk factors or population structure.

The Cohort Allelic Sums Test (CAST) (Morgenthaler and Thilly 2007), Combined Multivariate and Collapsing (CMC) method (Li and Leal 2008), and Weighted Sum Statistic (WSS) (Madsen and Browning 2009) are collapsing methods developed specifically for case–control data. CAST (Morgenthaler and Thilly 2007) tests for a difference in the number of individuals with at least one mutation, between cases and controls, and does not filter on variant frequency; the coding and weights are the same as for RVT2 of Morris and Zeggini (2010) with the exception that common variants are included. Testing is performed via a 2×2 contingency table of case–control status and the presence/absence of at least one mutation. This can be regarded as a logistic regression of case–control status against the presence/absence of at least one mutation. Due to the absence of MAF filtering, if there are many common mutations within a region, the majority of individuals may carry a mutation. In turn, this will adversely affect the power to detect a difference in the number of individuals with at least one mutation. The CMC method (Li and Leal 2008) overcomes this issue by introducing an MAF threshold and group variants below the threshold together, while treating each variant of higher frequency as a group containing the single variant. They also consider a collapsing method COLL in which only variants below the MAF threshold are aggregated together in a single group, and common variants are excluded from analysis.

Implementation of the CMC method involves collapsing each group of multiple variants to an indicator variable for the presence/absence of any rare allele; groups consisting of a single variant do not require collapsing. The null hypothesis that none of these groups are disease associated is then tested by using a multivariate test, such as Hotelling's T^2 or logistic regression. The logistic regression form fits the framework of model (1) by letting G(•) be the logit function for the probability that $Y_i = 1$ (individual i is a case) and following the setup of RVT2, allowing multiple summations for each group of variant(s). Li and Leal (2008) demonstrate via simulation studies that the CMC method combines the strengths of collapsing methods

(high power for rare variant analyses) and multivariate methods (robust against inclusion of non-causal variants).

In the WSS approach, each variant is weighted such that those of lower frequency are up-weighted, following the assumption that rare variants exhibit stronger effects than common ones (Madsen and Browning 2009). Genotype code x_{ij} is based on the number of minor alleles at the variant. In addition, the assigned weight for each x_{ij} is inversely proportional to the estimated standard deviation of the total number of mutations across both cases and controls, under the assumption of the null hypothesis that the mutation frequencies are not associated with disease status. That is, letting n_i be the number of cases and controls, each variant is given weight $w = \left[n_j q_j \left(1 - q_j\right) \right]^{-1/2}$, where

$$q_j = \frac{\sum_{j=1}^{n_i^0} x_{ij}^0 + 1}{2n_i^0 + 2},$$

n_i^0 is the number of controls, and x_{ij}^0 the number of minor alleles at variant j for control subject i. The proportion q_j is an adjusted estimate of the MAF in controls, which includes correction factors to avoid numerical problems from zero estimates, which may occur if the minor allele only appears in cases. The allele frequency is based on controls rather than the entire sample to circumvent deflation of a true signal due to an excess of minor alleles in cases. This weighting scheme assigns larger weights to very rare variants and Price et al. (2010) have shown that this scheme implicitly assumes that the log odds ratio is approximately inversely proportional to the square root of the allele frequency. The individual-specific genetic scores from the entire sample are then ranked together and the test statistic is formed from the sum of the ranks from case subjects. Re-sampling procedures are then used to obtain a p-value for the association test.

When all SNPs are causal and have the same direction of effect, the WSS approach has been demonstrated to have higher power than the CMC method. However, with the addition of non-causal low-frequency SNPs (MAF 0.02 or 0.05) the CMC test achieves a higher power (Han and Pan 2010). When both protective and deleterious variants are present, the WSS approach has a reduction in power, irrespective of nonfunctional variant inclusion, while the CMC test is able to achieve a higher power (Han and Pan 2010). The impact of different effect directions on the power of the CMC approach is likely not as extreme as for the WSS approach because variants in the CMC test are collapsed into an indicator variable for the presence/absence of a rare minor allele. In contrast, each variant contributes to the WSS test statistic.

A further pooling statistic is the Cumulative Minor-Allele Test (CMAT) (Zawistowski et al. 2010), which requires selection of an MAF threshold and is computationally efficient. It makes use of the same genotype coding as the WSS method, which is based on the number of minor alleles at the variant. The weight function for each variant is flexible and may be chosen according to any underly-

ing assumptions for the study. Specific weight functions considered in the simulation study of Zawistowski et al. (2010) are the weight function of Madsen and Browning (2009), as well as a simple weighting scheme that is an indicator function based on the filtering criteria for the variants. An example of filtering criteria includes a fixed MAF threshold and annotation: $w_j = I\{\text{MAF}_j \leq 0.05$ and SNP_j annotated as missense, nonsense, or splice-site mutation or an untranslated region$\}$, where MAF_j is the MAF of variant j, SNP_j, and $1\{E\}$ is the indicator function, taking on value 1 when the event E occurs, and 0, otherwise. The genetic scores of (1) may then be viewed as the weighted minor allele counts for cases and controls, and the weighted major allele counts may be obtained by replacing the x_{ij} by $2-x_{ij}$,. These weighted allele counts for cases and for controls may then be used to test for independence with disease status in the form of a contingency table, with the exception that permutation is required to assess significance of the χ^2-like statistic. That is, the CMAT statistic takes the form of a χ^2, but does not follow a χ^2 distribution, since there is dependency among the allele counts due to LD. Similar to CAST (Morgenthaler and Thilly 2007), this procedure may be regarded as a logistic regression and fit in the general burden test framework (1). In particular, CMAT may be expressed as a logistic regression of case–control status against weighted minor allele counts, $\text{logit}\left(p_{ij}\right) = \alpha + \sum_j w_j x_{ij} s$, although permutations are required for evaluating significance.

Under various settings of probabilities for the incorrect inclusion of non-causal variants and the correct inclusion of causal variants, comparisons of CMAT, WSS (Madsen and Browning 2009), and the collapsing method COLL of Li and Leal (2008) reveal that CMAT and WSS have nearly identical power performance and are the most powerful at all misspecification levels considered. COLL was able to attain a similarly high power when the proportion of neutral variants included in analysis was below 2 %, but the power loss increases with this proportion (Zawistowski et al. 2010).

As burden tests experience a loss in power in the absence of a homogenous direction of effect, knowledge of the risk alleles is a requirement that would alleviate this limitation. This is addressed by the data-adaptive sum test (Han and Pan 2010), which involves two stages of association testing and employs unit weighting across all variants. First, single-variant tests are conducted at each site, regardless of MAF. The marginal direction of effect may then be estimated at each variant, based on a $\{0,1,2\}$ additive genotype-coding x_{ij}. Sites with negative effect estimates and association p-values below a prespecified threshold are then reversely coded such that the effect estimate becomes positive, i.e., x_{ij} is re-coded as $2-x_{ij}$ at such variants. A common effect for the group of SNPs may then be tested for association in a logistic regression setting with unit weights in the framework of model (1). Alternatively, to allow different effects (magnitude and direction) for rare and common variants, two groups may be formed according to low/high MAF, and a separate effect estimate may be obtained for each group. As all common variants are aggregated into a single group, this differs from the CMC approach, which includes multiple single common variant groupings. The single group of common variants has fewer degrees of freedom than the CMC approach, which may result in an

increase in power. However, this advantage will only be noticeable if the assumption of the same direction of effect is appropriate for the set of common variants. Implementing unit weights, model (1) is then fit in a logistic regression framework with two regression coefficients, coinciding with each variant grouping:

$$\text{logit}\left(\Pr\left\{Y_i = 1\right\}\right) = \alpha + \beta_r \sum_{j \in R} x_{ij}^* + \beta_c \sum_{j \in C} x_{ij}^*,$$

where R is the set of low-frequency/rare variants, C is the set of common variants, and x_{ij}^* is the data-adaptive genotype code for variant j of individual i. Due to the data-adaptive coding of the variants, permutation is required to assess significance for both data-adaptive approaches. In general, the approach in which variants are partitioned into groups according to MAF achieves higher power than the single aggregation. In the scenario where non-causal rare variants (MAF sum 0.01 or 0.05) are present, the power of the data-adaptive sum tests surpasses the CMC test, whether or not direction effects differ among the causal variants. However, in the ideal case of a common effect direction, and the addition of non-causal low-frequency variants (MAF 0.02 or 0.05), the CMC test occasionally has a power improvement over the data-adaptive sum tests (Han and Pan 2010).

Rather than limiting a collapsing analysis to MAF filtering based on a fixed threshold, as with most burden tests, a variable threshold (VT) collapsing approach has been introduced (Price et al. 2010). This flexible approach is motivated by the thought that there exists some optimal MAF threshold for which variants with MAF below the threshold are considerably more likely to play a functional role than those with MAF above it. Across several plausible MAF thresholds a test statistic is calculated, which employs 0/1 weights according to MAF above/below the specific threshold and variants are coded according to the allele count in cases. In particular, model (1) is fit with phenotype regressed against mutation counts satisfying the MAF threshold. In a variant of the VT approach, VTP, each rare variant (MAF < 0.01) is weighted according to its posterior probability of being functional, as inferred from its PolyPhen-2 (Adzhubei et al. 2010) probabilistic score. These scores are only considered when MAF < 0.01 since PolyPhen-2 predictions of functional effect are most effective at rare variants (Price et al. 2010). In doing so, signals from low-frequency and common variants are not excluded, and the test is not at risk of losing power from mis-prediction at low-frequency variants. In simulation studies of both binary and quantitative traits having causal variants with the same effect direction, at significance level 0.05, the VTP approach attains the highest power, followed by VT. Lower powers are obtained by fixed threshold approaches (MAF threshold 0.01 or 0.05) and a weighted approach that uses weights similar to WSS, $w_j = 1/\sqrt{p_j\left(1-p_j\right)}$, where p_j is the MAF of variant j. These two methods have very similar performance in terms of power.

A summary of the assumptions for each burden test is provided in Table 1. The data-adaptive approach (Han and Pan 2010) is the only collapsing approach that is able to overcome the caveat of power loss when both risk-increasing and

Table 1 Summary of burden tests and their assumptions

Burden test	Assumptions
CAST (Morgenthaler and Thilly 2007)	Case–control
	Same direction of effect
	Unit weights
	Few common mutations
CMC (Li and Leal 2008)	Case–control
	Same direction of effect
	Unit weights
WSS (Madsen and Browning 2009)	Case–control
	Same direction of effect
	Rare variants stronger effect than common
RVT1; RVT2 (Morris and Zeggini 2010)	Same direction of effect
CMAT (Zawistowski et al. 2010)	Case–control
	Same direction of effect
	Unit weights
VT (Price et al. 2010)	Same direction of effect
	Exists optimal MAF threshold T such that variants with MAF$<T$ are more likely to be functional
Data-adaptive (Han and Pan 2010)	Unit weights

risk-decreasing variants are present in the analysis region. Considering the numerous simulation studies involving different comparisons of collapsing methods under various scenarios, the VT approach appears to be one of the most powerful in the ideal setting of common effect direction and irrespective of the presence of neutral variants (Ladouceur et al. 2012).

Extended Methods

Several burden tests have been extended to account for genotype uncertainty due to imputation, as well as variant quality. In general, genotype probability calls from imputation may be used to obtain an expected genotype call, $\sum_{x=0}^{2} x \Pr\left(x_{ij} = x\right)$, referred to as the dosage. These probabilistic genotype calls have been incorporated in CMAT (Zawistowski et al. 2010) and GRANVIL (http://www.well.ox.ac.uk/GRANVIL), which implements and extends the RVT1 burden test of Morris and Zeggini (2010). The latter test has also been extended in the Accumulation of Rare variants Integrated and Extended Locus-specific test (ARIEL), to incorporate variant quality scores, such that variants of lower quality are appropriately down-weighted (Asimit et al. 2012).

Simulations were used by Mägi et al. (2012) to compare the power of GRANVIL under different strategies for assaying rare genetic variation: (1) re-sequencing of all samples; (2) imputation of GWAS data to a high-density reference panel; (3) directly

genotyping all variants present on the reference panel; (4) genome-wide association study (GWAS) data. Although the gold standard re-sequencing approach attains the highest power, it is only slightly better than the imputation approach with an appropriate reference panel. This suggests that the application of extended collapsing methods to imputed data has the potential to detect rare variant associations, without the high costs of sequencing.

Power is evaluated under various settings that dictate the spectrum of causal variants and for three different sizes of reference panel (ascertained from the same population as the cohort sample): 120, 500, and 4,000. A panel of 120 individuals is equivalent to the CEU sample of the 1000 Genomes Project, while size 500 coincides with the European samples (CEU, FIN, IBS, TSI, GBR) of the 1000 Genomes Project, and 4,000 individuals corresponds to what will be available from the UK10K Project. Causal variants are randomly selected based on two parameters: the maximum MAF of any causal variant and the maximum total MAF of the causal variants in aggregate.

As anticipated, the burden test applied to GWAS data alone consistently performs poorly to detect rare variant associations and has the lowest power among the four approaches to assaying rare genetic variation, as very few rare variants within the region are typed directly. As the reference panel size increases, the power based on genotyping all variants on the reference panel approaches that of sequencing and the power based on imputation approaches that of genotyping. Moreover, imputation leads to substantial power gains over GWAS data alone.

In the scenario in which there are only very rare causal variants (MAF<0.005) and the total MAF of the causal variants is only 0.02, there is a general reduction in power for all approaches, in comparison to higher frequency causal rare variants (MAF<0.01 and maximum total MAF 0.05). There is a general loss of power in the rarer causal variants scenario, since this restriction implies a higher proportion of non-causal rare variants, which results in lower power for the burden test, irrespective of the approach to access rare variants for association testing. Power differences between the imputation and genotyping approaches are larger in the rarer causal variants scenario since the distribution of causal allele frequencies is more skewed to the rarest variants, which are anticipated to be most difficult to impute, irrespective of the reference panel size.

Results of the simulation study suggest that when there are many very rare causal variants (e.g., maximum causal variant MAF 0.005), the relative power of imputation and directly genotyping all variants present on the reference panel are both sensitive to the number of individuals composing the reference panel. Genotyping all variants in the reference panel results in a slight power reduction relative to sequencing, with smallest differences when a 4,000-individual reference panel is employed. As demonstrated by Mägi et al. (2012), such a large reference panel is expected to capture most of the rare variation in the study sample, and hence the power approaches that of sequencing. For a reference panel of size 500, this direct genotyping approach has noticeable reductions from sequencing in the scenario for which causal variants are at the very low end of the frequency spectrum (e.g., maximum individual MAF 0.005 and total MAF 0.02 for causal variants), but negligible differences when the maximum MAF is increased to 1 %. That is, the power

gained in using a reference panel of 500 in comparison to size 4,000 is larger in the rarer causal variants scenario.

Identified Associations

The burden test implemented in GRANVIL was used to re-assess the evidence for association with rare variants for 14,000 cases of seven complex diseases and 3,000 shared controls from The Wellcome Trust Case Control Consortium (2007), imputed using the "all ancestries" reference panel from the 1000 Genomes Project Phase 1 reference panel (June 2011 interim release) (Mägi et al. 2012). An accumulation of rare variants (MAF < 0.01) in *PRDM10* was identified as having genome-wide significant evidence of an association with a decreased risk of coronary artery disease. In addition, genome-wide rare variant associations with type 1 diabetes were identified in ten genes (nine risk-increasing, one protective) within the MHC.

Rare variant associations have also been identified in the application of a burden test in a whole-genome sequence-based analysis of high-density lipoprotein cholesterol (Morrison et al. 2013). The burden test in this analysis is a variant of the RVT1 test of Morris and Zeggini (2010), such that the counts of the number of rare minor alleles (MAF < 0.01) for each individual are input in the linear regression model, rather than the proportion. Various regional units of analysis are tested, including annotated regulatory units and sliding windows of 4 kb. The most significant burden test result is from an application to sliding windows, for which the Bonferroni significance threshold is 3.73×10^{-8} based on the number of sliding windows, resulting in the identification of a statistically significant region on chromosome 4, near *PARM1* (p-value 2.69×10^{-8}).

Genes that were identified in a GWAS as associated with hypertriglyceridemia (HTG) were re-sequenced to gain insight on any role of rare variants in this complex trait (Johansen et al. 2010). A simple Fisher's exact-type burden test revealed that, across the four HTG-associated genes (APOA5, GCKR, LPL, APOB), there is evidence of an excess of rare variants in individuals with HTG (p-value 6.2×10^{-8}). Moreover, the proportion of variation contributing to HTG was increased by incrementally considering rare variants in these genes, illustrating that both common and rare variants share a role in susceptibility to HTG. A similar approach was applied to a candidate gene (and ten of its protein interaction partners) for schizophrenia and related neuropsychiatric disorders in the investigation of susceptibility in an isolated northern Swedish population (Moens et al. 2011). Patients were found to harbor an increased burden of non-synonymous rare variants (p-value 0.018) and the significance was even more so when the cases were restricted to the early onset patients (p-value 0.0004).

In a GWAS candidate gene rare variant burden analysis for various diseases, involving the collapsing to the presence/absence indicators of variants with MAF < 0.005 and predicted to be functionally damaging, a noteworthy association was identified for multiple sclerosis (Nelson et al. 2012). Upon adjusting for the

number of genes tested, *IL6* attained p-value 7.1×10^{-3} for an association of functionally damaging rare variants with multiple sclerosis.

Discussion

Prior to implementing rare variant burden tests, several decisions are required. As the unit of analysis is at the region level, consideration of region definition is required for all methods. Region definition for burden test analyses is often based on functional units, such as genes. Alternatively, sliding windows of a fixed length or fixed number of rare/low-frequency variants may be tested for an association. The majority of collapsing approaches rely on the choice of an MAF threshold for categorizing variants to be aggregated together, and several incorporate information from functional annotation, which relies on choosing an appropriate bioinformatics tool.

Selection of a test for rare variant associations is heavily dependent on the underlying genetic architecture, and burden tests have been shown to be powerful, provided that certain assumptions are met. In particular, an implicit assumption of burden tests is that all (weighted) rare variants have the same magnitude and direction of effect on the phenotype, and there is a degree of power loss when this assumption is not satisfied. Collapsing approaches are able to achieve high power when the majority of the variants are either protective or deleterious, but are prone to extreme loss of power when both directions of effect are present in the region of interest. That is, collapsing will enrich the signal when all of the rare variants have the same affect on disease risk, since a single shared effect is estimated for all of the rare variants in the region. Conversely, by estimating a single shared effect, the signal will be weakened if variants that increase disease risk are collapsed with those that decrease disease risk. The presence of non-causal rare/low-frequency variants also has a detrimental impact on the power to detect an association, as neutral variants have a null effect, which dilutes the signals of causal variants.

Collapsing approach rare variant methods have been designed for unrelated individuals. However, power to detect rare variant associations with quantitative traits may be improved by considering family-based designs, which are more likely to be enriched for rare variants; in families for which a minor allele is carried by one parent, half of the children are expected to possess it as well (Shi and Rao 2011). Powerful adaptive-weighted rare variant association tests that are robust to different directions and magnitudes of effect within the locus, as well as population stratification, have been proposed by Fang et al. (2012) in a non-collapsing framework. Powerful non-burden tests have also been developed as a means of overcoming the burden test limitations of power loss due to different effect directions and/or inclusion of nonfunctional variants.

It is not clear if rare variants in the same gene will have the same direction of effect on disease, as the genetic architecture of complex traits is not well understood. It follows that under some disease models, burden tests will still provide a powerful approach to detect rare variant associations. As a means of maximizing power to detect these associations, burden tests may be applied in parallel with tests that allow different directions of effect.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Asimit JL, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Asimit JL, Day-Williams AG, Morris AP, Zeggini E (2012) ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73:84–94

Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773–785

Dering C, Hemmelmann C, Pugh E, Ziegler A (2011) Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol 35:S12–S17

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434:857–863

Fang S, Sha Q, Zhang S (2012) Two adaptive weighting methods to test for rare variant associations in family-based designs. Genet Epidemiol 36:499–507

Haller G, Torgerson DG, Ober C, Thompson EE (2009) Sequencing the IL4 locus in African Americans implicates rare noncoding variants in asthma susceptibility. J Allergy Clin Immunol 124:1204–1209, e9

Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70:42–54

Johansen CT, Wang J, Lanktree MB et al (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42:684–687

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB (2012) The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. PLoS Genet 8, e1002496

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Lopes MC, Joyce C, Ritchie GR, John SL, Cunningham F, Asimit J, Zeggini E (2012) A combined functional annotation score for non-synonymous variants. Hum Hered 73:47–51

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2), e1000384. doi:10.1371/journal.pgen.1000384

Mägi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP (2012) Genome-wide association analysis of imputed rare variants: application to seven common complex diseases. Genet Epidemiol 36:785–796

Moens LN, De Rijk P, Reumers J, Van Den Bossche MJA, Glassee W et al (2011) Sequencing of DISC1 pathway genes reveals increased burden of rare missense variants in schizophrenia patients from a northern Swedish population. PLoS One 6(8), e23450

Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615:28–56

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193

Morrison AC, Voorman A, Johnson AD, Liu X, Yu J et al (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45:899–901

Nelson MR, Wegmann D, Ehm MG, Kessner D et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in14,002 people. Science 337:100–104

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–4

Pauws E, Moore GE, Stanier P (2009) A functional haplotype variant in the TBX22 promoter is associated with cleft palate and ankyloglossia. J Med Genet 46:555–561

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137

Ramensky V, Bork P, Sunyaev S (2002) Human nonsynonymous SNPs: Server and survey. Nucleic Acids Res 30:3894–3900

Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of non-coding sequence variants. Nat Methods 11:294–296

Shi G, Rao D (2011) Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genet Epidemiol 35:572–579

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and3,000 shared controls. Nature 447:661–678

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S (2010) Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617

Rare Variant Association Analysis: Beyond Collapsing Approaches

Han Chen and Josée Dupuis

Introduction

Because most studies do not have sufficient power to detect association with rare single nucleotide variants (SNVs), a number of approaches to jointly analyze SNVs have been proposed. The earlier approaches consisted of simply counting the number of rare alleles within a gene or pathway carried by each participant, and evaluating whether the count of rare alleles was associated with a trait or disease of interest. More sophisticated approaches followed, introducing weights to allow for some SNVs to have larger effects on the trait, and using of different definition of "rare" based on minor allele frequencies, described in detail in Chap. 13. However, these approaches had highest power when all rare SNVs had the same direction of effect on the trait studied, meaning that all SNVs were either detrimental or beneficial, and were seriously underpowered in situations where both detrimental and beneficial SNVs had an influence on the trait of interest, or a large proportion of SNVs were neutral.

To remedy the shortcoming of the earlier collapsing approaches, a number of methods allowing for different direction of effects were proposed and have been evaluated in simulation settings. In the next section, we outline these approaches, with emphasis on their commonality, advantages, and disadvantages in the analysis of rare SNVs.

H. Chen • J. Dupuis (✉)
Boston University School of Public Health, Boston, MA, USA
e-mail: dupuis@bu.edu

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_11

Methods

All approaches described in this section start from the following basic model:

$$g\left[E\left(Y_i\right)\right] = \gamma_0 + \sum_c \gamma_c z_{ic} + f\left(G_i\right) \tag{1}$$

where Y_i is the trait of interest, either a quantitative trait or a binary disease indicator, z_{ic} is the value of the cth covariate in individual i, γ_c is the effect of the cth covariate on the trait Y, G_i is the genotype at all SNVs within a functional unit (gene or pathway) for individual i, and $f(G_i)$ is a function on the genotypes. The function $g(\cdot)$ is a generalized linear model link function. For example, one may use the logit link function for binary traits and the identity link for quantitative traits.

More specifically, if $f(\cdot)$ is a linear function, then

$$g\left[E\left(Y_i\right)\right] = \gamma_0 + \sum_c \gamma_c z_{ic} + \sum_j \beta_j G_{ij} \tag{2}$$

where G_{ij} is the number of rare alleles carried by individual i at SNV j and β_j is the effect of SNV j on the trait.

In joint tests of association, the typical hypothesis of interest can be written as H_0: $\beta_j = 0$ for all j, although the specific form of the null hypothesis and the choice of test statistic vary according to the approach. For example, a general collapsing test statistic may be obtained by setting $\beta_j = \beta w_j$, where w_j is a weight assigned to the jth SNV. The w_j are assumed to be known, although in practice they are often estimated from the observed data. When assuming $\beta_j = \beta w_j$, (2) can be written as

$$g\left[E\left(Y_i\right)\right] = \gamma_0 + \sum_c \gamma_c z_{ic} + \beta \sum_j w_j G_{ij} \tag{3}$$

and the null hypothesis becomes H_0: $\beta = 0$. A Wald test, score test, or likelihood ratio test can be used to test the null hypothesis in a regression context. Using the notation and model defined in (1), we describe a number of methods for joint analysis of rare SNVs that go beyond the collapsing methods described in Chap. 13.

The Data-Adaptive Sum (aSum) Test

The data-adaptive sum (aSum) test proposed by Han and Pan (2010) is one of the earliest approaches developed for the scenario when both deleterious and protective SNVs are present. The original model used by Han and Pan reduces to (3) without covariates although it is simple to extend the approach to include covariates. The novelty of Han and Pan's approach rests in the definition of the vector of weight w_j, which depends on the observed data in the following way. Han and Pan defined $\hat{\beta}_{Mj}$

as the estimate of the effect of SNV j in the model with a single SNV included (M stands for marginal model), and P_{Mj} as the p-value for the test H_0: $\beta_{Mj}=0$. Then, for a pre-specified cutoff α_0, Han and Pan suggested setting $w_j=-1$ if $\hat{\beta}_{Mj}<0$ and $P_{Mj}\leq\alpha_0$, and $w_j=1$ otherwise. The choice of threshold α_0 will influence the power of the test. In the case of $\alpha_0=0$, all $w_j=1$ and the approach reduces to an unweighted collapsing test, where the rare SNV count is tested for association with a trait. In the case of $\alpha_0=1$, w_j is set to the sign of $\hat{\beta}_{Mj}$, the marginal effect of each SNV.

Han and Pan recommended using a score test to evaluate the association between $\sum_j w_j G_{ij}$ and the trait of interest. However, because the w_j's are selected based on the significance and sign of the single SNV estimated effects, using the asymptotic distribution to assess the significance of the score test would lead to inflated type-I error rate. To surmount this problem, Han and Pan proposed a permutation approach, where phenotypes (and covariates if applicable) are permuted among unrelated individuals and the procedure is repeated, selecting the most appropriate w_j for each permuted dataset and computing the score statistic for association. Because significance thresholds in gene-based genome-wide studies are typically in the order of 10^{-6}, a large number of permutations would need to be performed in order to get accurate permutation p-values, which could render this procedure impractical. To alleviate this issue, Han and Pan evaluated a second approach to estimate the significance of their adaptive test by assuming that the distribution of the score statistic follows a shifted chi-square distribution of the form $a\chi_1^2+b$, where a and b are parameters estimated from the permutation distribution. Estimation of a and b can be performed with a few hundred permutations, and this greatly increases the efficiency of the procedure. In their evaluation, Han and Pan used only 100 permutations to estimate a and b, and compared the p-value obtained under the shifted χ_1^2 assumption to a more typical permutation test with thousands of permutations.

Han and Pan performed extensive simulation studies, showing that their approach outperforms collapsing tests in many scenarios. Although the evaluation of aSum using the reduced number of permutations and the shifted χ_1^2 assumption appears to yield the correct type-I error, they cautioned that this approach should be more thoroughly studied and that the permutation distribution without this shifted χ_1^2 assumption is preferable, when feasible, to assess the significance of the test statistic. Given that Han and Pan explored the accuracy of the shifted χ_1^2 distribution at the $\alpha=0.05$ level only, and not in the tail when the accuracy is most important, this warning by the authors seems warranted.

The greatest advantage of the Han and Pan's approach is the gain in power over collapsing approaches when both deleterious and protective SNVs influence the trait of interest. However, there are a number of shortcomings to the approach. First, the permutation procedure greatly increases the computational burden. Second, the method is only applicable to unrelated individuals because the permutation procedure assumes that observations are interchangeable, and hence independent. This assumption will be violated in family samples and may be too restrictive in unrelated samples with cryptic relatedness, as would be present in population isolates. Finally, Han and Pan's approach will be most powerful when all SNVs

have the same magnitude of effects because of the simple +1/−1 weighting scheme. Because it is expected that some SNVs may have a large effect on the trait of interest, and that some SNVs may have no effect at all, a number of approaches were proposed to address this weakness.

Step-Up Method

Hoffmann et al. (2010) proposed a general step-up approach to allow SNVs to have different effect on the trait, taking into consideration that some SNVs may have no effect at all. Model (3) is also the basis for the step-up approach, although their original model does not allow for inclusion of covariates. However, the approach could easily accommodate covariate adjustments. Again, the difference in the step-up approach from other proposed rare SNV methods comes down to 1) the choice of test statistic, and 2) the choice of weights w_j.

To evaluate the association between SNVs and trait, Hoffman et al. (2010) suggested using the score test with empirically derived variance:

$$T = \frac{\left[\sum_{i=1}^{N} U_i\right]^2}{\sum_{i=1}^{N} U_i^2}$$

where $U_i = \left(Y_i - \bar{Y}\right)\sum_{j=1}^{J} w_j \left(G_{ij} - 2\hat{p}_j\right)$, $\bar{Y} = \sum_{i=1}^{N} Y_i / N$, and $\hat{p}_j$ is the estimated minor allele frequency of SNV $j\left(\hat{p}_j = \sum_{i=1}^{N} G_{ij} / \left[2N\right]\right)$. This statistic can be computed efficiently for both binary and quantitative traits. Inclusion of covariates can be accommodated by replacing $\bar{Y}$ by $\hat{\mu}_i$, where $\hat{\mu}_i = \hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_{ic}$ for continuous traits and $\hat{\mu}_i = \text{logit}^{-1}\left(\hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_{ic}\right)$ for binary traits, with $\hat{\gamma}_0$ and $\hat{\gamma}_c$ estimated under the null hypothesis of no rare variant influence on the trait. When the weights are known, this score statistic follows asymptotically a χ_1^2 distribution. However, the optimum weighting scheme is usually unknown, and the authors proposed various ways of setting the weights w_j to maximize power.

Hoffman et al. (2010) proposed to use weights of the form $w_j = a_j s_j v_j$, where a_j is a continuous weight, s_j depends on the direction of effect, as in the Han and Pan's approach above, and v_j is an indicator variable specifying whether SNV j belongs in the model (i.e., has a nonzero effect on the trait studied). This model addresses two of the shortcomings of the Han and Pan's approach. First, it takes into account that some SNVs may be "noise" SNV and have no effect on the trait. Secondly, the a_j's allow for SNVs to have different effect sizes. Although this is a very general model, one has to define a_j, s_j, and v_j in order to perform a test of

association. In the next paragraph, we describe some options that Hoffmann et al. (2010) proposed for setting the components a_j, s_j, and v_j.

The term a_j allows for SNVs to have different magnitude of effect on the trait. If one assumes that rarer SNVs have a larger effect on the trait, a natural choice for a_j is the Madsen–Browning weight function (2009) that depends on the allele frequency and are proportional to $\frac{1}{\sqrt{\hat{p}_j\left(1-\hat{p}_j\right)}}$, where $\hat{p}_j$ is the estimated minor allele frequency of SNV j. A more general form for a_j is the beta function with parameters α and β. Setting $\alpha=\beta=½$ is equivalent to the Madsen–Browning weight (2009). Wu et al. (2011) proposed using $\alpha=1$ and $\beta=25$. If one assumes that all SNVs have the same effect on the trait, then one should give equal weights to all SNVs by setting $a_j=1$. A comparison of these three weighting schemes is presented in Fig. 1.

The other components of the weighting function, s_j, allow for SNVs to have different direction of effect $\left(s_j=-1 \text{ or } +1\right)$. Values of s_j are usually set based on the observed data. As described above, Han and Pan (2010) proposed an approach for setting s_j based on the sign (and significance) of the regression coefficient. Hoffmann et al. (2010) proposed a modified approach that is computationally more efficient when there are no covariates. For binary traits, $s_j=-1$ when the SNV is more prevalent in controls, and +1 otherwise. For continuous traits, s_j is the sign of the correlation coefficient between the additively coded SNV and trait.

The final components of the weighting function, v_j, determine which SNVs are allowed to enter the model and hence are assumed to influence the trait. Values of v_j may be determined using prior information, such as functional annotation (e.g., v_j for non-synonymous SNVs), or may be data-driven (e.g., $v_j=1$ if $\hat{p}_j<0.01$). Hoffmann et al. (2010) proposed an iterative procedure for setting v_j called the

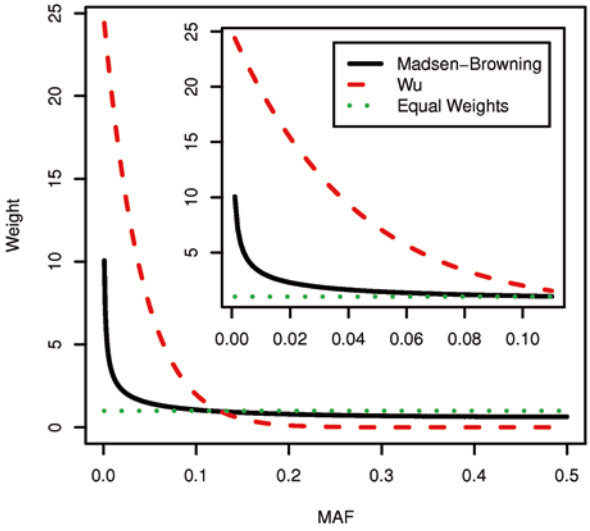

Fig. 1 The weighting schemes as a function of minor allele frequency (MAF)

"step-up" approach that is akin to forward selection in regression. First, all models with only one SNV are evaluated and the model with the largest score statistic is selected. Then, all models including that first selected SNV and one other SNV are evaluated. The score statistic for the best model with two SNVs is compared to the score statistic including only the best SNV; if the model with two SNVs has a higher score statistic than the model with one SNV, the procedure continues including SNVs in the model in this iterative fashion until the score statistic no longer increases.

Statistical significance of the final score statistic is evaluated empirically, permuting the trait values among all individuals and performing the step-up procedure for each permutation. The final *p*-value is the proportion of permutation datasets with a score statistic higher than the observed score statistic. The procedure has been implemented in an R package thgenetics (http://cran.r-project.org/web/packages/thgenetics/index.html). Although the R package does not allow for covariate adjustment, there is nothing in the theoretical development of the approach that would prevent inclusion of covariates. The R package is fairly efficient when analyzing a moderate number of SNVs (~20), but becomes highly computationally intensive with larger number of SNVs (~100) although the implementation allows the users to analyze subsets of SNVs that are then combined into a single test statistic (the "pathway" option).

The step-up approach is very general and encompasses many of the previously described collapsing tests and approaches. For example, if s_j is set according to the sign and significance of the regression coefficient from the marginal model, with $v_j = 1$, we are back to the Han and Pan's approach. If the a_j is set to the Madsen–Browning weights, also with $v_j = 1$, then we get the Madsen–Browning test. However, the permutation procedure required by both the Hoffman et al. and the Han and Pan's approaches poses a challenge for their genome-wide implementation. Moreover, the permutation approach is valid when observations are independent and therefore not appropriate for family samples without omitting related samples or adapting the permutation procedure to account for correlated observations, an issue that remains a challenge.

Sequence Kernel Association Test

Despite both the Han and Pan (2010) and Hoffman et al. (2010) approaches not having an implicit assumption that all SNV effects are in the same direction, the computational limitation imposed by the required permutation procedure is a drawback. Wu et al. (2011) proposed the sequence kernel association test (SKAT), a method that accommodates SNVs with different direction of effects and does not require permutation. SKAT is based on model (2), and the null hypothesis of interest is H_0: $\beta_j = 0$ for all j. However, because β_j cannot be reliably estimated for rare SNVs, Wu et al. (2011) assume that each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j^2\tau$, where w_j is a known weight for SNV j and τ is a variance component.

A test of H_0: $\beta_j = 0$ for all j is equivalent to testing H_0: $\tau = 0$. Wu et al. (2011) propose to perform a test of this latter hypothesis using a variance-component score test for a mixed model, assuming γ_c are fixed effects and β_j are random effects. The test statistic $Q = \left(Y - \hat{\mu}\right)' GWWG' \left(Y - \hat{\mu}\right)$, where $\hat{\mu}$ is the predicted mean of Y under the null hypothesis, $K = GWWG'$ is the weighted linear kernel matrix, W is a matrix whose diagonal elements are w_j and non-diagonal element are 0, and G is the $N \times J$ matrix of additively coded genotype. Note that $\hat{\mu} = \hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_c$ for continuous traits and $\hat{\mu} = \text{logit}^{-1}\left(\hat{\gamma}_0 + \sum_c \hat{\gamma}_c z_c\right)$ for binary traits. In the special case where Y is binary and there are no covariates, the SKAT statistic is equivalent to the C-alpha test proposed by Neale et al. (2011). In the C-alpha statistics, each rare SNV has the same probability of occurring in cases and controls under the null hypothesis of no association. Excess occurrence in cases or in controls is taken as evidence for association. A measure of excess occurrence is aggregated over all SNVs to create the C-alpha statistic. The SKAT statistic can be seen as a generalization of the C-alpha test, allowing for continuous traits and covariates or equivalently, the C-alpha test is a special case of a SKAT statistic. Under the null hypothesis, Q follows a weighted sum of χ_1^2 statistics, $Q \sim \sum_{j=1}^{J} \lambda_j \chi_{1,j}^2$ with λ_j estimated from the eigenvalues of a function of the weighted genotype covariance matrix. Therefore, evaluation of the significance of Q can be achieved analytically without resorting to permutation. The Q statistic can be re-written as the sum of the score test for each individual SNV: $Q = \sum_{j=1}^{J} w_j^2 S_j^2$, where $S_j = \sum_{i=1}^{N} G_{ij}\left(Y_i - \hat{\mu}_i\right)$. When using equal weights (W is the identity matrix, all $w_j = 1$), the SKAT statistic is equivalent to the sum of squares of the marginal score statistics (SumSqU, or SSU) proposed by Pan (2009). This form of the Q statistic is extremely useful when analyzing multiple cohorts. For example, one could use inverse variance weighted meta-analysis to obtain a pooled estimate of the score statistic for each variant, and use the meta-analyzed scores in the computation of the Q statistics. Similarly, the asymptotic distribution of the meta-analyzed Q could be obtained by pooling the genotype covariance matrix to evaluate significance.

More generally, instead of the linear function in model (2), SKAT can also take a more flexible function $f(G_i)$ in model (1), thus allowing for interactions among variants. Assuming the vector $f(G)$ of size N follows a distribution with mean 0 and covariance matrix τK, the test statistic $Q = \left(Y - \hat{\mu}\right)' K \left(Y - \hat{\mu}\right)$ may be used to evaluate the null hypothesis $H_0 : \tau = 0$.

SKAT offers many advantages over other approaches. First, the computational efficiency that results from using asymptotic rather than empirical distribution of the test statistic under the null hypothesis makes it feasible to apply to genome-wide studies. Moreover, the robustness of the test statistic to the direction and magnitude of effects offers increased power in scenarios where both deleterious and protective SNVs are at play. However, when most SNVs have the same direction of effect, SKAT has been shown to be less powerful than a simpler burden tests. For this reason, a combination of burden test and SKAT statistic may offer better power.

SKAT-O

When most SNVs included in the analysis are functionally related to the trait of interest and have the same direction of effect, then a burden test may outperform SKAT. Lee et al. (2012) proposed an extension to the SKAT statistic to deal with this scenario. They proposed a different class of kernels to use in the SKAT test, and the resulting Q statistic derived from this class of kernels is equivalent to a linear combination of the burden test and SKAT statistics:

$$Q_\rho = (1-\rho)Q_{\text{SKAT}} + \rho Q_{\text{burden}}, \quad \text{with} \quad 0 \le \rho \le 1.$$

When $\rho = 0$, Q_ρ reduces to the SKAT statistic; when $\rho = 1$, Q_ρ reduces to the burden test statistic $Q_{\text{burden}} = \left(\sum_{j=1}^{J} w_j S_j \right)^2$, which is the square of the score test statistic for H_0: $\beta = 0$ in model (3). For a fixed value of ρ, the distribution of Q_ρ follows a weighted sum of χ_1^2 distribution, with weights estimated from the eigenvalues of a function of the weighted genotype covariance matrix. However, Lee et al. (2012) suggested a data-driven approach to setting the value of ρ to optimize power by finding the minimum p-value over all values of ρ. They provide a procedure to evaluate the significance of this new test statistic that takes into consideration the fact that the p-value was minimized over ρ, a nuisance parameter which is present only under the alternative hypothesis. Again, the procedure does not require permutation and is highly computationally efficient. The name of this new procedure is SKAT-O, where "O" stands for optimized. Via simulations, Lee et al. (2012) showed that this procedure has close to equivalent power to the burden test when a large proportion of the SNVs have the same direction of effect, and power close to the original SKAT statistic in the context of SNVs with different direction of effect.

Wang et al. (2012) also proposed a joint test (Score-Joint), combining a burden test, equivalent to the square-root of Q_{burden} above, with a test of the variance component parameters τ defined in the SKAT section. Compared with SKAT-O, it is a joint test on two parameters, and it requires permutation to evaluate significance.

The SKAT-O statistic offers some power advantage over the original SKAT procedure when the proportion of influential SNVs is large and most SNVs have the same direction of effect, at small cost of some added complexity in computation.

Score-Seq

Lin and Tang (2011) proposed a slightly different procedure to test for association between a group of SNVs and a trait of interest. As for all the previous approaches, the basis of the method is model (2). In the same spirit as many of the collapsing approaches, Lin and Tang's approach assumes that $\beta_j = \beta w_j$, where w_j is the weight assigned to SNV j, and the model reduces to model (3) previously described. To test

the null hypothesis H_0: $\beta=0$, assuming a known vector of weights w, Lin and Tang derived the score statistic, which is of the form $U=\sum_{i=1}^{N}\left(Y_i-\hat{\mu}_i\right)G_i w$ with variance $V=\hat{\sigma}_0^2\left\{\sum_{i=1}^{N}\left(G_i w\right)^2-N^{-1}\left(\sum_{i=1}^{N}G_i w\right)^2\right\}$ when there is no covariates involved (but a more complex form with covariates). Note that $\hat{\sigma}_0^2=\overline{Y}\left(1-\overline{Y}\right)$ for binary trait and the estimated variance of Y for quantitative trait. The statistic $T=U/\sqrt{V}$ may be used to determine if the rare SNVs have an effect on the trait Y. The power of the test will depend on the choice of w, with optimum power achieved when $w_j=\beta_j$, the true (but unknown) value of the effect size parameter. Lin and Tang's approach differs from the typical weighted collapsing method in setting the values of the weight vector. When considering weighting schemes, Lin and Tang proposed two ways to achieve maximum power: (1) Maximizing the test statistic over multiple weight vectors and (2) setting weights from the Estimated REgression Coefficients (EREC). We describe both sets of weights below.

Maximizing the Test Statistic Over Multiple Weight Vectors

Given L weight vectors, $w^1, \ldots w^L$, each of length J, that include the weights for each of the J SNVs in the analysis, one can compute L score statistics (T_l) to test the association between the trait and the weighted genotypes formed by Gw^l. Ling and Tang suggested using the maximum test statistic over all weight vectors $\left(T_{\max}=\max\left|T_l\right|\right)$ to test for association between the SNVs and the trait. They derive the asymptotic distribution of $T_{\max}$ by assuming that the T_l statistics follow a multivariate normal distribution with mean 0, and with an estimated covariance matrix that can be computed from the data and weight vectors. Significance of the test can be evaluated asymptotically using the equation:

$$\Pr\left(T_{\max}>t_{\max}\right)=1-\Pr\left(\left|T_1\right|<t_{\max},\ldots,\left|T_L\right|<t_{\max}\right).$$

For example, one could evaluate the T statistic for equal weight ($w_j=1$ for all j), the Madsen–Browning weight and the Wu weight, and use the maximum statistic over these three weight vectors, taking into account that the statistic was maximized over three weight functions when evaluating significance. This may offer increased power over collapsing approaches using a single set of weights. One could also define the weights with a variable threshold based on allele frequencies to determine inclusion of SNVs, and maximize over multiple allele frequency thresholds. This is akin to the variable threshold (VT) test proposed by Price et al. (2010), with the added advantage that significance may be evaluated without the need for computationally intensive permutations.

One of the greatest advantages of this approach is the ability to evaluate empirically the significance of the test statistics when multiple weight functions are evaluated.

In practice, because the trait etiology is often unknown and one does not know, a priori, which rare SNVs influence the trait, investigators often evaluate multiple weight functions, which may involve restricting which SNVs are included in the test, based on function or other annotation, or by relaxing the definition of "rare" to allow common SNVs to be included. However, correction for multiple testing is often performed using a simple Bonferroni correction, leading to overly conservative tests because the correction does not take the correlation of the test statistics into consideration. The ability to properly correct for multiple testing induced by the evaluation of multiple weights function is a great addition to the literature.

Nevertheless, the approach would still have low power in the presence of both deleterious and protective rare SNVs, prompting Lin and Tang to explore a different approach to determine the optimal weight vector.

Estimated REgression Coefficients

As noted earlier, the most powerful test would be obtained by setting $w_j = \beta_j$, the true but unknown value of the parameter. While β_j may be estimated from the data, it will likely be poorly estimated because of the low frequency of the tested alleles. Lin and Tang suggested setting $w_j = \hat{\beta}_j + \delta$, where δ is a given constant. This is similar to Han and Pan's earlier approach, where w_j was dependent on the significance and sign of the beta estimate, although Han and Pan (2010) ignored the magnitude of the effect estimates. Because the data is used in setting the optimum weights, significance is evaluated using a permutation approach, where the phenotype value Y (and covariates if applicable) are permuted among individuals, and both weights and test statistics are recomputed with permuted data. It is important to permute both trait and covariates together; the null hypothesis is evaluated by breaking the relationship between genotype and trait, but keeping the relationship between the trait and covariates intact. Lin and Tang implemented this approach into the software Score-Seq, with an adaptive permutation test that selects fewer permutation iterations for large p-values but increases the number of permutation iterations to get more precision for low p-values.

The authors recommend setting $\delta = 1$ for binary traits and $\delta = 2$ for standardized quantitative traits when the sample size is less than 2,000. The authors have not explored the effect of varying δ on power.

The authors compared the multiple weight evaluation approach and Estimated REgression Coefficients (EREC) method with other available methods, namely the collapsing approach by Madsen and Browning (2009), the variable threshold approach proposed by Price et al. (2010), and SKAT. They showed the advantage of evaluating multiple weight functions over most collapsing tests when all SNVs had the same direction of effect. They also showed that EREC has a clear advantage over SKAT when all SNVs have the same direction of effect with no neutral SNVs included, a fact that was acknowledged by Wu et al. (2011) and remediated with the introduction of the SKAT-O statistic. In the presence of both deleterious

and protective SNVs, Lin and Tang (2011) also demonstrated an advantage of EREC over the SKAT statistic, claiming that the gain in power is due to the overly conservative asymptotic evaluation of the significance of SKAT statistic, while their permutation evaluation is not conservative. However, they acknowledge that the SKAT method is more computationally efficient than the EREC test.

Kernel-Based Adaptive Cluster

Liu and Leal (2010) proposed the kernel-based adaptive cluster (KBAC) approach, which classifies genotypes into groups based on multi-locus genotype patterns. Their method can be formulated using model (1) defined earlier.

For a set of J variants, there are at most 3^J genotype groups. However, when testing rare variants, the number of observed genotype groups may drop dramatically because of the low minor allele frequency and linkage disequilibrium. Given J SNVs, the $M+1$ distinct genotype patterns are denoted by $P_0, P_1, \cdots P_M$, and P_0 represents a pattern with no rare alleles. Using the model defined in (1), Liu and Leal (2010) let $f(G_i)=\eta K_m$ for individual i with genotype pattern P_m, where the kernel K_m is estimated from the data. The null hypothesis H_0: $\eta=0$ is evaluated using a score test to determine if there is some association between genotype patterns and phenotype. Because the kernel is data-driven, a permutation procedure is implemented for p-value evaluation.

Liu and Leal proposed (2010) three types of kernels for case–control designs: hyper-geometric kernel, marginal binomial kernel, and asymptotic normal kernel. Their evaluation of the approach focused on the hyper-geometric kernel, defined as

$$K_m = \sum_{r \in \left\{0,1,\ldots,N_m^1\right\}} \frac{\binom{N_m}{r}\binom{N-N_m}{N^1-r}}{\binom{N}{N^1}}.$$

where N^1 and N^0 are the number of cases and controls, respectively, with $N = N^1 + N^0$, and N_m is the number of individuals with genotype pattern P_m among which there are $N_m{}^1$ cases and $N_m{}^0$ controls. The kernel is different from the kernel in SKAT, because it is data-driven and depends on the genotype–trait relationship. Appropriate kernels for quantitative trait analyses were not proposed.

When there are no covariates, the score statistic from the logistic regression model (1) reduces to the KBAC statistic (up to a constant scalar):

$$\mathrm{KBAC} = \left(\sum_{m=1}^{M} K_m \left(\frac{N_m^1}{N^1} - \frac{N_m^0}{N^0} \right) \right)^2,$$

Basu and Pan (2011) suggested that KBAC might not perform well when there are both deleterious and protective variants, and when the proportion of causal variants is small. However, compared with other approaches, KBAC is attractive in rare variants association analysis because it allows for interactions among variants, by testing genotype patterns of multiple variants as a group, rather than simply summing up genotypes or test statistics from individual variants.

Discussion

All approaches described in this chapter use the same underlying model linking a trait to rare SNV genotypes, described in (1). Many other approaches for rare variant analyses have been proposed in the literature. Two examples of non-regression-based approach include the replication-based test (RBT) proposed by Ionita-Laza et al. (2011) and the functional principal component analysis (FPCA) introduced by Luo et al. (2011)

The RBT was developed for case–control designs and looks for more frequent occurrences of mutations in either cases or controls. Enrichment in cases is measured by a weighted sum of indicators of higher allele frequency in cases compared to controls, where the weights are data-driven and are higher for variants with larger difference in allele frequency between cases and controls. Because rare variants may be protective, a similar statistic for enrichment in controls is computed, and the RBT statistic is defined as the maximum of the two enrichment statistics. Statistical significance is evaluated by permutation. Compared with burden tests, RBT is less sensitive to the presence of both deleterious and protective variants, but power is reduced when the proportion of causal variants is low.

Luo et al. (2011) proposed the FPCA approach, which takes both rare variants and their genomic locations into consideration. From a functional data analysis point of view, they treat the positions as a continuous variable and define the genotype of each individual as a function of positions. By using data reduction and smoothing techniques, FPCA overcomes the high-dimensionality and multicollinearity issues in multivariate tests and collapsing methods, and is less sensitive to sequence errors and missing data. However, the multivariate nature of the Hotelling's T^2 test performed after reducing the dimension of genotype data using principal components may hamper power over lower dimensional methods described in this chapter. When the correlation between rare variants is low, FPCA introduces extra computational burden, but may not have much power gain compared to multivariate tests on the original genotype data. Also, FPCA does not adjust for covariates and is not directly applicable to quantitative traits, although such extensions would be straightforward.

In an ideal world, one would have infinite data and would be able to assess the effect on the phenotype of each rare variant individually. However, because of limits in sample sizes imposed by budget constraints and also simply by the availability of cases for certain rare diseases, getting reliable estimate of the effect of rare SNV on

Table 1 Summary of non-collapsing rare variant association analysis approaches

Test	Binary	Quantitative	Covariates	*p*-Value	References
SSU	Yes	Yes	Yes	Analytical	Pan (2009)
aSum	Yes	Yes	Yes	Permutation	Han and Pan (2010)
KBAC	Yes	No	Yes	Permutation	Liu and Leal (2010)
Step-up	Yes	Yes	Yes	Permutation	Hoffmann et al. (2010)
RBT	Yes	No	No	Permutation	Ionita-Laza et al. (2011)
C-alpha	Yes	No	No	Either	Neale et al. (2011)
FPCA	Yes	No	No	Analytical	Luo et al. (2011)
SKAT	Yes	Yes	Yes	Analytical	Wu et al. (2011)
Score-Seq	Yes	Yes	Yes	Analytical	Lin and Tang (2011)
EREC	Yes	Yes	Yes	Permutation	Lin and Tang (2011)
Score-Joint	Yes	Yes	Yes	Permutation	Wang et al. (2012)
SKAT-O	Yes	Yes	Yes	Analytical	Lee et al. (2012)

the quantitative trait or disease of interest is often not feasible. Therefore, additional assumptions are needed in order to identify rare SNVs associated with a phenotype. The rare variant approaches included in this chapter differ in their assumptions. Obviously, the closer the assumptions are to the "truth," the more effective the approaches will be at identifying SNVs and genes that are important in disease etiology. The most powerful approach will often depend on the true trait model, which unfortunately remains unknown for most traits under investigations. To a lesser extent, the choice of test statistic will also affect the ability to identify the causal variants. Table 1 summarizes the non-collapsing rare variants association analysis approaches mentioned in this chapter. Below we discuss differences between the approaches presented in this chapter, and how these differences may affect the ability to identify SNVs and genes influencing a quantitative trait or disease of interest.

Test Statistic and Evaluation of Statistical Significance

The approaches described in this chapter differ by the test statistic used to evaluate the null hypothesis of no association. However, they all have one thing in common: they strive to use computationally efficient statistics that can be computed genome-wide. All approaches use a score test because it is less computationally intensive than a likelihood ratio test. Moreover, all approaches strive for efficient evaluation of their score test.

In aSum, although permutation is required for evaluation of the score statistic, Han and Pan (2010) investigated ways to decrease the computational burden of their permutation procedure. Because very small *p*-values are required when analyzing multiple genomic regions, a large number of permutations are typically required to estimate such small *p*-values. Han and Pan (2010) investigated approximation to the

permutation distribution by a scaled non-central chi-square, and used a small number of permutations to estimate the scaling and shift parameters.

Hoffmann et al. (2010) also used a score statistic and permutation. To improve upon Han and Pan's method in terms of computation efficiency, they determine the direction of effect based on the correlation coefficient, doing away with formal testing of each variant. In addition, they implemented an adaptive permutation approach, where a few initial permutations are used to assess the p-value, and additional permutations are performed only when the p-value is below a certain threshold. While it is certainly feasible to apply Hoffman et al.'s approach to a large number of genomic regions across the genome, the computational burden of the permutation approach prevents large-scale simulation evaluation of the approach.

The SKAT and SKAT-O statistics are also score tests, but with the advantage that statistical significance can be evaluated theoretically, without requiring time consuming permutation. However, Lin and Tang noted that SKAT can be conservative, and suggested that permutation evaluation could improve power, especially for small samples.

While both SKAT and EREC offer a general framework to test for association between a group of SNVs and a trait using a score test, the difference in their underlying assumptions lead to a different score statistics: SKAT assumes that β_j follows a distribution with mean 0 and variance $w_j^2\tau$, while Lin and Tang (2011) assumes that β_j is of the form βw_j. Both methods are univariate tests, but τ is a variance parameter with one-sided alternative in SKAT, and β is a location parameter with two-sided alternative in Lin and Tang (2011), leading to different statistics with different distributions. While the significance of both score statistics may be evaluated empirically, Lin and Tang further propose to set the weights empirically, and because the data is used in setting weights, asymptotic evaluation is no longer possible.

KBAC classifies individuals into different groups based on genotype patterns, and performs a test on the difference between the proportions of each genotype group in cases and in controls. The test is similar to a weighted χ^2 test of independence. Liu and Leal (2010) used permutation to evaluate statistical significance. Noting that the original KBAC statistic suffers when there are both deleterious and protective variants within a particular genotype pattern, Basu and Pan (2011) proposed a modified statistic to overcome this issue. KBAC is distinctive in rare variant analysis by allowing for interactions, but it may suffer from loss of power when the proportion of non-causal variants is high, as the number of genotype patterns increases dramatically.

Missing Data and Imputing Rare SNVs

While most of the methods discussed in this chapter have been evaluated using targeted or exome sequencing, application of the methods could be extended to imputed genotypes. Rare SNVs are often poorly imputed in unrelated samples because of the low linkage disequilibrium with nearby SNVs. However, familial

transmission information, if available, may improve imputation. In model (2), the genotypes could be defined as the expected number of rare alleles, or dosage, instead of a three category variable indicating the number of rare alleles a person carries. Theoretically, all approaches described in this chapter based on model (2) can accommodate the use of dosage genotype, although not all software implementation can do so.

In the regression framework of model (2), missing genotypes are not allowed. One needs to either exclude observations with one or more missing genotypes, or impute such missing data. As the number of SNVs included in the analysis increases, excluding observations with missing genotypes will greatly reduce the sample size and power, even when the genotyping call rate is high. Therefore, most software includes some approaches for imputing the missing values. For rare SNVs, one option would be to set all missing genotypes to the homozygous major allele, which is the most likely genotype. This imputation scheme is easy to implement. However, for more common SNVs, it will create bias in allele frequency estimates, which in turn could result in false positive results if the missing rate differs in cases and controls. For this reason, SNVs with high missing rates are often omitted from analysis. A second approach to fill in missing genotypes is to impute the mean genotype value, or dosage, which is equal to twice the rare allele frequency. While this will not bias the estimate of allele frequency, this may cause other types of bias. For example, if the missingness is not random and participants with missing data are more likely to be from the case or control set, or if they have lower or higher trait values, then imputing the average dosage may create false association because most observations will have a genotype of 0 rare allele, while missing observations will have a dosage value of twice the rare allele frequency. This could be more pronounced if the imputation is performed in cases and control separately. The third option is to impute the missing data using information on nearby SNV and familial transmission, if available. This approach capitalizes on linkage disequilibrium at nearby SNVs to more precisely impute missing genotypes. Unfortunately, this type of imputation works best for common SNV, but imputation quality for rare SNV can be poor, especially if no familial information is available. Again, SNVs with differential missingness in cases and controls, or missingness pattern related to a quantitative trait studied, could lead to false positive errors. To avoid such bias one can omit SNVs with high missing rate, but also test for differential missingness in cases or controls, or for association between proportion missing genotypes and a quantitative trait. Wu et al. (2011) showed that for small amount of missingness, imputing to the most likely genotypes did not decrease power considerably.

Choice of Weights to Maximize Power

Weighting schemes are used in most rare variant methods to try to improve power to detect association between SNVs and trait. To reach maximum power, a weighting scheme should give close to zero weights to SNVs without effect on the trait,

and weights proportional to the effect size for associated SNVs. Because it is believed that rarer SNVs will have a larger effect on the trait, several proposed weighting schemes depend on the rare allele frequency, such as the Madsen–Browning and Wu weights. Madsen–Browning weights decrease much more rapidly than the Wu weight as the minor allele frequency increases; see Fig. 1. As a consequence, the effect of including more common SNVs when using the Madsen–Browning weight should be small, while more common SNVs would contribute more substantially to the test statistic under the Wu or equal weighting scheme. Hoffman et al. (2010) also described ways to include functional annotation in determining the weights, assuming that SNVs that are more likely to be damaging or functionally important would have a larger effect on the trait. Such annotation can also be incorporated in the SKAT and score-seq weighting schemes, although the Han and Pan +1/−1 cannot be easily generalized to take functional annotation into consideration. FPCA can also take weighted genotypes instead of original additive genotypes and calculate principal components. As prior information becomes more precise, methods that can incorporate information on function annotation will be most useful.

Which SNVs to Include in Association Testing

Ideally, only SNVs influencing the trait of interest would be evaluated for association with the trait. Unfortunately, one does not know, a priori, which SNVs are causal or in LD with causal SNVs, and which SNVs have no effect on the trait. Inclusion of "noise" SNVs will lower the power of the test, as will failure to include some causal SNVs. Therefore, one has to strike a balance between including too many SNVs, with some noise SNVs, and too few SNVs, missing important variants. There are two separate issues to deciding which subset of SNVs to include in a test: (1) definition of the genomic region and (2) selection of SNVs within a region.

While one may wish to evaluate large regions for association, inclusion of too many SNVs, many likely to have no effect on the trait, will impede the ability to detect true associations. Therefore, it is common to divide large genomic regions into smaller analysis units. A natural unit of analysis is a gene level, or if a finer division is sought, exons or transcripts may be used to define a genomic region of interest. However, most Genome-Wide Association Study (GWAS) findings map outside of gene regions, and investigators may wish to evaluate rare SNV in the region around GWAS findings (Hindorff et al. 2009). Genomic region boundary could be based on conserved regions across species, recombination estimates around the GWAS finding, or more agnostically based on sliding windows across the region of interest. The sliding window approach could easily be accommodated in the framework from Lin and Tang (2011), where the test statistic used is the maximum test statistic over a number of weight functions. One can think of a sliding winding as putting a weight of zero to all SNVs outside the window being considered, and use the method describe in Lin and Tang to get the significance of the maximum test statistic over multiple windows within a region. As a clearer picture emerges of how

rare SNVs influence traits, we will be able to use prior information to determine the best size and boundaries to define genomic regions for investigation. In the meantime, one has to explore various ways of defining genomic regions in order to maximize the chance to detect true associations.

Once genomic regions have been selected, one needs to determine which SNVs within the region to include in an association test. Burden tests often restrict analyses to SNVs with a low rare allele frequency, using threshold of 1, 2, or 5 %, and similar thresholds may be applied to the methods in this chapter. The optimal allele frequency threshold will depend on the frequency of the true causal SNVs, and using a too stringent threshold will omit important SNVs and reduce power, while a threshold that is too liberal will include too many noise SNVs and also decrease power. Variable threshold approaches, such as the one developed by Price et al. (2010) or by Lin and Tang (2011), can overcome the issue of having to evaluate a single allele frequency threshold. Functional annotation may also be used to try to identify SNVs that are more likely to influence the trait. However, recent publications indicate that there are a lot of functional elements outside of genes, so restricting analyses to protein-altering SNVs may miss important functional variants. Other measures of potential functionality, such as how conserved the region around the SNVs is in other species may be fruitful. An alternative is to include all SNVs within a region, but to use a weighting scheme to up-weight SNVs that are more likely to influence to trait based on annotation, and to down-weight SNVs that are most likely neutral. Hoffmann et al.'s approach gives specific examples regarding inclusion of prior annotation information in the evaluation of the null hypothesis of no association. Incorporating this information can be easily done by using different weighting schemes in the SKAT, Score-Seq, or FPCA framework. Obviously, as our functional annotation improves, our ability to detect genes and SNVs influencing the trait will also improve.

Meta-analysis

Another consideration when selecting the most suitable approach for analysis of rare SNVs is the availability of meta-analysis approaches. In the GWAS context, most discoveries were achieved after the formation of large consortia, where meta-analysis of many cohorts uncovered loci with smaller effect on the traits of interest. The need for larger sample sizes may be even more pronounced in the analysis of rare SNVs, where a single cohort may have very few individuals carrying rare alleles for a particular SNV, so that joining forces with other studies will be crucial for discoveries of rare SNV association. Because all approaches provide evaluation of the significance of an association test in the form of a p-value, one can use a p-value-based approach, such as the Fisher or Stouffer approach, for combining results from multiple cohorts. However, methods that directly combine the beta estimates from model (2) may offer improved efficiency (Lee et al. 2013; Liu et al. 2014). Development of efficient meta-analysis approaches will be important in our quest to identify rare variants influencing traits of interest.

Other Types of Traits

While most traits studied fall in two categories, binary disease status or continuous measurements such as blood pressure, lipid levels, or fasting glucose, other phenotypes of interest may be time-to-event or ordinal/categorical measures. For example, one may be interested in studying time to recurrence of cancer, or age of development of type 2 diabetes. Some psychiatric disorders may have multiple levels of severity and may be best coded as ordinal variables (American Psychiatric Association 2000). Some of the approaches above naturally extend to other types of phenotypes. For example, Lin and Tang provide details on the application of their approach for time-to-event data, although their software implementation does not include this option. Chen et al. (2014) extended the SKAT statistic for survival traits. Other approaches, such as Han and Pan's or Hoffmann et al.'s method, could easily accommodate survival and ordinal traits using the typical regression framework (Cox proportional hazard model for survival and generalized linear model for ordinal data) because the significance is evaluated using permutation. The limiting factor is incorporation of these options into user-friendly and computationally efficient software that are easily accessible to investigators with these types of data.

Exome sequencing, exome chip, and whole genome sequencing have opened the floodgate on rare variants that were not investigated in the earlier GWAS era, when most studies focused on SNVs with frequency >1 %. Our success in identifying key genes that influence diseases and traits of interest will rest on the appropriate use of statistical tools, and gathering as much knowledge as possible on the potential function of the variants under study. Hopefully, the combination of these tools will lead to exciting new discoveries and will further our understanding of the architecture of complex traits.

References

American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders: DSM-IV-TR®. APA, Washington, DC

Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35(7):606–619

Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA et al (2014) Sequence kernel association test for survival traits. Genet Epidemiol 38:191–197

Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70(1):42–54

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106(23):9362–9367

Hoffmann TJ, Marini NJ, Witte JS (2010) Comprehensive approach to analyzing rare genetic variants. PLoS One 5(11), e13584

Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet 7(2), e1001289

Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–775
Lee S, Teslovich TM, Boehnke M, Lin X (2013) General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 93(1):42–53
Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89(3):354–367
Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6(10), e1001156
Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S et al (2014) Meta-analysis of gene-level tests for rare variant association. Nat Genet 46:200–204
Luo L, Boerwinkle E, Xiong M (2011) Association studies for next-generation sequencing. Genome Res 21(7):1099–1108
Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2), e1000384
Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7(3), e1001322
Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33(6):497–507
Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ et al (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832–838
Wang Y, Chen Y, Yang Q (2012) Joint rare variant association test of the average and individual effects for sequencing studies. PLoS One 7(3), e32485
Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93

Significance Thresholds for Rare Variant Signals

Celia M.T. Greenwood, ChangJiang Xu, and Antonio Ciampi

Introduction

To control the family-wise significance thresholds in genome-wide association studies (GWAS), thresholds ranging from 2.5×10^{-7} to 5×10^{-8} have been proposed (Browning and Thompson 2012; Dudbridge and Gusnanto 2008; Gao et al. 2010), and the latter threshold is in common use. These thresholds have been derived for univariate analysis of single nucleotide polymorphisms (SNPs), i.e., common genetic variation, and by studying patterns of linkage disequilibrium along the genome. The thresholds are well known to vary by population, since the patterns of linkage disequilibrium vary, as do the number of common variant sites.

With the recent arrival of large-scale sequencing studies, millions of extremely rare or unique genetic variants are being identified, and the number of variants seen

C.M.T. Greenwood (✉)
Lady Davis Institute for Medical Research, Jewish General Hospital,
3755 Côte Sainte Catherine, Montreal, QC, Canada H3T 1E2

Departments of Oncology, Epidemiology, Biostatistics and Occupational Health, and Human Genetics, McGill University, Montreal, QC, Canada
e-mail: celia.greenwood@mcgill.ca

C. Xu
Lady Davis Institute for Medical Research, Jewish General Hospital,
3755 Côte Sainte Catherine, Montreal, QC, Canada H3T 1E2

Department of Epidemiology, Biostatistics and Occupational Health,
McGill University, Montreal, QC, Canada

A. Ciampi
Department of Epidemiology, Biostatistics and Occupational Health,
McGill University, Montreal, QC, Canada

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_12

in each study increases with the sample size. Due to the rarity of the minor alleles, simple univariate tests have little power. In response, many region-based tests of genetic association have been developed that use simultaneously all the genetic variability in a defined window of the genome to test for association (see Chaps. 13 and 14 for overviews of many of the recently developed tests). Although each test has been shown to have valid type 1 error when used to analyze a single genomic region, the joint distribution of such test statistics has received little attention. In this situation, any single SNP or variant may participate in several different test statistics, by changing window lengths, or allowing adjacent windows to overlap. The thresholds established for GWAS of single polymorphisms will not be appropriate for this new context.

Here we discuss several issues that need to be considered in order to set genome-wide significance thresholds for whole-genome sequencing studies and region-based tests, and we focus particularly on the effective number of independent tests.

Effective Number of Independent Tests

The family-wise error rate (FWER) is defined as the probability of making one or more type 1 errors in a set of m tests. If the desired FWER is α_{FW} and all tests are independent, then the FWER can be controlled by testing at a smaller significance threshold, α_C, such that

$$\alpha_{\mathrm{FW}} = 1-\left(1-\alpha_C\right)^m . \tag{1}$$

For large m and for small α_{FW}, $\alpha_C \approx \alpha_{\mathrm{FW}} / m$, which gives the usual Bonferroni correction. However, when performing a large number of tests of association with genetic information, the tests are highly dependent, especially for nearby SNPs.

In order to control the FWER, therefore, when analyzing many SNPs, Cheverud (2001) proposed the idea of calculating an effective number of independent tests, m_e, and then using this number in a Bonferroni-style correction. Using the matrix of correlations between SNP genotypes, he argued that the eigenvalues of this matrix could be used to estimate the effective number of independent tests. Let R denote the matrix of genotype correlations between a set of m SNPs. If the genotypes are coded as (0, 1, 2) for the number of minor alleles, then the Pearson correlation coefficient is equal to one of the standard linkage disequilibrium (LD) measures. Conceptually, if all tests were independent, then all the eigenvalues of R would be 1.0. In contrast, if the tests were perfectly dependent, then there would be 1 eigenvalue with value m and the remaining $m-1$ eigenvalues would be 0. Cheverud (2001), therefore, proposed to estimate the effective number of independent tests, m_e, by a linear function of the variance of the eigenvalues, Var(λ) (see Table 1).

Table 1 Estimators of the effective number of independent tests for single-marker tests of association

Reference	Effective number of tests	Comments
Cheverud (2001)	$m_e = 1 + (m-1)\left(1 - \frac{1}{m}\mathrm{Var}(\lambda)\right)$ $\mathrm{Var}(\lambda) = \frac{1}{m-1}\sum_{i=1}^{m}(\lambda_i - 1)^2$	Overly large Eigenvalues of correlation matrix
Li and Ji (2005)	$m_e = \sum_{i=1}^{m}\left[I(\lambda_i \geq 1) + (\lambda_i - \lambda_i)\right]$ Equivalently, $m_e = m - \sum_{i=1}^{m}\left[I(\lambda_i > 1)(\lambda_i - 1)\right]$	Intuitive, large Eigenvalues of correlation matrix $\sum_i \lambda_i = m$
Patterson et al. (2006)	$m_e = \frac{(n+1)\left(\sum_i \lambda_i\right)^2}{(n-1)\sum_i \lambda_i^2 - \left(\sum_i \lambda_i\right)^2}$ n is sample size	Small Eigenvalues of population correlation matrix
Gao et al. (2008)	$m_e = \min_{1 \leq k \leq m}\left[\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i} > c\right]$ $c = 99.5\%$	Accurate Eigenvalues of correlation matrix
Moskvina and Schmidt (2008)	$\mathrm{FWER} \leq 1 - (1-\alpha)^{m_e}$ $m_e = 1 + \sum_{i=2}^{m}\kappa_i$ $\kappa_i \approx \sqrt{1 - \left(\max_{1 \leq k \leq i-1}\lvert r_{ki}\rvert\right)^{-1.3\log_{10}\alpha}}$	
Chen and Liu (2011)	$R_i = \sum_{j=1}^{m}\lvert r_{ij}\rvert^k$ $m_e = \sum_{j=1}^{m}\frac{1}{R_i}$	$K=7$ recommended Close to Li and Ji estimate by simulations
Li et al. (2012)	$m_e = m - \sum_{j=1}^{m}\left[I(\lambda_i > 1)(\lambda_i - 1)\right]$	Slightly less than Li and Ji's estimate Eigenvalues of correlation matrix of p-values

Many other estimators have since been proposed for estimating the effective number of independent tests when analyzing a large number of single SNPs (i.e., in GWAS) (Table 1). Several of these methods are also based on the eigenvalues from the matrix of SNP correlations (Gao et al. 2008; Patterson et al. 2006; Li and Ji 2005).

Among these, Patterson et al. (2006) proposed a moment estimator derived from the asymptotic distribution of the largest eigenvalue for a correlation matrix of normally distributed random variables; their focus was on population structure rather than significance testing. Some authors have used the correlations directly without calculating the eigenvalues (Moskvina and Schmidt 2008; Chen and Liu 2011), and in fact the variance of the eigenvalues can be expressed as a function of the trace of the correlation matrix (Cheverud 2001; Moskvina and Schmidt 2008). Recently, Li et al. (2012) proposed a method based on the correlation matrix of the *p*-values, rather than the genotypes, an estimator slightly altered from Li and Ji (2005), where eigenvalues greater than one are distinguished from those less than one when calculating m_e. After estimating the effective number of independent tests, formula (1) can be used to estimate the FWER (Šidák 1967) or inverted to estimate the necessary significance threshold for a desired value of FWER.

For rare genetic variation, however, the question of significance thresholds needs to be considered differently. Firstly, the correlations tend to be very low between rare genetic variants and any nearby SNP. For a singleton variant (e.g., seen in only person *i*), it is easy to derive that the correlation between the singleton and another SNP, denoted *x*, is

$$r = \frac{\sqrt{N}}{N-1}\frac{x_i - \overline{x}}{s_x}$$

where N is the number of people, x_i is the number of minor alleles carried by the person i at the SNP x, and $\overline{x}$ and s_x are the mean and standard deviation of the minor allele at the SNP x. For large N, this correlation will always be very small.

Although correlations between extremely rare genetic variants and more common variants are always low, testing for association with rare variants usually involves simultaneously assessing association with a set of variants in a defined region. Therefore, the more relevant correlation for rare variant tests is between two test statistics or *p*-values. In addition to the dependence on linkage disequilibrium patterns and minor allele frequencies (MAFs), these correlations will depend on the window size, the degree of overlap between adjacent windows, and the chosen test statistic including the weighting factors used for different variants. It is not necessarily straightforward to calculate these correlations, and the complexity varies with the chosen test statistic.

For a single choice of window sizes, genome-wide significance thresholds for window-based tests have recently been estimated (Zuk et al. 2014) using an extension of the approach of Lander and Botstein (Lander and Botstein 1989) that looked at the largest statistic as being the maximum deviation from an Ornstein-Uhlenbeck diffusion process (Uhlenback and Ornstein 1930). These authors assumed that statistics from nonoverlapping windows would be uncorrelated, and hence the correlation is determined solely by the degree of overlap. However, for several commonly used rare variant statistics, we have seen substantial correlations between nonoverlapping windows. Furthermore, this method cannot adjust for correlations between repeated analyses with different window definitions. Therefore, a more general approach would be useful.

Estimation of the Effective Number of Independent Tests for Whole-Genome Sequencing Region-Based Tests

We have recently proposed (Xu et al. 2014a) a computationally conservative empirical method for estimating the effective number of independent tests, given a chosen analytic strategy and for a specific data set. Essentially, we do not expect much correlation between region-based test statistics that are far apart or on different chromosomes. Therefore, we have proposed a strategy of evaluating the effective number of independent tests in smaller genomic regions and then extrapolating from these small regions to the whole genome.

Although most of the methods in Table 1 for estimating the effective number of tests were developed for use with single SNP analyses, we have compared the correlations, and the resulting estimates of m_e, for region-based SKAT (Wu et al. 2011) statistics. The SKAT test statistic is defined by

$$T = \frac{\left(y-\hat{\mu}\right)' K \left(y-\hat{\mu}\right)}{2\hat{\sigma}^2},$$

where y is the phenotype, $\hat{\mu}$ is the predicted mean of y under null hypothesis, and $K = \left(GW\right)\left(GW\right)'$ is the SKAT kernel matrix, which depends on the genotype matrix G, assumed to be centered, and a choice of variant weights W. Let T_i and T_j be two SKAT test statistics. Let $e = y - \hat{\mu}$ and $Q_i = e' K_i e$, where $K_i = \left(G_i W_i\right)\left(G_i W_i\right)'$. Under null hypothesis, $e \sim N\left(0, \sigma^2 I\right)$ and $\hat{\sigma}^2 \to \sigma^2$. Then under the null hypothesis, the correlations between two SKAT statistics can be analytically calculated as

$$\mathrm{cor}\left(T_i, T_j\right) \to \mathrm{cor}\left(Q_i, Q_j\right) = \mathrm{cor}\left(e' K_i e, e' K_j e\right) = \frac{\mathrm{tr}\left(K_i K_j\right)}{\sqrt{\mathrm{tr}\left(K_i^2\right)\mathrm{tr}\left(K_j^2\right)}}$$

Note that since the SKAT test statistic is a score test calculated under the null hypothesis, the correlation matrix does not depend on the phenotypes. Using these correlations, we have calculated m_e using all the methods listed in Table 1, using the pilot data on chromosome 3 from the UK10K project. We then estimated the significance threshold needed to control FWER at 0.05 by $0.05/m_e$.

The UK10K project is undertaking whole-genome sequencing and analysis of approximately 10,000 individuals from the UK with the goal of understanding the contribution of rare genetic variation to common traits and diseases (www.uk10k.org). For region-based analysis of rare variants in this consortium, an initial analysis plan defined regions to contain 50 rare variants, where "rare" is either MAF < 0.01 or MAF < 0.05. Adjacent regions were allowed to overlap by 25 rare variants. To study correlation patterns, we used a portion of this sequencing data (chromosome 3) from an interim release, including 2,432 individuals and 2,577,674 genetic variants. For an MAF threshold of 0.01, there were 74,156 regions defined, derived from 1,853,923 rare variants at this threshold.

For comparison, we also performed a simulation study to obtain empirical estimates of m_e. For each of the 2,432 sequenced individuals, we generated 1,000 sets of normally distributed phenotypes. For each set, all windows on chromosome 3 were analyzed with SKAT using the asymptotic p-value estimate based on the Davies method and hence leading to 1,000 sets of results under the null hypothesis, each spanning chromosome 3. An empirical estimate of the significance threshold, α_C, for a desired FWER of 0.05 (i.e., $\alpha_{\mathrm{FW}} = 0.05$) can then be obtained by (1) finding the minimum p-value for each of the 1,000 simulation sets and then (2) selecting the fifth percentile of these minimum p-values or the 50th smallest value. Then it follows that $\hat{m}_e = \alpha_{\mathrm{FW}} / \hat{\alpha}_C$.

To evaluate how estimates of m_e vary as a function of the number of windows analyzed, we divided chromosome 3 into subsections of varying size and calculated the effective number of independent tests in each subsection, using all the methods described in Table 1 as well as the simulations. Specifically, we divided chromosome 3 into as many as 1,024 equally sized subsections of 72–73 windows each and then into a range of larger subsections. Using the methods based on correlations and eigenvalues, the largest subsections that we examined contained 2,000 windows, but for the simulation-based approach, we were able to work with all of the 74,156 tests on chromosome 3 simultaneously.

In Fig. 1, the estimated significance thresholds controlling the FWER are plotted against Bonferroni thresholds as a function of the number of tests. Both axes show $-\log_{10} 0.05 / m^*$; for the x-axis, $m^* = m$, the number of windows being tested in a subsection of chromosome 3, and for the y-axis, m^*=$\hat{m}_e$, the estimated number of independent tests using different estimation methods. The points in Fig. 1 are the mean estimates of m_e across all subsections of the same size. Figure 1 shows, therefore, how the empirically derived estimates of the significance thresholds scale with the size of the genomic subsection are being analyzed, and it can be seen that for all methods, the relationship is linearly increasing with all slope estimates very close to 1.0. However, the various methods for estimating m_e from the correlations or eigenvalues give quite different results. Cheverud's (2001) estimate tends to be almost the same as the Bonferroni correction; in contrast, the estimate from the formula by Patterson et al. (2006) estimates a significance threshold that is substantially too large and extremely variable. The methods that seem to agree best with the simulations in Fig. 1 are either Li and Ji (2005) or Li et al. (2012).

We therefore can conclude that it is feasible to extrapolate from a small genomic region to the whole genome to obtain estimates of the necessary genome-wide significance thresholds. However, estimating m_e from the correlations requires evaluation of a very large number of correlations, and furthermore, calculation of eigenvalues of these matrices may prove too computationally challenging for some computer systems or very large numbers of windows. Therefore, the largest subregions we examined here using the correlation-based methods contained 2,000 windows. In contrast, although the simulations used substantial computer time, we were easily able to obtain estimates of significance thresholds for the full length of chromosome 3. Table 2 shows not only the numbers corresponding to the points in Fig. 1 but also the standard deviation measured across the subregions.

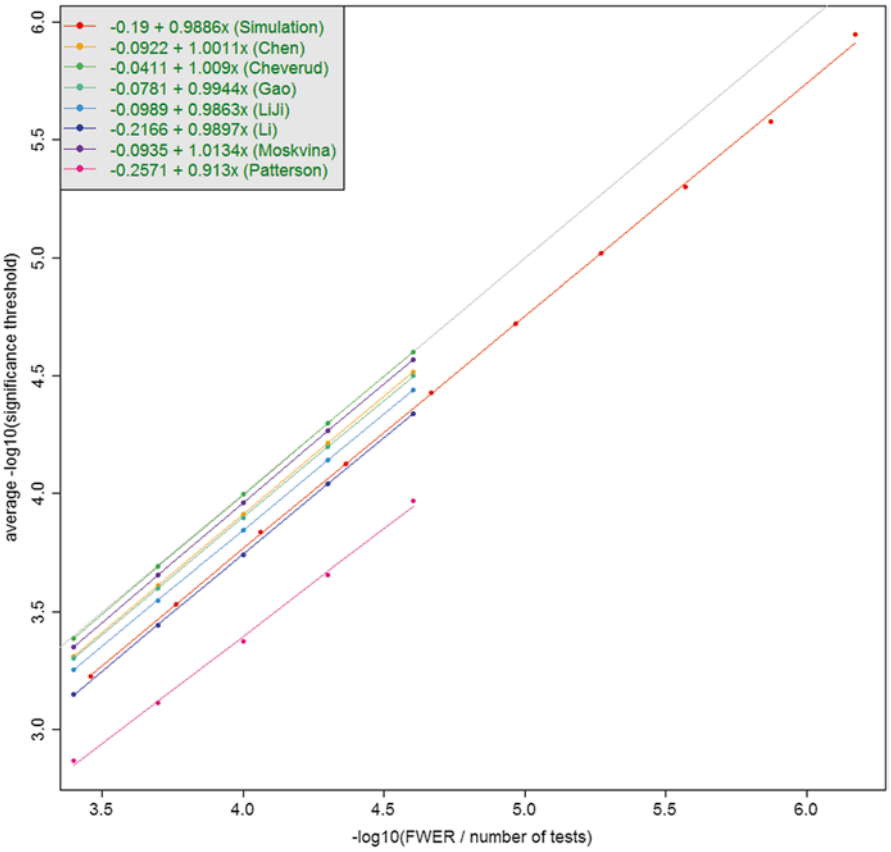

Fig. 1 Empirical significance thresholds estimated from chromosomal sections of different sizes, using several methods that are based on correlations and one simulation-based set of results

Table 2 Estimates of $-\log_{10}$ of the required significance threshold for subregions containing various numbers of tests

Method for estimating significance levels	Number of test statistics in the subregions			
	125	250	500	1,000
Bonferroni	3.40	3.70	4.00	4.30
Simulation	3.16 (0.102)	3.57 (0.090)	3.77 (0.080)	4.06 (0.069)
Cheverud (2001)	3.39 (0.018)	3.69 (0.011)	4.00 (0.006)	4.30 (0.002)
Li and Ji (2005)	3.25 (0.055)	3.55 (0.045)	3.84 (0.036)	4.14 (0.027)
Patterson et al. (2006)	2.87 (0.253)	3.11 (0.245)	3.37 (0.226)	3.66 (0.221)
Gao et al. (2008)	3.30 (0.013)	3.60 (0.012)	3.90 (0.010)	4.20 (0.007)
Moskvina and Schmidt (2008)	3.35 (0.025)	3.66 (0.019)	3.96 (0.015)	4.27 (0.010)
Chen and Liu (2011)	3.31 (0.067)	3.61 (0.055)	3.91 (0.045)	4.21 (0.028)
Li et al. (2012)	3.15 (0.047)	3.44 (0.039)	3.74 (0.032)	4.04 (0.023)

The means in the table correspond to the points in Fig. 1. Standard deviations are in parentheses

Table 3 Predictions for the effective number of independent tests genome wide, $\hat{m}_e$, and the corresponding genome-wide significance threshold, $\hat{\alpha}_C$, required to control FWER at $\alpha = 0.05$

Method	Number of tests in the largest subregions	Intercept	Slope	$\hat{m}_e$	$\hat{\alpha}_C$
Simulation	74,156	–0.190	0.9886	615,665	8.1213e-08
Simulation	2,000	–0.2215	0.9966	656,259	7.6189e-08
Cheverud (2001)	2,000	–0.0411	1.0090	1,225,467	4.0801e-08
Li and Ji (2005)	2,000	–0.0989	0.9863	730,252	6.8470e-08
Patterson et al. (2006)	2,000	–0.2571	0.9130	146,447	3.4142e-07
Gao et al. (2008)	2,000	–0.0781	0.9944	878,552	5.6912e-08
Moskvina and Schmidt (2008)	2,000	–0.0935	1.0134	1,171,691	4.2673e-08
Chen and Liu (2011)	2,000	–0.0922	1.0011	953,990	5.2411e-08
Li et al. (2012)	2,000	–0.2166	0.9897	590,030	8.4741e-08

Calculations are based on an MAF threshold for rare variants of 0.01 and analysis of 74,156 window tests on chromosome 3. The intercepts and slopes of each regression line are also shown in Fig. 1

Standard deviations for Patterson's estimates are very large, but all other methods are quite precise. Estimates from the simulations are associated with slightly larger standard deviations than the majority of the theoretical methods, and therefore several simulations should be performed to obtain an accurate estimate of the required significance levels.

Table 3 gives genome-wide predictions for the effective number of independent tests and the necessary genome-wide significance thresholds for $\alpha_{\text{FW}} = 0.05$, using an MAF threshold of 0.01 for both simulations and the various estimators based on the correlations or eigenvalues of the correlations. We note that since the uncertainty of a prediction increases rapidly outside the range spanned by the explanatory variable, our predictions based on simulations might be expected to be more precise than those based on the correlations and eigenvalues, simply because we are able to obtain estimates using the entire length of chromosome 3. However, we have not provided confidence intervals for these predictions, since the data on chromosome 3 are used multiple times to obtain the points forming the data for the regression line, and hence ordinary confidence intervals would be misleading.

This extrapolation approach is designed to be a computationally conservative way to estimate genome-wide significance thresholds for region-based tests, since whole-genome simulations or eigenvalue calculations are not required. Nevertheless, it may be necessary to repeat similar calculations for each different window definition, for substantially different sample sizes, and possibly for different test statistics and/or choices of weights. Changing window definitions will, of course, alter the number of windows tested. Sample size may make a difference; when more people are sequenced, more genetic variation is identified. Hence, if the window sizes are based on a fixed number of variants, then larger sample sizes imply more windows. However, we have shown in additional calculations that the predicted

significance thresholds are very similar for subsets of size 1,000, 1,500, or 2,000, randomly chosen from the 2,432 individuals (Xu et al. 2014a). Although an alteration in test statistic does not change the number of tests performed, it may change the correlation structure. In particular, an analysis giving equal weight to all rare variants can be expected to display very different correlations from an analysis with weights depending on genomic annotations. The impact of these factors on genome-wide significance thresholds has received little attention to date.

Single-Variant Analyses Combined with Window-Based Analyses

Even if a window-based analysis is planned for variants identified through whole-genome sequencing, most researchers are likely to continue to perform a set of univariate single-SNP analyses of the more common genetic variants. Therefore, it is of interest to know what significance thresholds should be used for a combined analytic strategy including both kinds of tests. In our recent manuscript (Xu et al. 2014a), we used a similar extrapolation approach for univariate tests of common SNPs, window-based tests of rare variants, and both kinds of tests together. The use of a correlation-based or eigenvalue-based approach for the combination strategy would be possible, but since there are large numbers of SNPs, the size of the matrices involved becomes rapidly extremely large. Hence, these analyses used exclusively the simulation-based approach to obtain estimates of the number of independent tests. In fact, since there are many more SNPs than windows (given our definition of 50 rare variants per window), we found that the necessary genome-wide significance threshold was largely driven by the univariate analyses, and we have recommended using a genome-wide threshold near 1e-08 for the SKAT test and a MAF threshold of 0.01 to define rare variants (Xu et al. 2014a).

Exome Sequencing

For rare variant analyses of whole exome sequencing data, most studies have analyzed each gene as a separate unit or window. As a result, many publications use a significance threshold that is adjusted simply for the number of genes, giving a necessary significance threshold near 2.5e-06. Some authors have adjusted for repeated analyses using different phenotypes. To give a few examples, in an exome-wide analysis of pain (Williams et al. 2012), the authors defined a genome-wide significance threshold of $p < 3e\text{-}06$, based on 17,129 tests performed. A study of several insulin phenotypes used a genome-wide threshold of 2.5e-07, adjusting for 19 phenotypes and 10,515 genes that could be tested. For exome sequencing analysis, guidelines have been suggested in a recent review; Do et al. (2012) suggested using a threshold

of approximately 5e-07 to adjust not only for the number of genes but also for several different analyses with different parameter settings. However, they also strongly recommended performing permutation analyses to obtain *p-values* that are accurately adjusted for the number of different analytic strategies and/or phenotypes tested.

Family Studies or Cancer Genome Analyses

Whole-genome sequencing can be used to identify new germline mutations occurring within families, by comparing the genome sequence of parents to sequence in one or more affected children (e.g., Awadalla et al. 2010; Girard et al. 2012). The issue of setting significance thresholds in this context is rarely formally addressed, since the inherently paired design drastically reduces the number of genetic variants of potential interest. If several different genomic regions are identified as interesting, they are usually prioritized for further investigation based on external knowledge and annotation. Given that such investigations often occur in a single family in a very exploratory setting, *p*-values are often ignored.

A conceptually similar study design involves comparing normal and tumor DNA from the same individual or comparing primary and metastatic tumors from the same individual. Again, in such a setting, the paired design reduces the number of genetic variants that might be of interest. In this situation, however, there may be a set of patients with mutations occurring in the same gene or genes. In either case, the paired study design can lead to a dramatic reduction in the number of genetic variants that are considered to be potentially interesting mutations. Nevertheless, an adequate estimate of the effective number of independent tests would be simply the number of regions that pass all quality control criteria and are further investigated.

Identity-by-Descent Considerations

Younger populations are expected to show more linkage disequilibrium (Labuda et al. 1996; Reich et al. 2001), and as a consequence, several methods for identifying causal genes have proposed studying patterns of identity-by-descent (IBD) inferred from genome-wide SNP data (Browning and Thompson 2012; Price et al. 2010; Sham et al. 2009; Allen and Satten 2009). By definition, two individuals share a chromosomal region IBD if this section of their chromosomes is derived from a common ancestor. Since rare variants are more likely to be of recent evolutionary origin, two individuals who carry the same rare variant are more likely to share the surrounding chromosomal region IBD, and furthermore the length of the shared region will be longer for recent variants than for more ancient variants. However, given the uncertainty in estimation of IBD status from genome-wide SNP

data, the power to detect IBD segments shorter than 2 cm tends to be poor (Browning and Thompson 2012; Browning and Browning 2011).

Does information on IBD sharing provide insight into the number of independent tests genome wide? It does allow, perhaps, inference of a lower bound on the number of independent tests that can be reliably inferred through an IBD model (3,000 cM genome/2 cM segment size). However, if sequencing data are available, then IBD estimation may no longer be useful: two individuals who carry the same variant can be assumed to have received it from the same ancestor (the probability of two identical mutations at the same location is very small) (Browning and Thompson 2012).

False Discovery Rates

As discussed above, most studies will be tempted to repeat analyses using several different test statistics, possibly incorporating information on gene structure or regulatory predictions by using different weights for annotated genomic variants. This will inevitably lead to a set of different results for the same genomic region, hence exacerbating the challenge of identifying an appropriate significance threshold.

It may, therefore, be worth considering alternatives to controlling the FWER. For example, control of the false discovery rate (FDR) may give increased power for detecting true associations while still providing some limits on the number of false associations. FDR methods have become very popular in gene expression experiments (Dudoit et al. 2003) where there may be a large number of true associations as well as substantial correlations between different genes (Benjamini and Hochberg 1995; Hochberg and Benjamini 1990; Tusher et al. 2001). Let R be the number of hypothesis tests where the null is rejected at a chosen significance level α. The FDR is defined as the proportion of these R tests where the null hypothesis is true. Since the proportion of false-positive results is controlled, the actual number of falsely rejected null hypotheses can increase as the number of tests increases. There are two noteworthy advantages of this approach. Firstly, the FDR is bounded when there is positive dependence between tests (Benjamini and Yekutieli 2001), and empirical investigations have shown that the FDR can be fairly accurately estimated in the presence of correlations ((Efron 2007), but note also (Schwartzman and Lin 2011)). Hence, repeated testing using different statistics or overlapping window definitions can be undertaken, yet FDR estimates can still be fairly accurate. Secondly, stratified FDR methods, where the test statistics are divided into different subsets with different prior probabilities of a true null hypothesis (Greenwood et al. 2007; Sun et al. 2006; Roeder et al. 2006), may be useful strategy for consideration. It may be possible to incorporate external information on cross-species conservation, predictions of amino acid changes, or other annotation through a stratified FDR approach.

However, using simulations, we recently showed very poor control of type 1 error when using FDR methods (Xu et al. 2014b). Using the data described above from the UK10K consortium, we performed a simulation study where causal rare variants were randomly selected in or near several up to 40 genes on chromosome 3.

Using these causal variants, a continuous phenotype was simulated using a linear model where each additional causal variant could increase the phenotype. Although some of the regions containing causal variants were detected with very high power, in general the sensitivity of detection was low. This is probably due to the fact that most of the causal variants were rare and therefore that the association signals were often weak. However, possibly of more concern, our estimated FDRs tended to be extremely optimistic relative to the simulated truth. For example, when using the Benjamini-Hochberg method for estimating FDR (Benjamini and Hochberg 1995) and when analyzing all variants in all windows, 98 % of the windows selected using an FDR threshold of 0.05 did not contain any causal variants. If we examined nearby windows in strong linkage disequilibrium (a correlation of 0.90 or higher between a causal variant and a variant in the selected window) to see if they contained causal variants, still 94 % of the selected windows were false positives (Xu et al. 2014b). This approach may warrant further consideration, however, if long-range patterns of linkage disequilibrium and their effects on region-based tests of association could be better understood.

Generalizing the Definition of a Window

To a large extent, the choice of the size of windows for region-based analysis can be quite arbitrary, and only a few papers have compared power associated with different window sizes (Yi and Zhi 2011; Zhou et al. 2010; Lin and Tang 2011; Li and Leal 2008; Xu et al. 2012). Recently however, there have been some proposals for optimizing the window size within a larger region of interest. In Brisbin et al. (2012), sliding windows of varying sizes, from 15 variants to over 120 variants, were used to analyze a candidate genomic region, and these results suggested which part of the overall region contained the most signal for association. In contrast, a clustering approach was used by Fier et al. (2012) to select an optimal set of rare variants—not necessarily contiguous—inside a broader region. These approaches, require permutation techniques in order to obtain valid estimates of statistical significance. Hence applying such techniques genome wide could be computationally daunting. Working with a smaller set of annotated variants all linked to a single gene may be more feasible, yet it will still be difficult to decide how to weight or include variants at promoters, enhancers, or other regulatory elements (Zuk et al. 2014).

Population Stratification and Admixture

Rare variant patterns and frequencies show substantial differences between populations. Many variants will be seen in only one population, and others will exist at variable population frequencies. Inflated type 1 error rates can therefore be obtained when performing tests of association without an appropriate correction for population

ancestry that is specific for rare genetic variation; it has been demonstrated that the usual approaches for adjustment with common SNPs are not adequate for rare genetic variants (Zhang et al. 2013; Mao et al. 2013).

Conclusions

Setting appropriate genome-wide significance thresholds for analysis of rare genetic variants is difficult, since the number of possible tests may be bounded only by the imagination (and computational capacity) of the researchers. Variants can be grouped based on physical proximity or by using some external annotation, and new methods are starting to appear that empirically estimate the best subset of variants for a joint test of association. For a given window definition, genome-wide significance thresholds can be estimated and extrapolation from smaller genomic regions to larger ones seems to work well. It may be worth considering alternative conceptual frameworks for multiple testing, possibly based on the ancestral relationships between chromosomes, or a Bayesian perspective that builds in genomic annotations (Stingo et al. 2011; Neath and Cavanaugh 2006). The former tolerates more false-positive associations and hence tends to increase power to find true signals; for the Bayesian context, good specifications of the prior distributions could provide a profitable way to incorporate genomic annotations or relevant variant groupings (Stephens and Balding 2009). However, either of these alternatives leads to an altered way of defining error rates and experimental success.

Acknowledgments The authors are supported by CIHR operating grant MOP-115110 to CG and AC and also by MITACS, the Mathematics of Information Technology and Complex Systems, part of the Canadian Networks of Centres of Excellence program. This study makes use of data generated by the UK10K Consortium, derived from samples from UK10K_COHORTS_TWINSUK (The TwinsUK Cohort) and UK10K_COHORT_ALSPAC (the Avon Longitudinal Study of Parents and Children). A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

References

Allen AS, Satten GA (2009) A novel haplotype-sharing approach for genome-wide case-control association studies implicates the calpastatin gene in Parkinson's disease. Genet Epidemiol 33(8):657–667

Awadalla P et al (2010) Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. Am J Hum Genet 87(3):316–324

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1):289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Statist 29(4):1165–1188

Brisbin A et al (2012) Localization of association signal from risk and protective variants in sequencing studies. Front Genet 3:173

Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. Am J Hum Genet 88(2):173–182
Browning SR, Thompson EA (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190(4):1521–1531
Chen Z, Liu Q (2011) A new approach to account for the correlations among single nucleotide polymorphisms in genome: wide association studies. Hum Hered 72(1):1–9
Cheverud JM (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity (Edinb) 87(Pt 1):52–58
Do R, Kathiresan S, Abecasis GR (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21(R1):R1–R9
Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genome wide association scans. Genet Epidemiol 32(3):227–234
Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Statist Sci 18(1):71–103
Efron B (2007) Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 102(477):93–103
Fier H et al (2012) 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinformatics 28(23):3027–3033
Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genet Epidemiol 32(4):361–369
Gao X et al (2010) Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol 34(1):100–105
Girard SL, Dion PA, Rouleau GA (2012) Schizophrenia genetics: putting all the pieces together. Curr Neurol Neurosci Rep 12(3):261–266
Greenwood CM, Rangrej J, Sun L (2007) Optimal selection of markers for validation or replication from genome-wide association studies. Genet Epidemiol 31(5):396–407
Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Statist Med 9(7):811–818
Labuda M et al (1996) Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. Am J Hum Genet 59(3): 633–643
Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121(1):185–199
Li J, Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95(3):221–227
Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311–321
Li MX et al (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet 131(5):747–756
Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89(3):354–367
Mao X et al (2013) Testing genetic association with rare variants in admixed populations. Genet Epidemiol 37(1):38–47
Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol 32(6):567–573
Neath AA, Cavanaugh JE (2006) A Bayesian approach to the multiple comparisons problem. J Data Sci 4:131–146
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190
Price AL et al (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459–463

Reich DE et al (2001) Linkage disequilibrium in the human genome. Nature 411(6834):199–204
Roeder K et al (2006) Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 78(2):243–252
Schwartzman A, Lin X (2011) The effect of correlation in false discovery rate estimation. Biometrika 98(1):199–214
Sham PC, Cherny SS, Purcell S (2009) Application of genome-wide SNP data for uncovering pairwise relationships and quantitative trait loci. Genetica 136(2):237–243
Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62(1):626–633
Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10(10):681–690
Stingo FC et al (2011) Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann Appl Statist 5(3):1978–2002
Sun L et al (2006) Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. Genet Epidemiol 30(6):519–530
Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98(9):5116–5121
Uhlenback GE, Ornstein LS (1930) On the theory of the Brownian motion. Phys Rev 36(5): 823–841
Williams FM et al (2012) Genes contributing to pain sensitivity in the normal population: an exome sequencing study. PLoS Genet 8(12):e1003095
Wu MC et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93
Xu C et al (2012) Multiple regression methods show great potential for rare variant association tests. PLoS One 7(8):e41694
Xu C et al (2014a) Estimating genome-wide significance for whole genome sequencing studies. Genet Epidemiol 38(4):281–290. doi:10.1002/gepi.21797
Xu C et al (2014b) Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation. Front Genet 5(11):1–13
Yi N, Zhi D (2011) Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35(1):57–69
Zhang Y, Guan W, Pan W (2013) Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 37(1):99–109
Zhou H et al (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26(19):2375–2382
Zuk O et al (2014) Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111:E455–E464

Power of Rare Variant Aggregate Tests

Manuel A. Rivas and Loukas Moutsianas

Introduction

Statistical hypothesis tests are commonly used in all quantitative sciences to make decisions using data from a study. In such tests, a result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a pre-determined significance level. The significance level may reflect a nominal significance level (e.g. $\alpha = 0.05$) and the number of independent tests, n, applied in the study, e.g. in genome-wide association studies (GWAS) of common variants $n \sim 1{,}000{,}000$. The Bonferroni correction is the classical method of adjusting for testing multiple hypotheses. The informal treatment of Bonferroni correction is that for each test that is applied to the data from a study a new level of significance is required to be achieved, $\alpha' = \alpha / n$. For GWAS, the new level of significance commonly employed is $\alpha' = \frac{0.05}{1{,}000{,}000} = 5 \times 10^{-8}$. For a more rigorous treatment please refer to Abdi (2007) and/or Casella and Berger (2002).

A hypothesis test of $H_0: \theta \in \Theta_0$ (the null hypothesis) versus $H_1 : \theta \in \Theta_0^c$ (the alternative hypothesis) is subject to two potential errors: Type I Error or Type II Error. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a Type I Error (false positive). If $\theta \in \Theta_0^c$ but the test decides to accept H_0 (the null), a Type II Error has been made (false negative). The *power* of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false, thereby not committing a Type II Error. The probability of a

M.A. Rivas (✉) • L. Moutsianas (✉)
Wellcome Trust Centre for Human Genetics Research, University of Oxford, Oxford OX3 7BN, UK
e-mail: rivas@well.ox.ac.uk; moutsian@well.ox.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_13

Type II Error occurring is referred to as the *false negative rate*, usually denoted as β. Power is equal to $1 - \beta$ and sometimes referred to as the *sensitivity* of a test.

Power analysis is a very important tool for human genetic studies. It has been widely used to evaluate contemporary study designs (Risch 1990) and propose future ones (Risch et al. 1996). Statistical power crucially depends on the study's sample size and the magnitude of effects, as well as other factors such as the penetrance of the trait and the linkage disequilibrium (LD) between the causal variants and the markers included in the study (Spencer et al. 2009). Careful consideration of the scientific question, taking these factors into account, can yield estimates of the parameters that need to be employed in order to increase the study's power (Altshuler et al. 2008). Fortunately, power analysis has been employed in settings where global investment was necessary to make scientific discoveries (de Bakker et al. 2005). Although evaluating all possible alternative scenarios can be challenging, power analysis can highlight properties of scientific studies which were unrealistic, as well as assess the plausibility of an unexpected result (Sebastiani et al. 2010). Furthermore, power analysis highlights the need for establishing large consortia to bring together cohorts from around the world (Manolio 2009), or initiate large human population-based biobanks to make genetic discoveries (Ollier et al. 2005; Chen et al. 2011).

Advances in DNA sequencing technologies are quickly transforming human genetic studies. Recent DNA sequencing studies of over 2,400 individual exomes and 14,000 samples for 202 targeted genes highlight an abundance of functional variants, most of which were rare (86% with a minor allele frequency less than 0.5%), previously unknown (82%), and population-specific (82%) (Nelson et al. 2012; Tennessen et al. 2012).

This explosion of rare variant catalogs has led to the development of statistical tests designed for the analysis of rare variants. In sequencing studies of complex traits, power to test rare and low frequency variants individually is weak. For example, reports of novel rare variant discoveries required over 30,000 participants for type 1 diabetes (Nejentsev et al. 2009) and over 45,000 participants for inflammatory bowel disease (Rivas et al. 2011). This highlights the challenge of validating the association of rare variants in the context of complex traits. In order to improve power, an approach that is increasing in utility is to combine statistical evidence from several genetic variants in a region. Tests following this approach, commonly referred to as *aggregate tests*, are changing how genetic association studies are undertaken, and are discussed in detail in this chapter and in Chap. 14.

In this chapter we present: (1) study design and variant properties to consider prior to the application of an aggregate test, (2) a comparison of aggregate tests and an evaluation of power and adequate sample size calculations for (3) a rare variant study of dichotomous traits, and (4) a rare variant study of continuous traits. We explore the influence of study size, statistical tests used, variance explained, and the inclusion of null variants in the power to detect associations. Furthermore, we make available simulated datasets along with a program to generate comparisons and evaluate power.

Model Building and Test Selection

Over the past few years, development of statistical methods for identifying rare variant association has been an active area of research.

Prior to applying these methods, one needs to identify the unit to be tested. In the setting of an exome study, the unit may simply be taken to be the gene. In whole genome sequencing studies, and the interrogation of the non-coding genome, the unit to be tested can be more difficult to define. Various variant groupings may be applied, which will have a direct impact on the power to detect associations. Once the unit to be tested has been defined, the null hypothesis about the variants within that unit should be formed. Annotation information is important in clearly stating the null hypothesis. For example, in the setting of an exome sequencing study, our unit of interest is the gene, and our null hypothesis may reflect the subtypes of variants within a gene which are of higher potential impact on the trait, e.g. protein altering variants:

H_0 : Rare protein altering variants (missense, nonsense, coding indels, splice) in *APOB* are not associated with early-onset myocardial infarction.

Annotation

Functional annotation is relevant for the application of aggregate statistical tests (Please see Chap. 7 and Fig. 1 for a more in-depth treatment of annotation). For example, an analyst may consider a more targeted approach to test for association by including variants annotated as truncating (Herman et al. 2012; Ruark et al. 2012; Hopper et al. 1999; Dodé et al. 2003) or commonly referred to as Loss of Function (LoF) (MacArthur et al. 2012). These variant types are usually singletons, private variants (Tennessen et al. 2012), and are likely to have a similar impact on protein function (for some exceptions see Ruark et al. 2012 and Isidor et al. 2011). In the LoF setting one should apply a test that scans for an overall shift in the number of rare variant copies in cases compared to controls.

On the other hand, consider applying an aggregate test to all variants discovered in a gene sequencing experiment. In a more inclusive approach, an analyst may want to include variants predicted to be regulatory, synonymous (silent) substitutions, missense, nonsense, as well as coding indels. Alternatively, one may want to remove variants from the association analysis that are unlikely to contribute to signal, e.g. silent substitutions; as we will see later, including variants with no impact on trait value may quickly impact power to detect association. Let's take into consideration the test one may want to employ. In 2,400 samples, a gene, on average, has approximately 20–30 coding variants. It is very likely that some of these variants will have no impact on trait value, whilst some may have protective or deleterious effects on the trait (in the setting of a dichotomous trait).

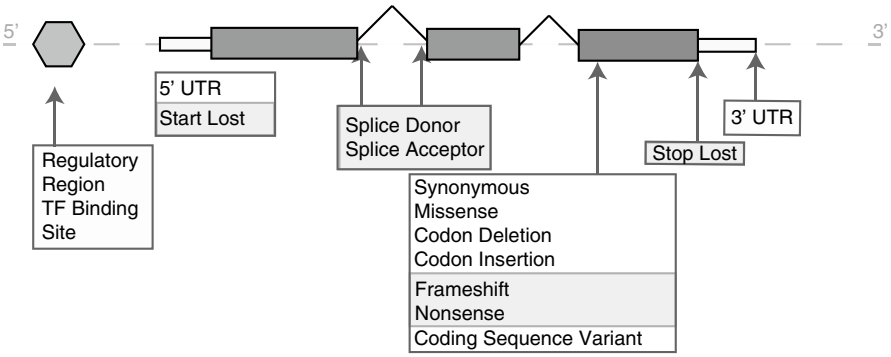

Fig. 1 Variant annotation. Diagram of variant annotation in rare variant studies. Truncating variants: Frameshift (*light gray* annotation box), Nonsense, Splice Donor, Splice Acceptor, Stop Lost, Frameshift, Start Lost and Protein altering variants: Truncating variants + Missense, Codon Deletion, Codon Insertion, Stop Lost, Start Lost. *Dark boxes* in figure are protein coding exons of a gene, *white boxes* of figure are the untranslated exons of a gene and *light gray* hexagon represents the regulatory region of a gene. Adapted from ENSEMBL

Classes of Aggregate Tests

The choice of the test or tests to apply depends on the alternative hypothesis, as different classes of tests address different hypotheses. To highlight some of the differences which may exist between them consider two tests which have been used in the literature: (1) FRQWGT (Madsen and Browning 2009) and (2) C-alpha (Neale et al. 2011). For a more complete list of aggregate tests available see Table 1 and for additional studies comparing aggregate tests see Basu and Pan (2011) and Asimit and Zeggini (2010).

FRQWGT is a test that scans for an excess of rare variants in cases compared to controls; weights for FRQWGT are assigned to variants according to their corresponding allele frequency. C-alpha, on the other hand, scans for a signal of overdispersion of the distribution of rare variants in a unit. The alternative hypothesis for C-alpha is that at least one of the variants is not binomially distributed according to the null parameter (null parameter for a case–control study of equal number of cases and controls is 0.5), whereas the alternative hypothesis for FRQWGT is that there is a total excess of rare variants in cases versus controls. Figure 2 a highlights the signal that aggregate tests considered as "collapsing" are well suited for, i.e. detecting a shift in the mean of the distribution. Figure 2b highlights the strength of aggregate tests in capturing dispersion signal, i.e. change in variance.

Table 1 A selection of available aggregate tests for rare variant analysis in PLINK/SEQ

Aggregate test	Description	Directionality	References
FRQWGT	Frequency-weighted test, in spirit of Madsen-Browning. Collapsing	Mean-based	Madsen and Browning (2009)
KBAC	Rare variant test in the presence of misclassification and gene interaction. Collapsing	Mean-based	Liu and Leal (2010)
SUMSTAT	Sum of single-site statistics. Dispersion	Variance-based	PLINK/SEQ
UNIQ	Count of case-unique rare alleles. Collapsing	Mean-based	PLINK/SEQ
VT	Variable threshold test. Collapsing	Mean-based	Price et al. (2010)
BURDEN	Excess of rare alleles in cases compared to controls. Collapsing	Mean-based	PLINK/SEQ
C-alpha	C-alpha test. Dispersion	Variance-based	Neale et al. (2011)
SKAT	Kernel based regression method and score-based variance component test. Dispersion	Variance-based	Wu et al. (2011)
SKAT-O	Mixture of collapsing and dispersion test	Variance-based	Lee et al. (2012)

Power of Rare Variant Aggregate Tests for Case–Control Association Studies

In this section, we focus on the power of aggregate tests to detect signals of association in dichotomous (binary) traits, such as in a case–control design. The power to detect signals in quantitative (continuous) traits is addressed in the section "Power of Rare Variant Aggregate Tests for Continuous Traits". We have used different simulation approaches to obtain an estimate of the power of aggregate tests to detect associations. Our choices of models and parameters were directly or implicitly informed by our current understanding and beliefs regarding the genetic architecture of complex disease, with sample sizes in the range of those observed in contemporary whole-genome sequencing studies. It is important to highlight that any estimate of power is affected by the choice of a wide range of parameters regarding both

- the way data is simulated, and
- the way data is tested.

For the former, these include the frequency of the risk alleles, the magnitude of effects, the number of variants in a unit, and the percentage of them that are causal. For the latter, these include annotation and frequency filters on the variants to test, which will have to be carefully thought and selected in real studies too.

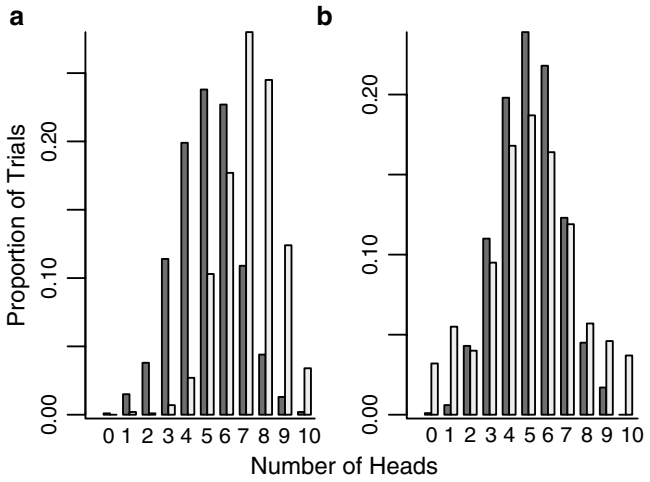

Fig. 2 Signal of aggregate tests. We use coin tosses as in Neale et al. (2011) to highlight the source of signal for aggregate tests. (**a**) *Black bars* represent the distribution of total heads from fair coin tosses (null variants), *white bars* represent the distribution of total heads from loaded coins (risk variants). We observe a shift in the mean of the distribution coming from the loaded coins compared to fair coins. (**b**) *Black bars* represent the distribution of total heads from fair coin tosses (null variants), *white bars* represent the distribution of total heads from a 10:80:10 mixture of loaded coins ($p = 0.1$, favoring tails), fair coins ($p = 0.8$), and loaded coins ($p = 0.1$, favoring heads). In this scenario we clearly observe that there is no shift in the mean; however, there is a shift in the variance of the distribution indicating *overdispersion*. In Neale et al. (2011) and Wu et al. (2011) the authors highlight how a burden test may fail to detect signal in the setting of a mixture of null variants and risk variants

In the section "A Simulation Study Based on 1000 Genomes Project", we estimate the power for a balanced dataset consisting of 2,000 samples, of which 1,000 samples have the trait and the rest are controls from the general population. In the section "A Simulation Study for Sample Size Calculations: Dichotomous Traits", we discuss how estimates of power change by sample size and suggest numbers which may be required to attain this power under various hypotheses.

A Simulation Study Based on 1000 Genomes Project

While simulation efforts to investigate the power of rare variant aggregate tests commonly introduce fixed combinations of the number of causal variants and effect sizes (e.g. Basu and Pan 2011; Ladouceur et al. 2012), we decided to allow these to vary and fix the percentage of the total variance explained by each locus instead.

For the calculation of the variance explained by each locus, we employed the method followed by So et al. (2011). This approach assumes a multifactorial liability threshold model (Falconer 2007). According to this model, the overall liability to disease is a continuous function of a number of contributing genetic variants with various effects, as well as of other risk factors. It is assumed to follow a standard normal distribution. A fraction of individuals from the general population, whose liability exceeds a certain threshold, will develop the disease. Each variant which has been assigned an effect will explain part of the variance in liability, with the total variance in variability explained by a unit being equal to 1 %. Throughout this chapter, and unless otherwise stated, the term "variance explained" should be taken to mean variance in liability.

To estimate the percentage of variance explained by a single variant, the model takes three parameters as input: disease prevalence, frequency of the risk variant in the general population, and genotype relative risk. Assuming independence between the different risk variants at each unit, the total percentage of the variance explained is taken to be the sum of the variance explained by each of the variants with introduced effects. Simulated datasets were generated using Hapgen2 (Su et al. 2011), which employs the Li and Stephens model (Li and Stephens 2003). This is a haplotype reshuffling approach, where simulated (unobserved) haplotypes are assumed to be an imperfect mosaic of actual (observed) ones, generated using a Hidden Markov Model. Hapgen2 introduces deleterious effects by over-sampling haplotype segments which contain the variants to which the effects are introduced, based on the relative risk assigned to them.

For the present study, the reference panel consisted of 379 European individuals from the 1000 Genomes project (1000G) (Consortium 2010), and was annotated using CHAoS (http://www.well.ox.ac.uk/~kgaulton/chaos.html). Data was simulated for 1,000 cases and 1,000 controls at eight genes implicated in type 2 diabetes (T2D): *ING1*, *RAPH1*, *GPATCH2*, *CACNA2D2*, *KLK11*, *DCAF16*, *LTBR* and *PLCL1*. Datasets were simulated under three alternative models, all of which explain 1 % of the variance in liability:

- M1: Only deleterious effects in a unit
- M2: An equal mixture of deleterious and protective effects in a unit
- M3: Only protective effects in a unit

The effects are assigned at random to a subset of the annotated variants at each locus. These include synonymous, missense, nonsense, splice, 3′ UTR and 5′ UTR variants. Deleterious effects were assigned to variants with a frequency of up to 1 %, whereas protective ones to variants in [0.5 %, 5 %]. The relative risk introduced to each variant is drawn at random from [1, 5], as stronger effects were deemed unlikely to occur frequently. For protective variants, the relative risk was drawn at random from the [0.2, 1] interval. Disease prevalence was assumed to be 8 %, consistent with recent estimates for T2D (personal communication).

Results

We have compared representative tests of both the mean- and the variance-based approaches across the three models. We have also included SKAT-O (Lee et al. 2012) to the comparison, which is a mixture of the two approaches. The purpose of this simulation study, rather than to serve as an exhaustive comparison between all available tests, is to showcase the differences in performance one may expect between the two main subsets of tests (mean- and variance-sided) under different scenarios. Even when only deleterious effects are introduced (M1), we find C-alpha to perform better than its mean-based alternative FRQWGT. This somewhat surprising result is likely to reflect the loss of power that mean-based tests face in the presence of null variation. Power is greater than 15 % for a study of a dichotomous trait with 1,000 cases and 1,000 controls at $\alpha = 0.01$ for variance explained = 1 %. As expected, we find that while C-alpha and SKAT-O maintain stable power across the three models, the power drops sharply for the collapsing approach. The variance in the results, as observed in Fig. 3, highlights the sensitivity of the test to unit-specific parameters such as the number of variants being tested.

We have demonstrated a comparison of representative rare variant aggregate tests from groups with a different design philosophy, for case–control analysis under a few alternatives regarding the direction of the effect of the risk variants. Other tests of the same group are broadly expected to have comparable power, but see Moutsianas et al. (2014) for a more nuanced treatment. For instance, SKAT (Wu et al. 2011) and C-alpha (Neale et al. 2011) are expected to have similar power across all alternatives, as SKAT is a generalization of C-alpha. To achieve optimal power an analyst should carefully consider the alternatives they wish to test.

A Simulation Study for Sample Size Calculations: Dichotomous Traits

To determine the adequate sample sizes needed to achieve 80 % power to detect association for case–control analysis of rare variants, we conducted a simulation study focused on SKAT and SKAT-O.

Dataset

We used the SKAT package that provides a dataset which contains a haplotype matrix of 10,000 haplotypes over a 200 kb region (see the section “Links”). The haplotypes were simulated using a calibrated coalescent model that mimicks LD structure of European ancestry. We carried out sample size calculations using the haplotypes with the following parameters:

1. Subregion length = 3 k base pairs (default).
2. Prevalence = 0.08, a value of disease prevalence for T2D (personal communication).

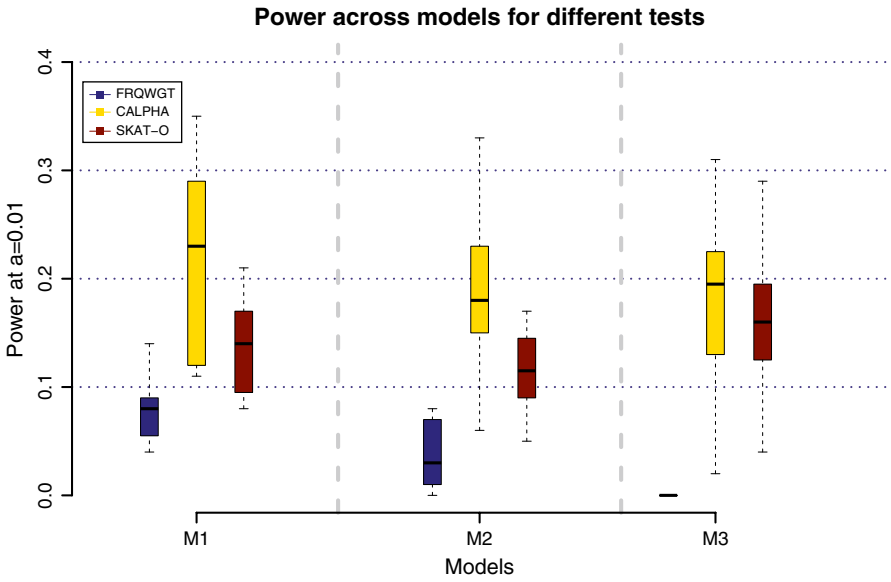

Fig. 3 Power comparison for different aggregate tests. Comparison of power for a selected subset of the tests implemented in PLINK/SEQ across three different models: M1. Only deleterious effects in a unit, M2. An equal mixture of deleterious and protective effects in a unit, M3: Only protective effects in a unit. Surprisingly, C-alpha and SKAT outperform FRQWGT for model M1. We postulate that this may be due to a proportion of variants tested being neutral. Hence, in the presence of a mixture of null and deleterious variants, power loss for C-alpha and SKAT may be slower than for mean-based alternatives

3. For dichotomous traits, we set the maximum effect size to OR = 4. For SKAT-O we set r.corr (the ρ parameter of new class of kernels) to 2 to allow grid search.
4. $\alpha = 2.5 \times 10^{-6}$ for exome-sequencing datasets (n is assumed to be 20,000; approximate number of genes tested).

Figure 4 shows that as the proportion of null variants included in the test increases, the sample sizes required to achieve power to detect association for a unit also increases (Ladouceur et al. 2012). A 3 kb locus with 10 %, 20 %, 50 %, and 100 % of the variants as causal, with maximum effect size of 4, requires 116 k, 20 k, 6 k, and 2 k, respectively. In other words, if we were to aggregate all variants discovered in a 2 k sample sequencing experiment in a 3 kb segment, and *all* variants in the 3 kb segment were causal and associated with phenotype, then we would achieve 80 % power to detect association. If we subset the variants in a 3 kb region, we are effectively decreasing the size of our subregion. If we change the subregion size to 500 base pairs and use the same requirements described above, it will require over 6,000 cases and 6,000 controls when *all* variants in the segment are causal.

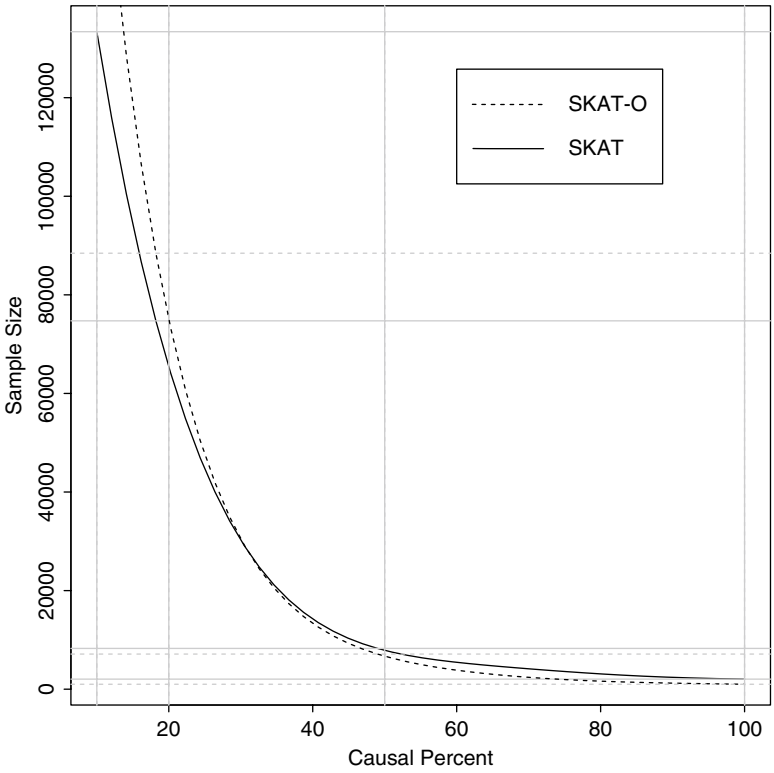

Fig. 4 Sample size required to achieve 80 % power to detect association for aggregate tests

Power of Rare Variant Aggregate Tests for Continuous Traits

In this section we turn our attention to rare variant aggregate tests for continuous traits. More specifically, we perform power analysis for SKAT (variance-based) and SKAT-O (a mixture of mean- and variance-based tests). We focus on the power of aggregate tests to detect signals of association for different levels of phenotypic variance explained for a locus in the section "A Simulation Study to Evaluate Power". Next, we calculate the adequate sample sizes require to achieve significance for an exome study in the section "A Simulation Study for Sample Size Calculations".

A Simulation Study to Evaluate Power

Typically, the power function of a test will depend on the sample size n. If n can be chosen for a study design, power analysis might help determine if the sample size is appropriate for the study and place bounds on the effects that may be detected. In the setting of continuous traits, an analyst will want to evaluate the phenotypic

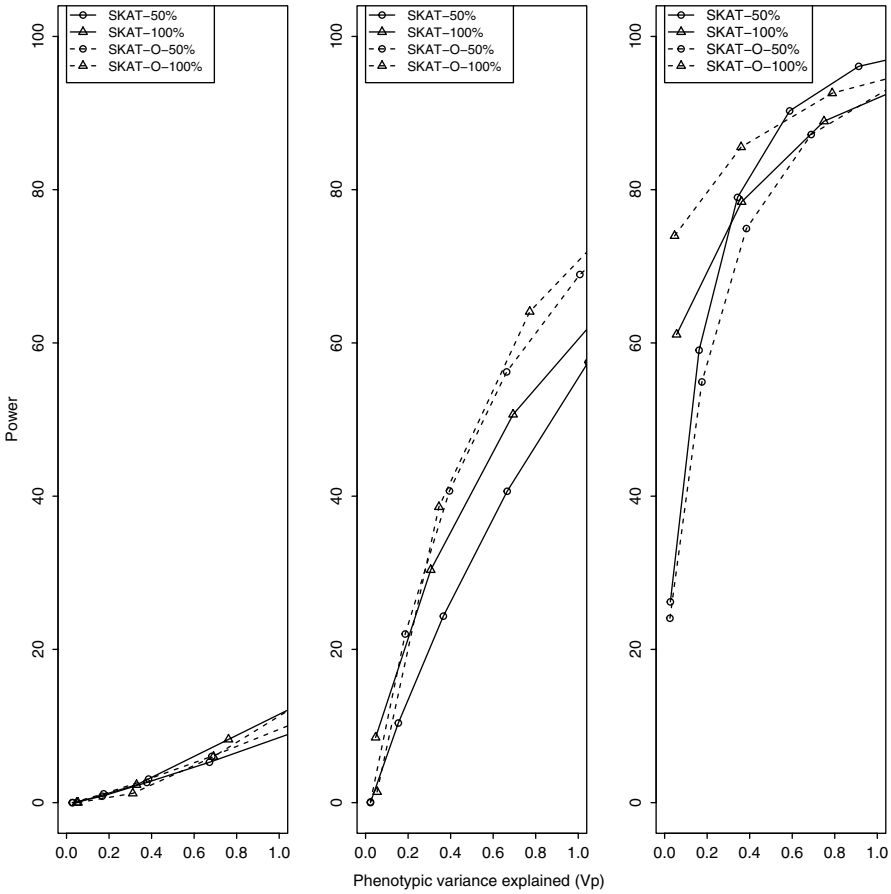

Fig. 5 Power of rare variant aggregate tests for continuous traits at different levels of phenotypic variance explained. Power for different levels of phenotypic variance explained is evaluated for sample size n = (**a**) 1 k, (**b**) 10 k, and (**c**) 100 k (from *left* to *right*). *Dashed lines* and *solid lines* are power estimates for SKAT-O and SKAT, respectively. *Circle* and *triangle point symbols* represent power evaluated at percent causal variants = 50 % and 100 %

variance explained by a locus that may be detected in a study of n samples (Fig. 5). We present results from a simulation experiment with similar parameters as the case–control simulation study, i.e.:

1. Subregion length = 3 k bp
2. Causal percentage = 50 % and 100 %
3. For continuous traits, we set the maximum effect size to be $\beta = 1.6$. For SKAT-O we set r.corr = 2 to allow grid search and maximize between a dispersion and a collapsing test.
4. $\alpha = 2.5 \times 10^{-6}$ for exome-sequencing datasets.

We vary the phenotypic variance explained for the subregion from 0.1 % to 1 %, and evaluate power for sample sizes n = 1 k, 10 k, 100 k. For a locus explaining 0.5 %

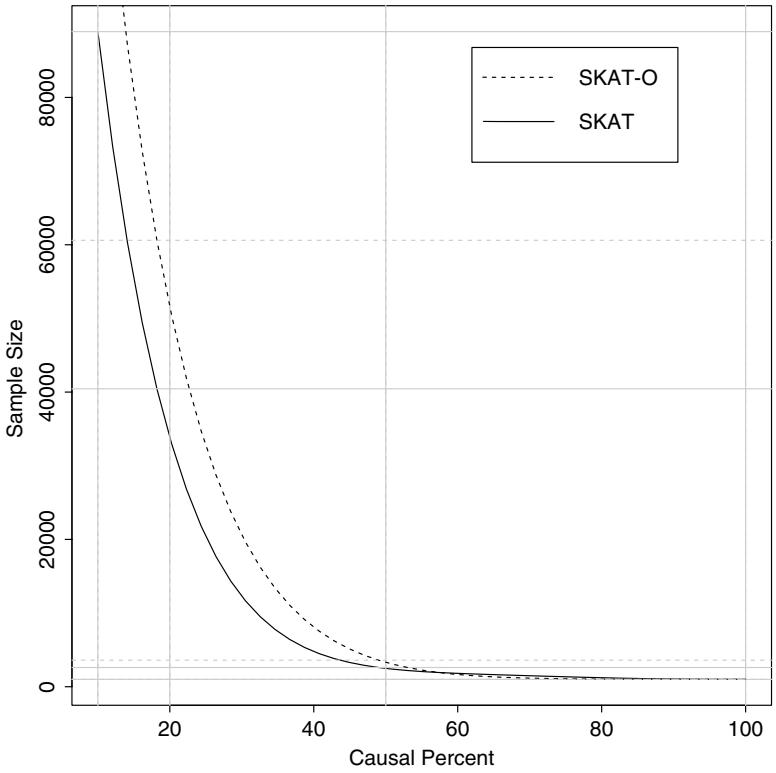

Fig. 6 Power to detect association as a function of sample size. Adequate sample size required to achieve 80 % power is shown for varying levels of causal percent variants in a 3 kb locus tested for association to a continuous trait using SKAT (*solid line*) and SKAT-O (*dashed-line*)

of the phenotypic variance (a substantial but realistic amount, compared to reported findings, e.g. *PCSK9* and LDL levels Cohen et al. 2006) we achieve 9 %, 58 %, and 100 % power, respectively, to identify association using SKAT with only causal variants being tested.

A Simulation Study for Sample Size Calculations

To illustrate the sample sizes needed to achieve 80 % power to detect association for the analysis of rare variants, we focus on a simulation study of rare variants using SKAT and SKAT-O.

Using the same haplotype dataset and the same parameters as in the section "A Simulation Study for Sample Size Calculations: Dichotomous Traits", we find that over 40,000 samples are required to detect signal when 20 % of the variants are causal and 4,000 samples are required to detect signal when 50 % of the variants in a 3 kb region are causal (Fig. 6).

Conclusion

Rare variant aggregate tests have an important role to play in the analysis of data from re-sequencing studies, since the power of single variant tests to detect associations for rare variants is weak. The design of such tests is an active area of research, but many are readily available. No single aggregate test achieves optimal power. Power analysis of rare variant aggregate tests highlights the difference between mean- and variance-based tests. We estimate from our simulation study that at least 20 k and 4 k samples will be required to achieve power to detect associations at exome-wide level of significance for rare variant studies of dichotomous and continuous traits respectively, based on parameters we chose to reflect our experience with exome sequencing studies of complex traits. Our observations are in line with recent reports in literature (Moutsianas et al. 2015; Zuk et al. 2014). Alternative study design strategies (not discussed in this chapter) may improve power to detect association, e.g. selecting samples from the tails of the distribution (Guey et al. 2011) or analyzing multiple phenotypes (Korte et al. 2012). Well-formed hypotheses and clearly defined testing units should be carefully considered before applying an aggregate test to re-sequencing data.

Links

Here are some useful links to software and simulated data which can be used for the analysis of power or rare variant aggregate tests:

PLINK/SEQ:
http://atgu.mgh.harvard.edu/plinkseq/
T2D HapGen Simulated dataset:
http://www.well.ox.ac.uk/~rivas/p1redo.tar.gz
SKAT Package:
http://cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf

Acknowledgements We would like to thank Mark McCarthy for invaluable feedback and the editors Andrew Morris and Eleftheria Zeggini.

References

Abdi H (2007) Bonferroni and šidák corrections for multiple comparisons. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Thousand Oaks, Sage

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322(5903):881–888

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35(7):606–619

Casella G, Berger RL (2002) Statistical Inference (2nd edn), Duxbury, Pacific Grove, CA
Chen Z, Chen J, Collins R, Guo Y, Peto R, Wu F, Li L (2011) China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 40(6):1652–1666
Cohen JC, Boerwinkle E, Mosley Jr TH, Hobbs HH (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 354(12):1264–1272
Consortium GP (2010) A map of human genome variation from population scale sequencing. Nature 467(7319):1061–1073
de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nat Genet 37(11):1217–1223
Dodé C, Levilliers J, Dupont J-M, De Paepe A, Le Dû N, Soussi-Yanicostas N, Coimbra RS, Delmaghani S, Compain-Nouaille S, Baverel F et al (2003) Loss-of-function mutations in FGFR1 cause autosomal dominant Kallmann syndrome. Nat Genet 33(4):463–465
Falconer DS (2007) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann Hum Genet 29(1):51–76
Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B et al (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35(4):236–246
Herman DS, Lam L, Taylor MR, Wang L, Teekakirikul P, Christodoulou D, Conner L, DePalma SR, McDonough B, Sparks E et al (2012) Truncations of titin causing dilated cardiomyopathy. N Engl J Med 366(7):619–628
Hopper JL, Southey MC, Dite GS, Jolley DJ, Giles GG, McCredie MR, Easton DF, Venter DJ (1999) Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and BRCA2. Cancer Epidemiol Biomark Prev 8(9):741–747
Isidor B, Lindenbaum P, Pichon O, Bézieau S, Dina C, Jacquemont S, Martin-Coignard D, Thauvin-Robinet C, Le Merrer M, Mandel J-L et al (2011) Truncating mutations in the last exon of notch2 cause a rare skeletal disorder with osteoporosis. Nat Genet 43(4):306–308
Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M (2012) A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44(9):1066–1071
Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PloS Genet 8(2):e1002496
Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–75
Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165(4):2213–2233
Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6(10):e1001156
MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828
Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384
Manolio TA (2009) Cohort studies and the genetics of complex disease. Nat Genet 41(1):5–6

Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, The GoT2D Consortium, McVean G, Boehnke M, Altshuler D, McCarthy MI (2015) The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet 11(4): e1005165

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925):387–389

Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean PS, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D et al (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337(6090):100–104

Ollier W, Sprosen T, Peakman T (2005) Uk biobank: from concept to reality. Pharmacogenomics 6(6):639–646

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832–838

Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46(2):222

Risch N, Merikangas K et al (1996) The future of genetic studies of complex human diseases. Science (AAAS-Weekly Paper Edition) 273(5281):1516–1517

Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagace C, Neale B, Lo KS, Schumm P, Torkvist L, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, Lettre G, Franke A, D'Amato M, McGovern DP, Cho JH, Rioux JD, Xavier RJ, Daly MJ (2011) Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43(11):1066–1073

Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, Brough R, Rodrigues D, Renwick A, Seal S, Ramsay E, Duarte S, Rivas M et al (2013) Mosaic ppm1d mutations are associated with predisposition to breast and ovarian cancer. Nature 493(7432):406–410

Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH et al (2010) Genetic signatures of exceptional longevity in humans. Science 10:1126

So H-C, Gui AH, Cherny SS, Sham PC (2011) Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. Genet Epidemiol 35(5):310–317

Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5(5):e1000477

Su Z, Marchini J, Donnelly P (2011) Hapgen2: simulation of multiple disease snps. Bioinformatics 27(16):2304–2305

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES (2014) Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci 111(4):455–464

Replicating Sequencing-Based Association Studies of Rare Variants

Dajiang J. Liu and Suzanne M. Leal

Background

Currently there is worldwide interest in studying the role of rare genetic variants in the etiology of complex traits. Compared to common variants which are for the most part functionally neutral, it should be easier to interpret identified rare variant association signals. Many sequencing studies have already provided evidence for the involvement of rare variants in the etiology of complex traits, including colorectal adenomas, plasma lipid levels, blood pressure, and diabetes-related quantitative traits (Cohen et al. 2004, 2006; Ji et al. 2008; Romeo et al. 2007, 2009; Huyghe et al. 2013).

Indirect association mapping using tagSNPs is underpowered to detect associations with rare variants due to the weak correlations between higher-frequency tagSNPs and rare variants (Li and Leal 2008). Instead direct association mapping should be applied, where variants are discovered and directly tested for associations. With the rapid development of cost-effective next-generation sequencing (NGS) technologies, sequence-based genetic association studies of complex traits have been made possible. Although only sequencing-captured genetic regions can be cost and time effective, the high cost of sequencing is still a concern. Additionally, NGS can have higher error rates than those obtained from genotyping arrays (Harismendy et al. 2009). Therefore, in order to replicate associations to candidate genetic regions, it is

D.J. Liu
Department of Public Health Sciences, College of Medicine,
Pennsylvania State University, Hershey, PA 17033, USA

S.M. Leal (✉)
Department of Molecular and Human Genetics, Center for Statistical Genetics,
Baylor College of Medicine, Houston, TX 77030, USA
e-mail: suzannemleal@gmail.com

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_14

of interest to explore alternative technologies to NGS, such as using customized genotyping or exome arrays.

Recently, in an effort to cost effectively obtain data on rare variants, an exome array was developed based upon a collation of variant sites identified from NGS of ~12,000 exomes and genomes. Missense variants which were observed at least three times in two different cohorts and nonsense and splice sites which were observed at least two times in two different cohorts were selected for inclusion on the exome array. The exome array also contains additional content, e.g., ancestry informative markers (AIMs) and SNPs which are associated with complex traits. Compared to sequencing targeted regions, the exome array can be much more cost-effective. In addition, genotyping is a more mature technology and can potentially have better accuracy compared to NGS. However, by design, very rare variants, e.g., singletons, were not included on the exome array and therefore cannot be integrated.

In addition to technology advancements in generating sequence data, there has been a plethora of new statistical association methods developed specifically to analyzing rare variant data. Based upon the observation that analyzing rare variants individually or performing multivariate tests can be either underpowered or numerically unstable, gene-level association tests have been proposed. Instead of analyzing each variant by itself, the unit of analysis is a gene or another functional unit. For example, it is possible to test for an association between the trait of interest and a "burden" of rare variants (i.e., the total number of rare variants in a gene region). This gene-level association analysis strategy avoids the repeated analysis of multiple very low-frequency variants and can therefore be more powerful than traditional methods which are used to analyze common variants.

Similar to the analysis of common variants, it is necessary to confirm the association signal in the original study (stage 1 sample) using an independent dataset (stage 2 sample), in order to guard against spurious associations. In this chapter, strategies for replicating gene-level associations are discussed and the plausibility of using genotyping or sequencing in replication studies from a combined genetic epidemiology and population genetics perspective is explored. A comparison of replication strategies is performed through simulation studies and also by association analysis of energy metabolism traits and sequence data from the Dallas Heart Study (DHS) on the *ANGPTL3* [MIM 604774], *ANGPTL4* [MIM 605910], and *ANGPTL5* [MIM 607666] genes. In addition, we also point the reader to a few resources that can be beneficial for designing replication studies.

Replication Strategies for Gene-Level Association Test

For mapping rare variants, gene-based tests are usually performed, which aggregate multiple rare variants in a given gene region and test for their associations with the trait of interest. Representative methods include the combined multivariate and collapsing (CMC) test (Li and Leal 2008), gene- or region-based analysis of variants of

intermediate and low frequency (GRANVIL) test (Morris and Zeggini 2010), weighted sum statistic (WSS) (Madsen and Browning 2009), kernel-based adaptive cluster (KBAC) (Liu and Leal 2010a), and sequencing kernel association test (SKAT) (Wu et al. 2011).

Three Possible Replication Strategies

To replicate significant findings in stage 1 studies, three different strategies can be used. As a first strategy, only the variants at the nucleotide sites uncovered from the original sample are followed up. Using this strategy, novel nucleotide sites that are present only in the stage 2 sample, but not in the stage 1 sample will not be incorporated in the replication study. This constitutes a replication in a "strict" sense, i.e., both the gene region and the variants uncovered in the stage 1 sample are followed up in the replication sample. When only variants uncovered in the stage 1 sample are of interest, genotyping is sufficient. We will refer to this replication strategy as *variant based*. Second, given the cost-effectiveness and the fact that many rare coding variants are represented on the exome array, it is possible to follow up the candidate region by genotyping the replication samples using exome arrays. We call this *exome-array-based* replication. Compared to the first strategy, exome-array-based replication can be much more cost-effective, but it can be less powerful if causal variants identified in the stage 1 study are not included on the exome array. Finally, the third and the most comprehensive strategy is to follow up the entire gene region identified in the stage 1 sample. This can be performed by sequencing the gene region. For this design, analysis of the stage 2 sample is not restricted to the nucleotide sites uncovered in stage 1. Variants from novel sites in the replication sample are also assessed for their associations with the phenotype of interest. We will refer to this design as *sequence-based* replication. With this strategy, sequencing the target gene in the stage 2 sample is necessary. While it is clear that exome-array-based replication is the most cost-effective strategy, followed by variant- and sequence-based replication, their power for replicating genuine associations needs to be evaluated.

Factors That Influence the Power for Replication

The power to replicate for variant-based replication is mainly dependent on the percentage of causal variants sites that were uncovered for the gene region in the stage 1 sample, while for exome-array-based replication, power is driven by how many causal variant sites are available for genotype on the exome array. If the stage 1 sample is small, there can be an advantage to sequence- or exome-array-based replication, since many low-frequency variants may not have been observed. However, the difference between variant- and sequence-based replication strategies will diminish if a majority of causal variants can be uncovered in stage 1.

The proportion of causal variants identified in the stage 1 sample can be affected by (1) the sample size in the stage 1 study, (2) the genetic architecture of the trait of interest, (3) as well as the sequencing error rate. In general, the proportion of causal variants identified will increase as a larger sample is sequenced. For diseases which are driven by low-frequency variants [e.g. variants with minor allele frequency (MAF) < 5 %], sequencing 200 cases will identify >99 % of the low-frequency causal variants. However, if a trait is mainly driven by very rare mutations (e.g., variants with MAF < 0.1 %), a much larger number of samples will need to be sequenced. Finally, the actual detection of causal variants may also be affected by sequencing errors and false-negative calls may lead to the under-detection of causal variants and therefore a loss of power.

Additional factors which play a role in the ability to replicate regardless of the type of replication used include effect sizes of the variants within a region. The winner's curse may cause an overestimation of the effect size in the stage 1 study due to sampling and this should be taken into consideration when designing a replication study. Additionally the sample size for the stage 2 replication study will also play an important role in the ability to replicate.

Power for Replicating Sequence-Based Association Studies: A Mathematical Formulation

We are interested in the power to replicate, i.e., the probability of successfully replicating an association identified in stage 1 using an independent sample. To define necessary notations, the test statistics used for the stage 1 and sequence-based stage 2 studies are denoted by T^{S1} and T^{seq}, respectively. Specifically, the following probability $P_{H^A}\left(\left|T^{seq}\right| > z_{1-\alpha_{S2}/2} \middle| \left|T^{S1}\right| > z_{1-\alpha_{S1}/2}\right)$ is considered, where α_{S1} and α_{S2} are significance levels used for the stage 1 and the stage 2 replication study.

Similarly, the power for variant-based replication is given by $P_{H^A}\left(\left|T^{var}\right| > z_{1-\alpha_{S2}/2} \middle| \left|T^{S1}\right| > z_{1-\alpha_{S1}/2}\right)$ but unlike for sequence-based replication, the test statistics T^{var}, T^{S1} are not conditionally independent. Under the alternative hypothesis, the distribution of T^{var} depends on K which is the set of rare variant sites uncovered in stage 1.

For notational convenience, the ratio of total frequencies of uncovered rare variants to the total frequencies of all locus rare variants (including those that are not uncovered) is denoted by

$$f_{MAF} = \sum_{s \in K} P\left(X_i^s = 1\right) / \sum_{s=1}^{S} P\left(X_i^s = 1\right)$$

In addition, the ratio

$$f_{PAR} = \sum_{s \in K \cap C} P\left(X_i^s = 1 \middle| Y_i = 1\right) / \sum_{s \in C} P\left(X_i^s = 1 \middle| Y_i = 1\right)$$

represents the proportion of locus population attributable risk (PAR) that can be explained by the uncovered causal variants in stage 1. This is asymptotically

equivalent to the epidemiological definition of PAR which is the reduction of disease incidence rate that would be observed if the population were unexposed, i.e., if there were no carriers of locus causative variants.

Simulation Comparison for Three Replication Strategies

In order to conduct simulations that maximally reflect the true variant frequency spectrums, we use allele frequencies estimated from the Exome Sequencing Project (ESP) (Tennessen et al. 2012). Specifically, minor allele frequency information from 15,585 genes obtained on European Americans from the Exome Variant Server (EVS) (http://evs.gs.washington.edu/EVS/) were utilized. Genotypes for each individual were simulated according to the estimated variant frequency in EVS, assuming no linkage disequilibrium between variant sites and that the genotypes at each site are in Hardy–Weinberg equilibrium within the general population. In addition, exome-array genotypes were simulated according to variant sites that were incorporated in the exome array (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). The detailed frequency spectrum of coding sequence variants can be found in the article of Tennessen et al. (2012).

Phenotypic effects of rare non-synonymous (NS) variants are assumed independent of their allele frequencies (Pritchard 2001). Fifty percent of the rare NS variants (with MAF ≤ 0.01) were randomly picked to be causal and affect the binary phenotype of interest. Based upon surveys of multifactorial diseases (Bodmer and Bonilla 2008), the following phenotypic model was considered. The genetic effects of causal variants are inversely correlated with their MAFs. It is assumed that causal variants with the smallest (or largest) MAFs (i.e., $p_{\min}$ or $p_{\max}$) have the largest (or smallest) log odds ratio (log-OR) of $\beta_{\max}$ (or $\beta_{\min}$), respectively. For a causal variant with MAF p_i, the log-OR follows the interpolation relation:

$$\beta_i = \beta_{\max} + \left(\beta_{\max} - \beta_{\min}\right) / \left(p_{\max} - p_{\min}\right) \times \left(p_i - p_{\min}\right), i \in C$$

The ORs for causal variants thus satisfy an exponential relationship with their MAFs. A choice of $\beta_{\max} = \log(10), \beta_{\min} = \log(2)$ was used. A baseline penetrance of 0.01 is assumed, which gives $\beta_0 = \log\left(0.01/\left(1-0.01\right)\right)$. In order to mimic the design of exome chips, we only analyzed variant sites that are designed to be genotyped on the exome array.

Under the simulation framework, we first compare the rare variant discovery rates. All comparisons were made under the assumption that sequencing data is of perfect quality (Table 1). When sequencing is not perfect, the fractions of uncovered variants will be lowered by the false-negative rate. At the same time, a portion of observed variants can be false positives.

Under the variable genetic effect model, when an exome-wide significance level $\alpha_{S1} = 2.5 \times 10^{-6}$ (an $\alpha = 0.05$ corrected for testing 20,000 genes) (Kryukov et al. 2009), and when a sample of 1,000 cases and 1,000 controls was analyzed, 60.8 %

Table 1 The discovery of rare variants in genetic studies

Number of cases/controls in stage 1 and 2 samples	Proportion of rare variant sites uncovered[a]		Proportion[a]	
	All	Causal	Locus PAR explained by uncovered causal rare variants	Causal variant sites among all uncovered rare variant sites
Variable effects phenotypic model				
1,000/1000[b]	0.608	0.777	0.601	0.594
5,000/5000[b]	0.772	0.827	0.776	0.589

[a]Results are based upon 2,000 replicates where for each replicate variant sites and frequencies were obtained from the Exome Variant Server for European Americans by randomly selecting 10,000 genes
[b] $\alpha_{S1} = 2.5 \times 10^{-6}$ Stage 1 α level (α = 0.05 with a Bonferroni correction for testing 20,000 genes)

of rare variant nucleotide sites are present in the dataset and a majority of (77.7 %) causal nucleotide sites can be uncovered. These uncovered variants explain nearly 100 % of the locus PAR (Table 1). Therefore, in principle, when a large stage 1 sample is analyzed, the advantage of sequencing for novel SNP discoveries diminishes as long as the stage 2 samples are drawn from the same population.

Since affected individuals are enriched in a case–control sample, nucleotide sites containing causal variants have a much higher probability of being uncovered than noncausal variant sites. Next, we compared the power for sequence-, variant-, and exome-array-based replication strategies under different combinations of false-positive/false-negative variant discovery rates, genotyping assay success rates, and error rates.

In the ideal scenario where both sequencing and customized genotyping qualities are perfect, the power for sequence- and variant-based replication strategies are jointly affected by the sample size, the proportions of rare variants uncovered, and the fractions of uncovered rare variant sites that contain causal variants. For most of the examined scenarios, the power of sequence-based replication is consistently better than variant-based replication when CMC is used. For example, under the variable effects model (Table 2), for a sample size of 2,000 cases and 2,000 controls, the power for sequence-based replication is 72.7 % while the power of variant-based replication is 69.5 %. The exome-array-based replication has slightly better power than variant-based replication (71.0 %). This is because most of the variants identified in a small stage 1 sample are relatively common and are included in the array. A larger fraction of causal variants can be analyzed in the second stage when the exome array is used for replication. It should be noted that these results are also somewhat biased, causing replication using the exome array to be more powerful than it may actually be, since many of the samples which are included in ESP were used in the design of the exome array.

When the sample size is increased to 3,500 cases and 3,500 controls, the power hardly differs between sequence- and variant-based replication. This is because a large proportion of variant sites are uncovered in the stage 1 sample, and the

Table 2 Power comparisons of sequencing-, variant-, and exome-array-based replication under the variable effect model

Number of cases/controls in stage 1 and 2 samples	Rates[a]				Power for replication[b]		
	False positive	False negative	Assay success	Error ratio	Sequence based	Exome-array based[c]	Variant based
2,000/2,000[d]	0	0	1	1	0.727	0.710	0.695
	1 %	4 %	0.9	0.5	0.713		0.680
				1			0.675
3,500/3,500[d]	0	0	1	1	0.899	0.779	0.898
	1 %	4 %	0.9	0.5	0.865	0.780	0.850
				1			0.823

[a]The power was empirically estimated using 2,000 replicates where for each replicate 10,000 genes were randomly selected from the Exome Variant Server

[b]Significance levels used for stage 1 and stage 2 studies $\alpha_{S1} = 2.5\times10^{-6}$ and $\alpha_{S2} = 2.5\times10^{-6}$

[c]Only variants on the exome array were analyzed in the stage 2 study

[d]The impact of different combinations of false-positive/false-negative rate, assay success rate, and genotyping and sequencing error rate ratio on the replication power is examined

uncovered variants account for nearly 100 % of the locus PAR. In this case, the exome-array-based replication is slightly less powerful, possibly because low-frequency causal variants are not included in the array and therefore cannot be analyzed in the replication study.

Comparisons were also made using the WSS for analysis of both the stage 1 and 2 datasets. Results are similar to those when CMC is used to analyze the data (data not shown).

Comparison of Replication Strategies Using Sequence Data from the Dallas Heart Study

In order to illustrate the relative efficiency of sequence-based versus variant- and exome-array-based replication strategies, a dataset from the DHS was used. The dataset is a multiethnic population-based sample (1,830 African Americans, 601 Hispanics, 1,045 European Americans, and 75 individuals from other ethnic groups) of Dallas County residents whose lipid and glucose metabolism have been characterized and recorded (Browning et al. 2004; Victor et al. 2004). In order to investigate how sequence variations in *ANGPTL3*, *ANGPTL4*, *ANGPTL5*, and *ANGPTL6* influence energy metabolism in humans, coding regions of the four genes were sequenced using DNA samples obtained from 3,551 participants in DHS (Romeo et al. 2007). A total of 348 nucleotide sites of sequence variations were uncovered in the four genes. Most of them are rare and 86 % of them have MAFs < 1 % (Romeo et al. 2007). Nine phenotypes were measured and tested for their associations with rare genetic variants, i.e., *b*ody *m*ass *i*ndex (BMI), *dias*tolic *b*lood *p*ressure (DiasBP),

*sys*tolic *b*lood *p*ressure (SysBP), *t*otal *c*holesterol *l*evel (TCL), *l*ow-*d*ensity *l*ipoprotein (LDL), *h*igh-*d*ensity *l*ipoprotein (HDL), *t*rig*l*yceride (TG), *v*ery *l*ow-*d*ensity *l*ipoprotein (VLDL), and glucose. For the stage 1 study, individuals with quantitative trait values in the lower 10 % and upper 90 % of the phenotypic distributions were used to form a "case–control" dataset ($N=710$). For the replication study, individuals with intermediate quantitative trait values ($N=1{,}776$), i.e., in the range of the lower 10–35 % and upper 65 %–90 % of trait values, were analyzed. Sequence-, variant-, and exome-array-based replication strategies were compared using the replication dataset.

For the first analysis, the stage 1 and 2 data from the *ANGPTL3, ANGPTL4*, *ANGPTL5*, and *ANGPTL6* genes are analyzed. Although a small sample size was used for the stage 1 study, multiple (novel) associations were detected using the CMC test (Table 3), i.e., (a) TCL with *ANGPTL3* $\left(p_{\mathrm{CMC}}=0.0283\right)$, (b) LDL with *ANGPTL* 4 $\left(p_{\mathrm{CMC}}=0.0208\right)$, (c) TG with *ANGPTL4* $\left(p_{\mathrm{CMC}}=0.0269\right)$, (d) VLDL with *ANGPTL4* $\left(p_{\mathrm{CMC}}=0.0373\right)$, (e) BMI with *ANGPTL5* $\left(p_{\mathrm{CMC}}=0.0287\right)$, (f) HDL with *ANGPTL5* $\left(p_{\mathrm{CMC}}=0.0252\right)$, and (g) BMI with *ANGPTL6* $\left(p_{\mathrm{CMC}}=0.0013\right)$. Among these, the association between BMI and *ANGPTL*6 is significant even after performing a Bonferroni correction for testing multiple genotypes and phenotypes; however, the association could not be replicated. For most of the analyses, approximately 25–40 % of the nucleotide sites observed in the entire DHS sample are also observed in stage 1. Since the stage 2 replication sample consists of individuals with less extreme quantitative trait values, to ensure that the power of the stage 2 replication sample is adequate, a much larger sample size is chosen for stage 2 ($N=1{,}776$) compared to stage 1 ($N=710$). Two of the seven identified associations in the stage 1 sample were successfully replicated by sequence-, variant-, and exome-array-based replication strategies, i.e., associations between TG and *ANGPTL4* as well as between VLDL and *ANGPTL4*. Given that the associations are mainly driven by the relatively common E40K variant, exome-array-based replication strategy has a slightly smaller p-value than the other two approaches.

For the second analysis, the empirical power for replicating the validated association between TG and rare variants in *ANGPTL4* gene was compared for variant-, sequence-, and exome-array-based replication strategies. For this analysis, phenotype and variant data for individuals with TG levels in the range of the lower 10–35 % and upper 65–90 % trait values ($N=1{,}776$) were sampled with replacement to form replicates each with a sample size of 710 individuals, which is the same sample size used for the stage 1 study. Each replicate was tested for an association using the CMC and the power was determined by the proportion of 2,000 replicates with a p-value <0.05. For sequence-based replication, all variants in the stage 2 sample were analyzed, and for the variant-based replication, only those variants observed in the stage 1 study were analyzed. While for the exome-array-based replication, only those variant sites which are available on the exome array were analyzed. The empirically estimated power for sequence-, variant-, and exome-array-based replication strategies are 65.3 %, 62.7 %, and 63.4 %, respectively. The power for sequence-based replication is only slightly better than the other replication strategies. This result is compatible with observations from simulated data.

Table 3 Analyses of sequence data from the *ANGPTL3*, *ANGPTL4*, *ANGPTL5*, and *ANGPTL6* genes using CMC test

Trait	*P*-values				Proportion	Ratio	Number of rare variants observed		
	Stage 1 analysis[a] (CMC)	Sequence-based replication[b] (CMC)	Variant-based replication[b] (CMC)	Exome-array-based replication[c]	Nucleotide sites uncovered in stage 1	Rare variant freq. in stage 1 sample/rare variant freq. in entire sample	Sequence-based replication	Variant-based replication	Exome-array-based replication
ANGPTL3									
TCL	0.028	0.522	0.726	0.832	0.30	0.87	46/51	39/40	37/39
ANGPTL4									
LDL	0.021	0.272	0.508	0.27	0.35	0.94	78/62	70/60	64/57
TG	0.027	0.025	0.039	0.017	0.26	0.92	77/51	69/46	69/46
VLDL	0.037	0.031	0.031	0.031	0.26	0.92	75/51	69/46	69/46
ANGPTL5									
BMI	0.029	0.464	0.451	0.437	0.5	0.95	67/71	63/67	64/66
HDL	0.025	1.0	0.772	0.854	0.5	0.95	63/66	61/60	54/56
ANGPTL6									
BMI	0.001	0.909	0.794	0.801	0.21	0.78	42/40	33/30	27/23

[a]For each phenotype analyzed, individuals with quantitative trait values from the top 10 % and upper 90 % were used as a stage 1 sample
[b]Individuals with quantitative trait values in the range of the lower 10–35 % and upper 65–90 % were used as the replication sample
[c]For exome-array-based replication strategy, only those variant sites available on the exome-array were analyzed in the replication sample

Resources for Designing Replication Studies

Many sequencing studies have publically released their frequency information, which can be useful for research investigators to design replication studies. For instance, the nucleotide site and frequency information have been released for the ESP from the National Heart, Lung, and Blood Institute on the EVS (http://evs.gs.washington.edu/EVS/) (Fu et al. 2013). This information may be useful for estimating power for a variant-based replication study. Additionally a complete list of variant sites represented on the exome array is available at http://genome.sph.umich.edu/wiki/Exome_Chip_Design. User-friendly software has been made available, such as SimRare (Li et al. 2012), and can be very useful for designing variant-based replication studies.

Conclusions and Discussions

In this chapter, we extended the work of Liu and Leal (2010b) to reflect recent development in sequence-based genetic studies. We evaluated strategies for sequence-, variant-, and exome-array-based replication for complex trait rare variant association studies and compared them using a rigorous population genetic framework. It is demonstrated that in the ideal scenario where sequencing and genotyping are both of perfect quality, sequence-based replication is consistently more powerful. However, since the uncovered variants can account for a large proportion of locus PAR even for a stage 1 study with only a few hundred samples, the advantage in power can be very small if stage 1 and stage 2 samples are drawn from the same population. The power of sequence- and variant-based replication studies is negatively impacted by sequencing and genotyping errors. For currently attainable levels of sequencing errors, the impact is minimal, and the advantage of using sequence-based replication studies remains. Using exome arrays for replicating gene-level association is also a cost-effective option; however, this option can be poorly powered if the variants from the stage 1 study are not well represented.

It has been found previously that rare variants tend to be population specific (Bodmer and Bonilla 2008). Many studies have suggested that disease-associated variants in different populations can have very different frequencies. For example, the E40K variant in the *ANGPTL4* gene was shown to be associated with TG levels; the MAF is approximately 3 % in European Americans but is very rare in African Americans and Hispanics (Romeo et al. 2007). These differences can be observed in even more closely related populations; for example, rare variants in *CFTR*, *BRCA1*, and *BRCA*2 genes have higher frequencies in the Ashkenazi Jewish population compared to Sephardic Jews and non-Jewish European populations (Kerem et al. 1997; King et al. 1993). Population-specific diversity of variant frequencies and sites is more pronounced for rare variants than for common variants since rare variants tend to be younger and occur more recently in human history (Bodmer and

Bonilla 2008). When stage 2 samples are drawn from a different population than the stage 1 samples, the variant-based replication studies may be at a severe disadvantage and grossly underpowered. Given that the demographic and selection models incorporating complex migration and admixtures are still limited (Boyko et al. 2008), simulation studies for variant discovery using multiethnic samples still remain to be explored. Evaluating the benefits and drawbacks of replication studies using samples from different populations will be very important.

Sequencing-based genetic studies have an irreplaceable advantage over genotyping, which is to discover novel genetic variants. Human population experienced complex patterns of demographic expansion and purifying selection (Romeo et al. 2007; Nielsen et al. 2007). Large numbers of very rare variant nucleotide sites exist. Based upon the observations from our extensive simulations and real data, for moderate-sized stage 1 studies, only a limited proportion of rare variant nucleotide sites can be uncovered. Identifying and cataloging rare variants themselves can be of great importance in genetic studies. The novel rare causal variants which are uncovered will help enhance the understanding of genetic architectures for complex traits. They can also be useful for risk prediction and personalized medicine. As a result, even if a gene is replicated using variant- or exome-array-based replication, sequencing of the gene region should also be eventually performed to uncover additional variants. For large-scale genetic studies with thousands of cases and controls, most of the disease-causative variants can be identified in stage 1. Therefore, for replicating large-scale studies, customized genotyping can be a viable solution. In addition, customized genotyping can be advantageous to targeted sequencing in that multiple unlinked markers can be genotyped and used to control for population substructure/admixture. The advantage is particularly beneficial when GWAS data is not available for the replication sample. On the other hand, using exome-array-based replication can also be a viable approach for replicating both large-scale and small-scale studies. Prior to designing replication studies, it is possible to examine whether promising variant sites uncovered in stage 1 are represented on the exome array. When most of the variant sites uncovered are present, exome-array-based replication can be a cost-effective strategy. In particular, for small-scale studies, where most of the identified variants in stage 1 sample are relatively common, using the exome array can be both more powerful and more cost-effective than custom genotyping-based replication. It should be noted that, because the exome array was designed using mainly individuals of European descent, there may be poor representation of variants found only in non-European populations.

With the rapid large-scale application of NGS, understandings of genetic etiologies of rare variants will advance to an unprecedented level. Replications of significant findings will be an indispensable part of every genetic study. Sequence-based replication for both small- and large-scale genetic studies is advantageous and will eventually be affordable and widely applied. In the meantime, variant- or exome-array-based replication can be a temporary, more cost-effective solution for the replication of genetic sequence-based studies and will greatly accelerate the process of identifying disease-causative variants.

Acknowledgments This work was supported by National Institutes of Health grants HL102926 and MD005964.

References

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40(6):695–701

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR et al (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4(5):e1000083

Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, Grundy SM, Hobbs HH (2004) Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology 40(6):1387–1395

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685):869–872

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 103(6):1810–1815

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493(7431):216–220

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S et al (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):R32

Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, Stringham HM, Sim X, Yang L, Fuchsberger C, Cederberg H et al (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45(2):197–201

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40(5):592–599

Kerem B, Chiba-Falek O, Kerem E (1997) Cystic fibrosis in Jews: frequency and mutation distribution. Genet Test 1(1):35–39

King MC, Rowell S, Love SM (1993) Inherited breast and ovarian cancer. What are the risks? What are the choices? JAMA 269(15):1975–1980

Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A 106(10): 3871–3876

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311–321

Li B, Wang G, Leal SM (2012) SimRare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits. Bioinformatics 28(20):2703–2704

Liu DJ, Leal SM (2010a) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6(10):e1001156

Liu DJ, Leal SM (2010b) Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 87(6):790–801

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188–193

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8(11):857–868

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69(1):124–137

Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39(4):513–516

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119(1):70–79

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69

Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA et al (2004) The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol 93(12):1473–1480

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93

Meta-Analysis of Rare Variants

Ioanna Tachmazidou and Eleftheria Zeggini

Motivation for Meta-Analysis

Meta-analysis is the use of statistical methods to synthesize results of individual studies examining the same trait. A genome-wide meta-analysis primarily serves the purpose of combining data to increase power to obtain statistical evidence of association between disease and variants that would have otherwise escaped detection, for example because of their small effect sizes. For example, the power to attain a *p*-value of genome-wide significance (5×10^{-8}) for a common variant with 0.20 MAF and a small effect size (odds ratio 1.15) in a GWAS of 2,000 cases and 3,000 controls is 0.45 %, assuming disease prevalence of 1 %, a multiplicative disease model and that the causal variant is typed itself. In contrast, a GWAS meta-analysis of five similar homogeneous studies across 10,000 cases and 15,000 controls has 80 % power to identify risk variants at the genome-wide significance level. Chapman et al. (2011) investigated the way sample size affects the power of GWAS meta-analyses, in the presence and absence of modest levels of heterogeneity and across a range of different allelic architectures.

Genome-wide meta-analysis is facilitated by imputation (Marchini et al. 2007), which enables the combination of data across different genotyping platforms. Reference datasets, such as those emerging from the HapMap (www.hapmap.org), 1,000 Genomes (www.1000genomes.org) and the UK10K (www.uk10k.org) projects, can be used to impute genotypes for all variants at untyped positions in the target dataset of interest, using the GWAS genotypes as a scaffold. Because of limited overlap of markers genotyped between platforms, variants are likely to be imputed in some studies and directly typed in others. Meta-analysis and imputation are analytical tools, which have been widely employed in the field of complex

I. Tachmazidou (✉) • E. Zeggini
Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
e-mail: it3@sanger.ac.uk; Eleftheria@sanger.ac.uk

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_15

disease genetics. However, genotyping and imputation of low frequency and rare variants tend to be less accurate compared to common variation, and the current common practice in GWAS meta-analysis is to investigate association only at common variants to avoid false-positive association signals that are driven by statistical or genotyping artefacts.

Whole-exome or whole-genome sequencing marks the beginning of a new era for genetics with exciting possibilities, but also with new statistical challenges. Aggregate tests that combine information from low frequency/rare variants within a region are typically used to test association of a gene or other chromosomal unit with the trait of interest (Asimit and Zeggini 2010; Ladouceur et al. 2011; Chen et al. 2011), as single-point analysis of variants tends to have low power for variants at the low end of the MAF spectrum. The evaluation of different meta-analysis approaches to rare variant single-point and burden test association statistics is an active field of study catalyzed by method development.

Key Principles of Meta-Analysis and Practical Considerations

Often, when undertaking a GWAS meta-analysis, extensive information is required, such as summary statistics at each variant (e.g. quality control metrics, effect sizes and their standard errors, *p*-values, imputation accuracy scores, MAF), information on the analysis method and the covariates used, the size of the study (e.g. number of cases and controls, or the effective sample size), filtering of samples and variants, approaches taken to adjust for any population stratification (e.g. genomic control Devlin et al. 2001 or principal component adjustment Price et al. 2006), relatedness of samples within and between studies, and the strand and build of the human genome on which allele coding has been based (Zeggini and Ioannidis 2009). It is important to ensure that individual studies have been subjected to rigorous quality checks to avoid spurious associations. de Bakker et al. (2008), Zeggini and Ioannidis (2009), and Thompson et al. (2011) provide detailed descriptions of the different stages of conducting GWAS meta-analysis.

Meta-analysis of rare variants from sequencing studies require additional harmonization of the analysis protocol across collaborators, specifying not only analysis method, but also parameters such as the unit of interest in region-based tests, the type of variants examined within the region (e.g. all variants that fall within a MAF and/or functional annotation class), and any weighting schemes used (e.g. based on MAF, functional annotation, sequencing/genotyping/imputation accuracy scores, etc.). Quality control of sequencing studies is very important, as rare variant calls could be the result of sequencing errors. Synthesizing results from sequencing data that are produced using different technologies, at different depth and/or called using different pipelines and parameters also requires careful considerations.

Traditional meta-analysis techniques for GWAS can be effect size-, *p*-value-, or Bayes' factor-based. *P*-value based meta-analyses are typically only used if effect size estimates from individual studies and their respective standard errors are not

available (either because of accessibility issues, or because the test used does not return an effect size, e.g. SKAT Wu et al. 2011), or when they are known but they measure quantities that are not directly comparable (e.g. in different units). *P*-value based meta-analysis cannot provide meta-analytic effect sizes or traditional estimates of heterogeneity. For these reasons, effect size-based meta-analysis is preferable where feasible, and can be implemented using a fixed or random effects model. Fixed effects meta-analysis assumes that the true effect size of each variant is the same across all individual studies, and therefore no heterogeneity exists among the studies. In contrast, a random effects meta-analysis assumes that different effect sizes of the same variant drive association among the different studies, and therefore between-study heterogeneity exists. The fixed effects approach achieves the highest power to detect association, while the random effects approach produces estimates with larger uncertainty and lower statistical significance in the presence of true heterogeneity. Software packages, such as META (Liu et al. 2010) and GWAMA (Magi and Morris 2010), implement both the fixed and random effects model for single variants in a computationally efficient manner; they also align effects to the same strand, calculate heterogeneity statistics, and offer genomic control correction for population stratification, all important parameters in obtaining robust results. The traditional random effects model may be conservative, as it implicitly assumes heterogeneity under the null hypothesis of no disease association. Recently, Han and Eskin (2011) proposed a random effects model that assumes homogeneity under the null model, while a random effects component is used to inflate the variance of the estimated allelic effect of each variant under the alternative model, and can thus increase power in the presence of heterogeneity. Bayesian meta-analysis approaches incorporate uncertainty in prior beliefs about between-study heterogeneity, effect sizes, and genetic model (Verzilli et al. 2008; Newcombe et al. 2009; De Iorio et al. 2011). However, it is not clear how Bayesian meta-analysis approaches scale to GWAS. Recently, a computationally tractable Bayesian meta-analysis approach that uses a Markov chain Monte Carlo algorithm to calculate posterior probabilities that the effect exists in each study has been proposed (Han and Eskin 2012). Evangelou and Ioannidis (2013) give an overview of statistical methods for GWAS meta-analysis.

Meta-Analysis of Sequencing Studies

Ma et al. (2013) examined the efficiency of joint and meta-analysis for low frequency variants. They focused on case–control studies and the logistic regression-based Wald, score, likelihood ratio, and Firth bias-corrected (Firth 1993) tests. These tests control the type I error rate well and are asymptotically equivalent for common single-variant testing (Cox and Hinkley 1974). However, the asymptotic assumptions for logistic regression may not be valid for low frequency variants, making them conservative or anti-conservative for the study of such variants. Ma et al. (2013) studied the behavior of these tests in joint and meta-analysis of binary traits for low

frequency variants. They find that in studies with balanced number of cases and controls and for joint analysis, the Firth test has type I error rates close to the nominal threshold. In contrast, the score and Wald tests are very conservative, and the likelihood ratio test can be slightly anti-conservative. The Firth test has also the best power in this setting. For meta-analysis of balanced studies, the score test controls the type I error rate best (Firth and particularly Wald test results are more conservative, while likelihood ratio test can again be anti-conservative), although it is less powerful than Firth-test based joint analysis. For sufficiently unbalanced studies (e.g. 1:3 or 1:19 cases and controls), tests in joint analysis behave the same as for balanced studies, whereas in meta-analysis all tests can be highly anti-conservative. However, these results were obtained for homogeneous studies with no covariate adjustments, and it is not clear how the power of joint analysis will compare to meta-analysis when combining heterogeneous studies. Moreover, Ma et al. (2013) show that test calibration remains consistent when using a constant minor allele count threshold, below which tests begin to deviate substantially from the nominal significance threshold.

Asimit et al. (2012) examined the utility of traditional meta-analytical approaches in combining data across independent studies to increase power for region-based tests. Collapsing methods that combine low frequency/rare variants into one super locus (such as, for example, Morris and Zeggini 2010) provide both a global coefficient estimate and a *p*-value of significance, and therefore a sample size and an inverse variance based meta-analysis is possible, but not necessarily powerful. However, not all rare variants tests (such as, for example, Mukhopadhyay et al. 2010; Wu et al. 2011) provide global coefficient estimates that can be combined in a traditional inverse variance based meta-analysis. Moreover, a sample size based meta-analysis requires a global direction of effect, and can therefore be used only with rare variant tests that do not allow for different directions of effect within the locus of interest. A straightforward *p*-value based approach of meta-analyzing results from region-based analyses is applying Fisher's combined probability test (Fisher 19321) or Stouffer's z-score method (Stouffer et al. 1949).

Fisher's combined probability test (Fisher 19321) is a method that combines results from several independent tests of the same hypothesis. Fisher's method combines *p*-values from each study by summing their natural logarithm and multiplying the resulting quantity by − 2. When the combined *p*-values are independent and if all the null hypotheses are true, Fisher's test statistic has asymptotically a chi-squared distribution with $2 \times M$ degrees of freedom, where M is the number of combined studies. In particular, under the global null hypothesis:

$$X^2 = -2\sum_{M}^{m=1} \log_{\epsilon}(p\ \text{value}_m) \sim \chi^2_{2M}.$$

Weights can be easily introduced, in which case the test becomes a weighted sum of independent chi-squared statistics under the global null hypothesis. Fisher's null hypothesis is that all of the individual null hypotheses are true, whereas their alternative hypothesis is that at least one of the individual alternative hypotheses is true.

Therefore, Fisher's test does not allow for the possibility of heterogeneity, where the null hypothesis holds in a subset of the combined studies but not in all of them.

A similar approach to Fisher's test is Stouffer's z-score method (Stouffer et al. 1949), which is based on z-scores rather than *p*-values. Stouffer's method combines z-scores from each study by summing the inverse normal transformation of the individual *p*-values, which is asymptotically normally distributed under the global null hypothesis:

$$Z = \frac{\sum_{M}^{m=1} \Phi^{-1}(1 - \frac{p\ \text{value}_m}{2})(\text{direction of effect for study i})}{\sqrt{M}} \sim N(0,1).$$

Weights ω_m are also easily introduced, in which case the test still has a normal null distribution:

$$Z = \frac{\sum_{M}^{m=1} \omega_m \Phi^{-1}(1 - \frac{p\ \text{value}_m}{2})(\text{direction of effect for study i})}{\sqrt{\sum_{M}^{m=1} \omega_m^2}} \sim N(0,1).$$

However, region-level *p*-value based meta-analytical approaches, such as Fisher's and Stouffer's tests, are not necessarily powerful in combining data across independent studies for rare-variant association testing. Ideally, meta-analytical approaches for next generation sequencing studies result to no or little power loss as compared to a joint analysis approach, in the same way as meta-analysis of singe-tests for common variants.

Recently Liu et al. (2013a) compared the power of region-based analysis when the data across studies are meta-analyzed as compared to when the data are pooled together and then analyzed as if they came from the same study (mega-analysis). They conducted their meta-analysis by combining region-based *p*-values provided by the Sequence Kernel Association Test (SKAT, Wu et al. 2011) across studies using Stouffer's test, where the weights of the study-specific z-scores were chosen to be the study's sample size. The mega-analysis requires careful harmonization of the datasets to ensure that the distribution of rare variants is the same between studies. This was achieved by filtering variants based on call rate, read depth, and balance of alternative to reference reads. Homogeneity was ensured by achieving similar numbers of minor allele calls (MAC) per sample per gene for variants of MAF less than 1 %. Stringent filtering, however, can eliminate real signals along with false ones. The authors suggest sequencing a few samples in all centers that produced the data to monitor and evaluate the results of filtering, but this may not be always feasible. Population stratification can be a source of heterogeneity within and between studies. Population stratification at low frequency/rare variants might not be corrected by adjusting for principle components (PC) using common or low frequency variants, and evidence has been contradictory (Mathieson and McVean 2012; Zhang et al. 2012). Via simulation, Liu et al. (2013a) concluded that mega-analysis is more

powerful than Stouffer's meta-analysis technique. This is due to the fact that Stouffer's meta-analysis approach combines gene-level statistics, whereas mega-analysis combines all the available information in the datasets at the genotype level. On the other hand, appropriate meta-analytic approaches allow for heterogeneity between datasets, whereas mega-analysis explicitly assumes no heterogeneity between samples. Heterogeneity can be the result of different sequencing tools and protocols utilized for the studies being combined, and Liu et al. (2013a) suggest ways of removing these sources of heterogeneity. However, heterogeneity can be inherent between datasets, for example when combining datasets from different ethnic groups. Another disadvantage of mega-analysis is that it requires access to the full genotype data instead of association summary statistics, which makes this approach infeasible in cases where individual-level data cannot be shared. Mega-analyses can also be impractical not only because they require transferring large amounts of data, in contrast to meta-analysis where only summary statistics are shared, but also when different studies have collected and adjusted for different covariates.

Heterogeneity and the Effects on Power

It is expected that different causal variants of low frequency will be found to reside within the same functional units, and that different alleles will be carried by different individuals across studies (allelic heterogeneity) (Cirulli and Goldstein 2010; Eichler et al. 2010). Allelic and locus heterogeneity is expected to be a particularly important consideration in meta-analysis of trans-ethnic studies.

Asimit et al. (2012) evaluated different meta-analysis approaches in the presence of allelic heterogeneity. They simulated case–control data from different populations where the causal variants were population-specific, and for each population separately, they performed an association analysis using the collapsing method of Morris and Zeggini (2010) and the allele-matching association test KBAT that allows for different directions of effect (Mukhopadhyay et al. 2010). Subsequently, they performed a meta-analysis using the traditional inverse variance based technique with the odds ratio estimate from the collapsing method, and Fisher's meta-analytic technique with the *p*-values of the collapsing method and KBAT. They found that a *p*-value based meta-analysis of summary results from allele-matching locus-wide tests has some power advantages, although power remains low. Moreover, they found that for low-frequency variants with large effects (odds ratios 2–3), single-point tests have high power, but also high false-positive rates. They concluded that current strategies for the combination of genetic association data in the presence of allelic heterogeneity are insufficiently powered. New methodological approaches are required and are currently being developed for meta-analysis of sequencing studies to allow for locus and allelic heterogeneity (Lumley et al. 2013; Lee et al. 2013; Tang and Lin 2013; Liu et al. 2013b).

Lumley et al. (2013) recently developed a meta-analytical approach based on SKAT (Wu et al. 2011), and showed that their meta-analysis technique, called skatMeta, is as efficient as an analysis that pools individual-level data together. Within the chromosomal region of interest, for each cohort k and for each variant j, $j = 1,\ldots,m$, a score $\hat{\beta}_{kj}$ and its associated variance $\hat{s}_{kj}^2$ are calculated as:

$$\hat{\beta}_{kj} = \frac{\sum_{i=1}^{N_k} G_{kij} Y_{ki}}{2N_k p_j (1-p_j)}, \text{and } \hat{S}_{kj}^2 = \frac{1}{N_k} \hat{\sigma}_k^2 \, 2p_j(1-p_j),$$

where G_{kij} and Y_{ki} are the genotype and phenotype of individual i and variant j in study k, N_k is the sample size of study k, $\hat{\sigma}_k^2$ is the variance of the phenotype in study k, and p_j is the average MAF of variant j across all studies. If a variant is absent in some studies, then it is assumed to be observed in these studies with its regression coefficient reduced to zero.

Then a pooled score $\hat{\beta}_j$ and a pooled score variance $\hat{s}_j^2$ is calculated for each variant j by a standard inverse variance based meta-analysis across cohorts, and therefore a pooled score test statistic w_j is obtained. The skatMeta statistic is the sum of squares of the pooled score test statistics across all low frequency/rare variants in the unit of interest, weighted by a function of their pooled MAF across all cohorts:

$$Q = \sum_{j=1}^{m} \omega_j \hat{z}_j^2,$$

where ω_j is the weight for variant j. Asymptotically, the skatMeta statistic is distributed as a sum of chi-squared statistics that depends on the number of variants observed and the average linkage disequilibrium between variants. Therefore, skatMeta uses score statistics and the MAF of variants from each study, and the study-specific genotype covariance matrix, which makes it applicable even when individual-level data cannot be shared. As a regression based test, skatMeta can handle both binary and continuous traits. Although it can be adjusted for covariates, such as principal components, the meta-analysis is less exact. In a simulation study, Lumley et al. (2013) showed that skatMeta utilizes all the information in the data.

A simpler approach of meta-analysing SKAT results is by summing SKAT test statistics Q_k from each cohort k weighted by a cohort weight ν_k (proportional for example to sample size). This weighted sum is asymptotically distributed as a linear combination of chi-squared tests:

$$Q^{\star} = \sum_{k=1}^{K} \nu_k Q_k \sim \sum_{k,j} \lambda_{kj} \chi_1^2.$$

Lumley et al. (2013) showed that this simplistic approach is significantly less powerful than skatMeta. In fact, its power is comparable to the power of Fisher's and Stouffer's method.

The approach of Lumley et al. (2013) is an extension of the fixed-effect meta-analysis model for single variants. Lee et al. (2013) independently developed a meta-

analysis approach of rare variants applicable to burden tests (Li and Leal 2008; Morris and Zeggini 2010), SKAT, and the unified SKAT-O test (Lee et al. 2012), called MetaSKAT, which can assume both homogeneous and heterogeneous genetic effects across studies, corresponding to a fixed and random effects meta-analysis model respectively. In the same spirit as Lumley et al. (2013), the fixed-effects meta-analysis version of SKAT by Lee et al. (2013) first accumulates the weighted score of each variant j across K studies and then sums the squared accumulated score statistics across the m variants in the region of interest. In particular, if S_{kj} is the score statistic and ω_{kj} is the weight of variant j in study k, then the SKAT-O test statistic for a fixed-effects meta-analysis is given by a weighted average of the SKAT and burden meta-analysis test statistics:

$$Q_{\text{hom-meta}}(\rho) = (1-\rho)Q_{\text{hom-meta-SKAT}} + \rho Q_{\text{meta-Burden}},$$

where

$$Q_{\text{hom-meta-SKAT}} = \sum_{j=1}^{m}\left(\sum_{k=1}^{K}\omega_{kj}S_{kj}\right)^2, \quad Q_{\text{meta-Burden}} = \left(\sum_{j=1}^{m}\sum_{k=1}^{K}\omega_{kj}S_{kj}\right)^2, \text{and}$$

ρ is interpreted as the pair-wise correlation among the coefficients of genetic effects β_{kj}. When $\rho=0$, $Q_{\text{hom-meta}}(0)$ corresponds to a joint analysis of the k studies using SKAT, whereas for $\rho=1$, $Q_{\text{hom-meta}}(1)$ corresponds to a joint analysis of the studies using a burden test.

To allow between-study heterogeneity, the effect size of variant j in the combined studies are assumed independent and to have a common distribution. The random-effects meta-analysis SKAT-O statistic is given by Lee et al. (2013) as:

$$Q_{\text{het-meta}}(\rho) = (1-\rho)Q_{\text{het-meta-SKAT}} + \rho Q_{\text{meta-Burden}},$$

where

$$Q_{\text{het-meta-SKAT}} = \sum_{j=1}^{m}\sum_{k=1}^{K}\omega_{kj}^2 S_{kj}^2, \text{ where}$$

This method can accommodate different levels of heterogeneity of genetic effects across studies. For example, if a subset of the studies combined belong to the same ancestry group, $Q_{\text{het-meta-SKAT}}$ can be rewritten to allow homogeneity between those and at the same time heterogeneity between the studies from different ancestry groups. If the weighting scheme is based on MAF, the average MAF across studies can be used for a fixed-effects meta-analysis, whereas for studies grouped by ancestry, weighs can be based on ancestry specific MAFs. Moreover, if a variant is absent in some studies, then it is assumed to be observed in these studies with its score set to zero. It can be shown that the asymptotic null distribution of $Q_{\text{hom-meta}}(\rho)$ and $Q_{\text{het-meta-SKAT}}$ can be approximated as a mixture of chi-square distributions. As both SKAT and burden tests are implemented in a regression framework, MetaSKAT can analyze binary and continuous traits.

Lee et al. (2013) compared the power and type I error of MetaSKAT to Fisher's test using individual study SKAT-O p-values and inverse variance weighting based meta-analysis burden tests in a simulation study of varying levels of heterogeneity. Overall, MetaSKAT controlled the type I error rate well at lower significance levels, with slightly inflated and deflated errors for continuous and binary traits respectively for increasing significance levels. As expected, power for meta-analysis burden tests was substantially reduced in the presence of both deleterious and protective variants. Its performance is also not robust to the proportion of causal variants included in the analysis, in contrast to MetaSKAT. When genetic effects are homogeneous across studies, the power of meta-analysis using Hom-Meta-SKAT and Hom-Meta-SKAT-O is similar to those for joint analysis using SKAT and SKAT-O, while Het-Meta-SKAT-O had modest power loss. In the presence of heterogeneity between studies, Het-Meta-SKAT-O and Het-Meta-SKAT were the most powerful approaches. Fisher's test had overall similar or lower power than Het-Meta-SKAT-O.

Tang and Lin (2013) also developed a program, called Meta-Analysis of Score Statistics (MASS), that performs meta-analysis of rare variants by combining score statistics across studies. MASS implements three different tests that correspond to a variety of rare variants test, such as the burden test (Li and Leal 2008; Morris and Zeggini 2010), variable threshold test (VT) (Price et al. 2010; Lin and Tang 2011), C-alpha test (Neale et al. 2011), and SKAT (Wu et al. 2011). Let $U^{(k)}$, an $m \times 1$ vector if there are m variants under investigation in the region, denote the score statistic for study k and $V^{(k)}$, an $m \times m$ matrix, be the corresponding information matrix for study k. Then the score of each variant is collapsed across the studies, and therefore the overall score statistic U and overall information matrix V are calculated as:

$$U = \sum_{k=1}^{K} U^{(k)}, \text{ and } V = \sum_{k=1}^{K} V^{(k)}.$$

Under the null hypothesis that none of the variants are associated with the outcome in any of the studies, pertaining to a fixed-effects meta-analysis model, U asymptotically follows a multivariate normal distribution with mean 0 and covariance matrix V. It can be shown that U is the score statistic for testing the null hypothesis when using the joint individual-level datasets. Using U and V, MASS implements the quadratic statistic $Q = U^T V^{-1} U$ (which encompasses the CMC test), the maximum statistic $T_{\max} = \max_{j=1,\ldots,m} U_j^2 / V_j$ (which encompasses the VT test), and the weighted quadratic statistic $Q_\omega = U^T \Omega U$ (which encompasses C-alpha and SKAT). Under the null hypothesis, the quadratic statistic has a chi-squared distribution with m degrees of freedom, the distribution of the maximum statistic is determined by the multivariate normal distribution of U under the null model, while the null distribution of the weighted quadratic statistic is a linear combination of chi-squared tests of 1 degree of freedom that are proportional to the correlation of variants within studies.

Liu et al. (2013b) independently proposed an approach for meta-analysis of a number of popular gene-level association tests, such as burden tests, VT test and SKAT, weighted or un-weighted by MAF or predicted functional annotations. Liu et al. (2013b) approach is very similar to Tang and Lin (2013) method, and it is implemented in a software called RareMETAL. As with other meta-analysis methods of gene-level tests, RareMETAL requires single variants score statistics, MAFs, and correlation information within studies. A novel feature of RareMETAL is that apart from calculating asymptotic *p*-values, it also evaluates significance in an empirical and numerically stable way via Monte-Carlo simulations. Since, as discussed above, $U \sim \mathrm{MVN}(0,\mathrm{V})$, we can generate empirical distributions for gene-level statistics by sampling from this multivariate normal distribution. An adaptive approach can be used for computational efficiency, where a larger number of simulations are performed when assessing small asymptotic *p*-values and fewer simulations for assessing larger asymptotic *p*-values. Another unique feature of RareMETAL is its ability to conduct conditional meta-analysis of gene-level tests. Extending an approach used by Yang et al. (2012) for conditional meta-analysis of common variants, Liu et al. (2013b) facilitate meta-analysis of gene-level tests conditional on common variants in the gene. Liu et al. (2013b) illustrate in a simulation study that RareMETAL produces similar results when individual level data are shared from homogenous studies, or when study heterogeneity (e.g. in trait means and variance, or covariate effects) has been adjusted for. Moreover, RareMETAL is more powerful than Fisher's method for combining *p*-value, and it controls the type I error rate well.

After a meta-analysis has been conducted, interesting signals are prioritized for follow-up and replication, in order to achieve genome-wide significance. The replication stage could be a large meta-analysis itself. Typically the replication data are meta-analyzed with the discovery data to capture the totality of the evidence. The replication of previously established loci can serve as validation for the approach followed.

Concluding Remarks

Over the next few years the field of complex trait genetics is poised to witness large-scale collaborative efforts to meta-analyse sequencing studies. As whole genome sequencing costs continue to decline and approach those of dense GWAS arrays, the expectation is that future association studies will be truly genome-wide, assessing variation across the full allele frequency spectrum. Low depth whole genome sequencing experiments across hundreds of thousands of individuals and also across diverse ethnic groups will become a reality, potentially transforming the current understanding of complex trait architecture. Statistical genetics method development is currently an intensely active field, which will deliver efficient solutions to analytical and computational challenges associated with this unprecedented scale and structure of data.

References

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Asimit J, Day-Williams A, Zgaga L, Rudan I, Boraska V, Zeggini E (2012) An evaluation of different meta-analysis approaches in the presence of allelic heterogeneity. Eur J Hum Genet 20:709–712

Chapman K, Ferreira T, Morris A, Asimit J, Zeggini E (2011) Defining the power limits of genome-wide association scan meta-analyses. Genet Epidemiol 35:781–789

Chen H, Hendricks AE, Cheng Y, Cupples AL, Dupuis J, Liu CT (2011) Comparison of statistical approaches to rare variant analysis for quantitative traits. BMC Proc 5:S113

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight B (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:122–128

De Iorio M, Newcombe PJ, Tachmazidou I, Verzilli CJ, Whittaker JC (2011) Bayesian semiparametric meta-analysis for genetic association studies. Genet Epidemiol 35:333–340

Devlin B, Roeder K, Wassermanb L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Genet Rev 14:379–389

Firth D (1993) Bias reduction of maximum-likelihood-estimates. Biometrika 80:27–38

Fisher RA (1932) Statistical methods for research workers. Oliver and Boyd, Edinburgh

Han E, Eskin JP (2011) Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet 88:586–598

Han E, Eskin JP (2012) Interpreting meta-analysis of genome-wide association studies. PLoS Genet 8:e1002555

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Brent RJ (2011) The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. PLoS Genet 8:e1002496

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91:224–237

Lee S, Teslovich TM, Boehnke M, Lin X (2013) General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet 93:1–12

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321

Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89:354–367

Liu JZ et al (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet 42:436–440

Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, Lim E, Bodea CA, Muzny D, Reid JG et al (2013a) Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet 9:e1003443

Liu DJ, Peloso GM, Zhan X, Holmen O, Zawistowski M, Feng S, Nikpay M, Auer PL, Goel A, Zhang H et al (2013b) Meta-analysis of gene level association tests. http://arxiv.org/abs/1305.1318

Lumley T, Brody J, Dupuis J, Cupples A (2013) Meta-analysis of a rare variant association test. http://stattech.wordpress.fos.auckland.ac.nz/files/2012/11/skat-meta-paper.pdf

Ma C, Blackwell T, Boehnke M, Scott LJ, The GoT2D Investigators (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol. doi:10.1002/gepi.21742

Magi R, Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinf 11:288–294

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39:906–913

Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44:243–246

Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193

Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A (2010) Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol 34:213–221

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322

Newcombe PJ, Verzilli C, Casas JP, Hingorani AD, Smeeth L, Whittaker JC (2009) Multilocus Bayesian meta-analysis of gene-disease associations. Am J Hum Genet 84:567–580

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838

Stouffer SA, Suchman EA, DeVinney LC, Williams JRM (1949) The American soldier, volume I: adjustment during army life. Princeton University Press, Princeton

Tang ZZ, Lin DY (2013) MASS: meta-analysis of score statistics for sequencing studies. Bioinformatics 29:1803–1805

Thompson JR, Attia J, Minelli C (2011) The meta-analysis of genome-wide association studies. Brief Bioinform 12:259–269

Verzilli C, Shah T, Casas JP, Chapman J, Sandhu M, Debenham SL, Boekholdt MS, Khaw KT, Wareham NJ, Judson R, Benjamin EJ, Kathiresan S, Larson MJ, Rong J, Sofat R, Humphries SE, Smeeth L, Cavaller G, Whittaker JC, Hingorani AD (2008) Bayesian meta-analysis of genetic association studies with different sets of markers. Am J Hum Genet 82:859–872

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93

Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ, Frayling TM, McCarthy MI, Hirschhorn JN, Goddard ME, Visscher PM (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369–375

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201

Zhang Y, Guan W, Pan W (2012) Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 37:99–109

Population Stratification of Rare Variants

Emmanuelle Génin, Sébastien Letort, and Marie-Claude Babron

Introduction

It has long been recognized that there can exist, at least at some loci, differences in allele frequency between human populations. In 1919, Hirszfeld and Hirszfeld (1919) first reported differences in the ABO blood group frequency between soldiers from different ethnic origins. In their pioneer work, Cavalli Sforza et al. (1994) extended this observation to several different loci and showed a global pattern of allele frequency gradients over the world that follows human migrations. More recently, with the development of high-throughput technologies and the possibility to assess genetic variations at an unprecedented scale of hundreds of thousands of markers, it was shown that these differences exist at all the geographic levels: between continents, between countries within a continent and also between regions within a country.

Describing the genetic diversity between groups of individuals and understanding its origin are the basis of population genetics. Indeed, apart from migrations, different phenomena can lead to allele frequency variation between groups of individuals. These differences are essential as they are the bricks on which the species relies to ensure its survival. They need to be recognized and accounted for when studying the

E. Génin (✉) • S. Letort
Inserm UMR-1078, Génétique, Génomique fonctionnelle et Biotechnologies,
46 rue Félix Le Dantec, CS 51819, Brest Cedex 2 29218, France

Centre Hospitalier Universitaire de Brest Morvan, Brest, France
e-mail: emmanuelle.genin@inserm.fr

M.-C. Babron
Inserm UMR-946, Genetic Variability and Human Diseases, Paris, France

University Paris-Diderot, Sorbonne Paris Cité, UMRS-946, Paris, France

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_16

genetic determinants of complex traits to optimally design the study and to avoid false conclusions.

Indeed, in the context of genetic association studies, allele frequency differences between subgroups of individuals when coupled with differences in disease risks can lead to population stratification. Sampling cases and controls without accounting for population stratification exposes to a risk of falsely concluding that the alleles that have an increased frequency in the subgroups with a higher disease risk are involved in the disease either directly or because they are located in the vicinity of a disease risk locus. This problem is more likely to arise when the individuals are sampled from different ethnic groups as illustrated by the famous example of an association between a Gm haplotype and type 2 diabetes that was explained by Pima–Papago ancestry (Knowler et al. 1988). More recently with the advent of genome-wide association studies (GWAS), the problem becomes more crucial as very large samples of cases and controls are compared at hundreds of thousands of markers (Clayton et al. 2005). It was a stimulus for the development of new methods to detect and correct for population stratification in case–control data (for a review, see (Price et al. 2010)) that were shown to perform well under most scenarios. These novel methods however were developed to analyse common genetic variants and the few studies performed so far on rare variants agree in showing that they might not be as efficient to correct for rare variant stratification that differs in its pattern from common variant stratification.

In this book chapter, after recalling some of the basic principles of population genetics and the methods used to evidence population stratification and to correct for it in association studies of common variants, we review the different studies performed so far to assess the population stratification of rare variants and its impact on association tests.

Some Basic Principles of Population Genetics

Population genetics is the study of allele frequency variation in time and space. It was born in the 1920s with the founder works of R.A. Fisher, J.B.S. Haldane and S. Wright to reconcile the Mendelian theory of heredity and the Darwinian theory of evolution. Four main evolutionary forces play an important role in shaping human genetic diversity: mutation, natural selection, genetic drift and gene flow through migrations. To keep it simple, a new genetic variant is created by mutation. If this variant is neutral, it will most likely disappear because of genetic drift and because not all genetic variants are transmitted to the next generation. The rarer the variant, the less likely it will be transmitted. Its frequency could also increase by chance at a rate that depends on several factors but is mostly determined by the population effective size and its growth rate. If the variant is deleterious, there are even more chances that it will disappear very quickly from the population because of negative selection and because of the reduced fitness of carriers of this variant. On the contrary, if the variant is favourable and confers a selective advantage to its

carriers, it will more likely increase in frequency in the population at a quicker rate than the one expected for a neutral variant through genetic drift. Migrations can interfere in this process that would otherwise ultimately lead to population differentiation, by creating gene flows between populations and homogenization. Since migrations are more likely to occur between neighbouring populations, they can lead to some gradients of allele frequencies such as the one first described by Cavalli Sforza et al. (1994). Moreover, demography also plays an important role in shaping the spatial patterns of genetic diversity. Rapid population growth such as the one experienced by human populations over the past 400 generations (from a few million people 10,000 years ago to seven billion today) increases the load of rare variants that are generated by recent mutations (Keinan and Clark 2012).

From these basic principles of how the different evolutionary forces interact, it is clear that most rare genetic variants are young variants that have just arisen in the population through mutations and have had no time to increase in frequency and be dispersed in space. Therefore, they are expected to be seen only in a few restricted geographic areas. Some of the rare variants could be older variants that are under negative selection and maintained at low frequency because they induce some fitness reduction, either directly or indirectly, through linkage with deleterious variants. These older variants under negative selection could be found in wider geographic areas although their deleterious nature might also have prevented their dispersion.

Methods to Assess Population Differentiation

To quantify the degree of genetic differentiation observed at a given locus in different populations, Sewall Wright (Wright 1951) developed the theory of fixation indices or F-statistics. A population is assumed to be subdivided in subpopulations and three different F-statistics are defined. F_{IS} quantifies the amount of correlation between uniting gametes within an individual relative to the subpopulation the individual belongs to and is thus equivalent to the inbreeding coefficient of the individual. F_{IT} is the correlation between the gametes within the individual but relative to the entire population. F_{ST} is the correlation between two randomly chosen gametes in the same subpopulation relative to the entire population. It is the most common measure of population differentiation. Values of F_{ST} can typically vary between 0 in the absence of genetic differentiation and 1 when different alleles are fixed in the different populations. Several methods, reviewed in (Holsinger and Weir 2009), have been developed to estimate F_{ST} on real data. They are based on different modelling assumptions and can provide different results especially when the sample sizes are small. In addition, the comparison of F_{ST} measures between markers with different allele frequencies may be difficult as there is a strong dependency between F_{ST} and allele frequencies (Jakobsson et al. 2013).

Another way to evidence population stratification consists in performing a principal component analysis (PCA) of the individual genotype data to extract the

major axes of variation in the data (Price et al. 2006). PCA allows the visualization of the data in a space of reduced dimension and the identification of clusters of individuals who are genetically more similar and more likely to belong to the same population. Based on PCA results, it is possible to compute Euclidean distances between populations that are related to F_{ST} (McVean 2009). PCA results depend on the number of markers included and it may thus be difficult to compare results obtained with different sets of markers with different minor allele frequencies. It is indeed possible that the first principal component obtained on a sample of individuals for a set of markers is not well correlated with the first principal component obtained on the same sample of individuals with a different set of markers but correlates well with, say, the first ten principal components obtained on this latter set of markers.

More recently, with the availability of sequence data, multiple-population site frequency spectra (SFS) that record the joint distribution of single nucleotide polymorphisms (SNPs) in different populations have also been used to describe between-population structure. F_{ST} measures are summaries of multiple-population SFS. The observed multiple-population SFS can be compared to the hypergeometric distribution expected if individuals were randomly assigned to each population to see how it departs from this null distribution. More sophisticated models can also be used to test the fit to different demographic assumptions (Gutenkunst et al. 2009).

Methods to Correct for Population Stratification in Genome-Wide Association Studies

Different methods have been developed to genetically match cases and controls at common variants or to correct for the stratification in GWAS (Price et al. 2010).

The first and simplest method proposed to correct for stratification consists in computing the genomic control, λ_{GC}, defined as the median of the chi-squared association statistics computed across the different SNPs divided by its theoretical median under the null. The test statistics at each marker are then corrected by dividing them by λ_{GC} (Devlin and Roeder 1999). This correction assumes that the extent of stratification is the same over the entire genome. Thus, it works well especially in situations where the divergence between populations is mainly due to genetic drift. It does not perform as well in these areas of the genome where subpopulation differences are due to selective pressures because these regions exhibit different stratification patterns from the rest of the genome.

To avoid this assumption of a universal inflation, one relies on PCA-based methods that consist in first performing a PCA analysis of the joint samples of cases and controls and then adjusting for the top PCs in the association tests to basically match cases and controls based on the ancestry information captured by these top PCs (Price et al. 2006). A limitation of these methods however is that

they do not model family structure or cryptic relatedness that could exist within samples which would then result in type I error inflations.

On the other hand, mixed models that model the phenotypes as a mixture of fixed effects and random effects are able to deal with these complexities (Yu et al. 2006). Their applicability in GWAS is now possible thanks to the development of more efficient algorithms and their implementation in software (Zhang et al. 2010; Kang et al. 2010; Zhou and Stephens 2012).

Finally, another approach to correct for population stratification consists in first clustering individuals based on their genetic ancestries and then performing association tests within clusters (Devlin et al. 2001). Different models have been developed to cluster individuals based on genetic data; a few examples include Pritchard et al. (2000), Bouaziz et al. (2012) and Lawson and Falush (2012). Their application has however been limited because of computational constraints that make their use difficult when dealing with hundreds of thousands of markers. Recently, developments have leveraged these constraints and these approaches are now possible on GWAS data.

The different methods proposed to correct for population stratification have been extensively tested in simulation studies (see, e.g., (Bouaziz et al. 2011)) and have proven their efficiency in preventing false-positive results when testing for association with common variants.

Rare Variant Stratification on Real Data

In the last 10 years, several large-scale projects have been conducted to study the genetic diversity of different human populations. These projects have led to a major turn in population genetics that has moved from a theory-driven field to a data-driven field. It is now possible to confront the predictions obtained from population genetics theories with the observations on real data and thus to study how the theory fits with the reality (Pool et al. 2010). For this purpose, sequence data are of particular interest as, contrarily to genotyping data, they do not suffer from an assessment bias in favour of the most frequent variants.

The 1,000 Genomes Project by releasing sequence data on relatively large samples of individuals from different populations worldwide has made it possible to study rare variant stratification between and within continents (Abecasis et al. 2010). Focussing on the variants that are present twice across the entire sample (referred to as f2 variants) and that are thus probably, for most of them, recent variants revealed that they are found within the same population in 53 % of cases. When present in two different populations, they were most often seen in populations that have historical links such as the Spanish population and the populations from the Americas (Abecasis et al. 2012). To further explore this phenomenon, we downloaded the 1,000 Genomes release phase 1 low-coverage data on the 1,092 individuals from 14 different populations. After exclusion of 3 individuals who were the children of case-parent trios, we studied the distribution of variants present in various numbers

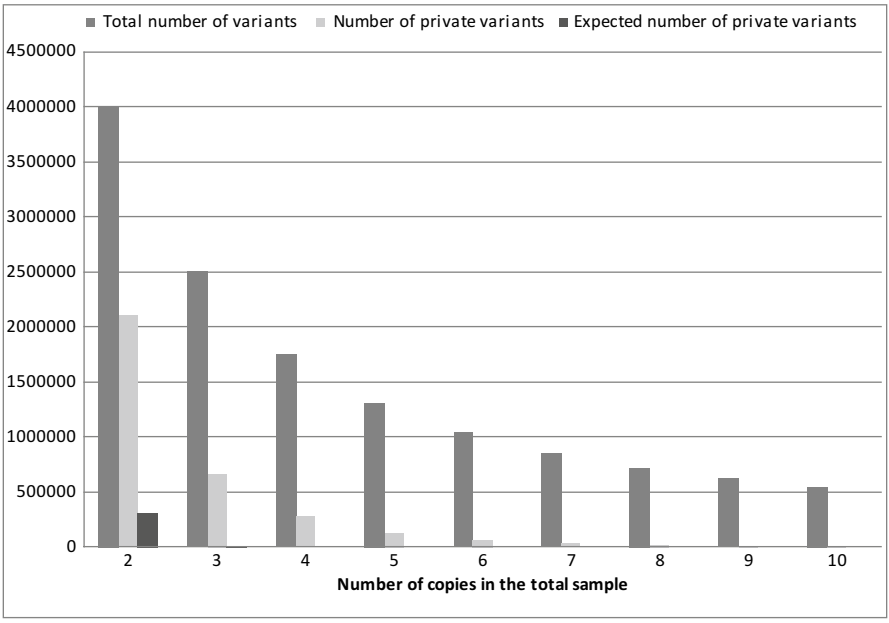

Fig. 1 The number of private variants in the different 1,000 Genomes Project populations. The number of private variants seen in 2–10 copies in the low-coverage data is reported and compared to the number expected under the null hypothesis that the alleles are randomly distributed between the populations. This null distribution is derived from the multivariate hypergeometric distribution: the expected probability $P_{p,k}$ for a variant to be seen in k copies in population p of samples size N_p individuals and to be absent in the other 13 populations of sample size N_j individuals is $P_{p,k} = \binom{2N_p}{k} / \binom{2\sum_{i=1}^{14} N_i}{k}$.

The expected number E_k of private variants seen in k copies is then $E_k = T_k \sum_{i=1}^{14} P_{i,k}$ where T_k is the total number of variants seen in k copies in the total sample

(from 2 to 10) in the remaining 1,089 individuals. We compared the number of variants that are private to a single population to the number expected from the multivariate hypergeometric distribution if variants were randomly distributed (Fig. 1). As expected and consistent with previous results, we found an important excess of private variants among the rarest variants and that a large proportion of them are not shared between the different populations from the 1,000 Genomes and even between the populations that are the closest geographically. These results might however be biased since a non-negligible fraction of the variants could be missed in the low-coverage data because of sequencing errors that can cause false negatives. The rate of false negatives depends on the real number of alleles that exist in the sample and is expected to decrease with the increasing number of alleles. These differences in false-negative rates however are not sufficient to completely explain the observed

patterns. Indeed, the fact that the lack of sharing is more pronounced for the rarest variants is still visible when sequencing errors are modelled as in the study by Gravel et al. (2011) that proposed a correction model for the multiple-population SFS. Even after this correction, the amount of sharing between continental populations for the rarest variants was significantly reduced compared to the amount expected under the hypothesis of a random assignment of individuals. Sharing was closer to the expectations for the most common variants.

Despite the efforts in the last few years towards sequencing larger numbers of individuals, the amount of sequence data needed to study population stratification at a finer scale than the continental scale is still too poor. However, information on fine-scale stratification may be gained from the different GWAS that have been performed so far and for which very large samples of individuals have been genotyped on SNP chips. Even if these studies suffer from an ascertainment bias in favour of the most common variants, they can still provide some clues on rare variant stratification in more geographically restricted regions. Taking advantage of the high content in rare variants of the Affymetrix 500 K SNP chip used in the WTCCC1 study, we were able to show that rare variants with a minor allele frequency (MAF) $\leq$1 % are not stratified in the same way in the UK control population as the common ones with MAF >5 % and that the low-frequency variants with an MAF in between these two values also display different stratification patterns (Babron et al. 2012). The regional maps obtained by plotting the first two principal components (PCs) of the PCA conducted on these different sets of variants are indeed very different (Fig. 2). The top PC extracted from the rare variant set shows very poor correlations with any PCs or combination of PCs from the two other variant sets and the rare variant stratification appears much stronger than the other two.

Impact of Rare Variant Stratification on Association Tests

The fact that rare variants are not stratified in the same way as common variants and display stronger stratification patterns calls for caution in the interpretation of association tests. Indeed, false-positive results are likely to arise due to a lack of appropriate genetic matching between cases and controls. Several studies converge to show that the performances of the methods developed to correct for common variant stratification are reduced when testing for association with rare variants.

Using the 1,000 Genomes Project sequence data on European and African samples, Zhang et al. (2013) found that PCs derived from low-frequency variants (1 % < MAF < 5 %) were better able to separate the two continental groups than those derived from the common variants. However, when used in association tests to correct for stratification, PCs derived from the low-frequency variants could lead to some power losses due to an overadjustment. Moreover, the top PCs derived from rare variants were found less efficient at capturing the continental stratification and rare variant association tests adjusted on these PCs had inflated type I error rates. Using simulated mini-exome data from the Genetic Analysis Workshop 17, He et al.

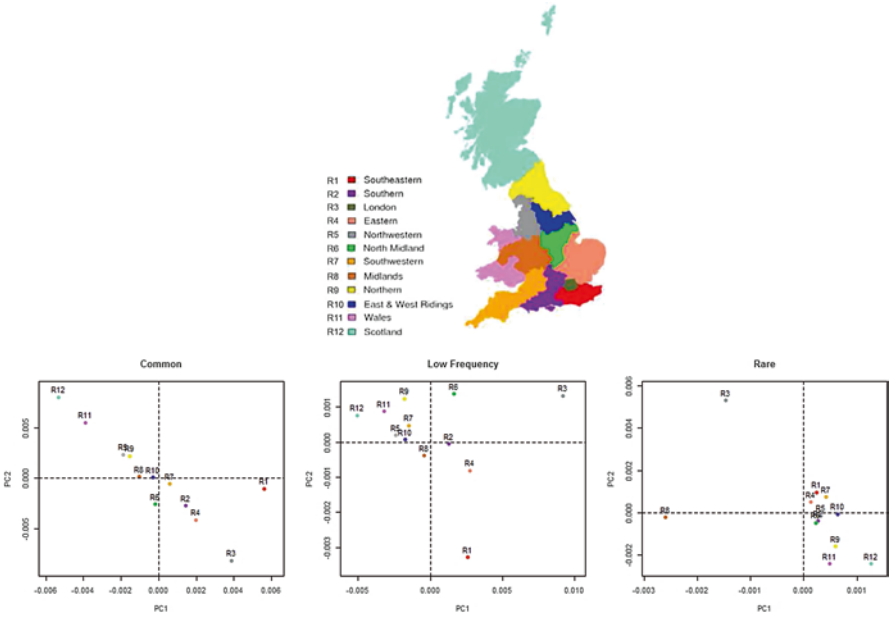

Fig. 2 Patterns of stratification for common, low-frequency and rare variants in the WTCCC1 control population from the UK. This figure displays the mean PC1 and PC2 scores in each of the 12 UK regions represented in the dataset. Common variants have a minor allele frequency (MAF) >5 %, the low-frequency variants have an MAF in the range between 1 and 5 %, and the rare variants have an MAF ≤1 % (Reproduced from Figure 3 in Babron et al. (2012))

(2011) also found that PCA was able to reduce false-positive rates much more effectively in common SNPs than in rare SNPs. If a total of 144 common SNPs and 44 rare SNPs were falsely declared significant before PCA adjustment, these numbers dropped respectively to 0 and 21 after adjustment. We reached the same conclusion concerning the difficulty to correct for rare variant stratification with PCA in our analysis of the WTCCC1 Affymetrix 500 K data stratified on the MAF (Babron et al. 2012). We found both on type 2 diabetes and on simulated data that, even if PCs computed on rare variants were used to adjust rare variant association tests, type I errors were inflated. This was also noticeable, albeit to a lesser extent, for the low-frequency variants. Mixed models such as the ones implemented in EMMAX (Kang et al. 2010) that were found to perform better than PCA for common variants, especially in situations where there exist cryptic relatedness, were expected to also perform better when studying rare variants as carriers of the same rare variants are likely to be remotely related. However, our results showed the opposite trend. The inflation factors obtained when using EMMAX on the rare variants were similar to those obtained without any stratification correction and thus slightly worse than the PCA-based corrections.

Using simulations of genotype and quantitative traits on spatial grids, Mathieson and McVean (2012) were also able to show that rare and common variants are expected to exhibit differential spatial distributions especially in situations where the

nongenetic disease risk has a sharp spatial distribution. In these situations, association tests performed with rare variants showed a stronger inflation under the null hypothesis than association tests performed with common variants. However, when the nongenetic risk had a wide and smooth spatial distribution, this was no longer the case and rare variants showed less inflation than common ones. The behaviour of the correcting methods for population stratification was also shown to depend on these distributions of nongenetic risks. They worked well in the latter situations where the nongenetic disease risk had a wide distribution and did not work for the small and sharp distribution of risk. Because many of the methods proposed to test for association with rare variants combine the information across multiple variants within a gene, the authors also investigated the effects of stratification on gene-based tests and found that in the sharp nongenetic risk situation, the problem still remained but was reduced as the number of aggregated variants within a gene increased. Moreover, population stratification adjustments derived for single SNPs may not be appropriate for gene-based tests as genes may have a different composition of rare and common variants and thus are likely to exhibit different stratification patterns. Through both theoretical and empirical investigations, Liu et al. (2013) showed that the inflation factor due to population stratification of gene-based tests was expected to depend on the number of variants within the genes and on their MAF distribution. Applying a genomic control correction for stratification has proven to be very inefficient in this context as this correction assumes a universal inflation over the different genes. PCA-based corrections were found to be superior in simple scenarios of two distinct populations but their performances decreased for more subtle scenarios with several subpopulations.

These different studies suggest that rare variant stratification remains a problem that still needs to be treated in association tests. New methodological developments are necessary to avoid false-positive conclusions. This is definitively an issue that investigators planning to use sequencing methods to assess rare variant association with complex traits need to be aware of and to account for in their study design.

References

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073. doi:10.1038/nature09534

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Babron MC, de Tayrac M, Rutledge DN, Zeggini E, Genin E (2012) Rare and low frequency variant stratification in the UK population: description and impact on association tests. PLoS One 7(10):e46519. doi:10.1371/journal.pone.0046519

Bouaziz M, Ambroise C, Guedj M (2011) Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PLoS One 6(12):e28845. doi:10.1371/journal.pone.0028845

Bouaziz M, Paccard C, Guedj M, Ambroise C (2012) SHIPS: spectral hierarchical clustering for the inference of population structure in genetic studies. PLoS One 7(10):e45685. doi:10.1371/journal.pone.0045685

Cavalli Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37(11):1243–1246. doi:10.1038/ng1653

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55(4):997–1004

Devlin B, Roeder K, Bacanu SA (2001) Unbiased methods for population-based association studies. Genet Epidemiol 21(4):273–284. doi:10.1002/gepi.1034

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108(29):11983–11988. doi:10.1073/pnas.1019276108

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5(10):e1000695. doi:10.1371/journal.pgen.1000695

He H, Zhang X, Ding L, Baye TM, Kurowski BG, Martin LJ (2011) Effect of population stratification analysis on false-positive rates for common and rare variants. BMC Proc 5(Suppl 9):S116. doi:10.1186/1753-6561-5-S9-S116

Hirszfeld L, Hirszfeld H (1919) Essai d'application des méthodes au problème des races. Anthropologie 29:505–537

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10(9):639–650. doi:10.1038/nrg2611

Jakobsson M, Edge MD, Rosenberg NA (2013) The relationship between FST and the frequency of the most frequent allele. Genetics 193(2):515–528. doi:10.1534/genetics.112.144758

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42(4):348–354. doi:10.1038/ng.548

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336(6082):740–743. doi:10.1126/science.1217283

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43(4): 520–526

Lawson DJ, Falush D (2012) Population identification using genetic data. Ann Rev Genom Hum Genet 13:337–361. doi:10.1146/annule-génome-082410-101510

Liu Q, Nicolae DL, Chen LS (2013) Marbled inflation from population structure in gene-based association studies with rare variants. Genet Epidemiol 37:286–292. doi:10.1002/gepi.21714

Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44(3):243–246. doi:10.1038/ng.1074

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5(10):e1000686. doi:10.1371/journal.pgen.1000686

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genom Res 20(3):291–300. doi:10.1101/gr.079509.108

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909. doi:10.1038/ng1847

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459–463. doi:10.1038/nrg2813

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Wright S (1951) The genetical structure of populations. Ann Eugenics 15:323–354

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208. doi:10.1038/ng1702

Zhang Y, Guan W, Pan W (2013) Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 37(1):99–109. doi:10.1002/gepi.21691

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4):355–360. doi:10.1038/ng.546

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44(7):821–824. doi:10.1038/ng.2310

Use of Appropriate Controls in Rare-Variant Studies

Audrey E. Hendricks

Introduction

When designing a case–control study, researchers must decide whether to gather and sequence or genotype controls in parallel with cases or whether to only sequence or genotype cases and to plan on using external, perhaps publically available control data. When using rare-variant genotyping chips, gathering and using internal controls may be the obvious choice as the cost per subject is relatively low although there is still the cost of recruiting or finding a suitable control set. There may be more reason to use external controls for sequencing due to its relatively high cost. Ultimately, the choice often comes down to a balance between available funds, resources, and the number and uniqueness of the cases available for sequencing or genotyping. In addition, other reasons may prompt researchers to use external controls. Researchers may choose to boost their study's sample size and potentially the power by adding in external, publically available controls to an existing internal control set.

Here, we address how to conduct an appropriate case–control study using internal as well as external controls.

Case–Control Studies

Careful consideration of appropriately using controls is especially important as systematic differences between cases and controls can cause bias, here defined as inaccurate genotype frequencies or differential genotype missingness, and manifest

A.E. Hendricks (✉)
University of Colorado Denver, Denver, CO, USA
e-mail: audrey.hendricks@ucdenver.edu

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_17

into false associations or even hide true associations. In this section, we first highlight the differences between internal and external controls and then comment on the possible sources and solutions of bias in general and those specific to external controls.

Internal Versus External Controls

Given limited resources and the cost of recruiting subjects and producing data for any study, there has always been a balance to the number of cases and controls to gather to achieve optimal power. This balance often depends on several parameters including the number of samples already recruited, the cost of recruiting further cases or controls, the prevalence of cases, the availability of appropriate external controls, and overall study resources.

As the prevalence of a case set decreases, so does the difficulty in recruiting further cases. In addition, for case sets with low prevalence, there are often fewer similar case sets from which to replicate findings. For instance, the UK10K Severe Childhood Onset Obesity Project (SCOOP) sample (described further in Chap. 9) used in the obesity arm of the UK10K project consists of Caucasian children with an age-adjusted BMI of greater than three standard deviations above the mean and an age at onset of less than 10 years. A similar BMI measure in adults would be classified as morbidly obese. The young age at recruitment means that the children have a relatively short duration of environmental exposure (compared to obese adults) and suggests that the sample may be enriched for genetic causes of obesity. In fact, severe childhood obese samples from which SCOOP is a subset have pathogenic mutations in *MC4R*, the gene that causes the most common form of monogenic obesity, at a rate of 2–5 times higher than the general population suggesting severe childhood obesity samples such as SCOOP are indeed enriched for rare genetic causes of obesity (Farooqi and O'Rahilly 2006). As SCOOP subjects are more unique than a set of obese or even morbidly obese adult subjects, the overall prevalence of SCOOP cases is much lower than common obesity. The uniqueness of the SCOOP sample is an example where there may be few similar samples with which to replicate findings. Thus, focusing resources on sequencing as many cases as possible while using or supplementing with externally available controls may be prudent.

Many possible confounders exist when conducting case–control analysis including differences due to ethnicity, technology, or platform. Using internal controls provides an opportunity to mediate or prevent some of the bias before completing the case–control analysis. Controls can be matched, upon entry to the study, to cases on reported or genetic (if available) ethnicity to help prevent population stratification. Further, cases and controls can be equally distributed on the same genotyping plates or mixed and sequenced in the same lanes and on the same days using the same technology to help prevent any bias due to differences in technology, targets, or other batch effects.

Below, we discuss further possible sources of bias as well as steps to prevent and address this bias. While some sources of bias will most likely only be seen when using external controls, there are other biases, such as population stratification, common to internal and external controls. Steps should be taken to prevent or correct possible sources of bias no matter which control set is used.

General Quality Control

As in all genetic association studies, sample- and variant-level quality control is important to arrive at an unbiased and robust result. Quality control (QC) is particularly important when using external controls where there is likely more non-trait-related difference between cases and controls that can lead to biased results.

In this section we outline steps that researchers can take to help prevent and control for bias in case–control studies. These steps can be taken as a broad guideline of QC and checks. Producing plots of various QC measures throughout can help identify and fix problems relatively early in the process. While we outline basic considerations, the particular exclusion values will and should be driven by the unique nature of each study.

Sample Quality Control

Common sample quality control includes detecting contaminated samples, finding cryptic relatedness, looking for ethnic outliers, completing concordance analysis with other available genetic datasets, and calculating descriptive statistics on the sample level such as the total number of variants, the number variants by minor allele frequency group, etc. Whether rare-variant data is coming from sequencing or rare-variant genotyping chips, there are usually enough common variants to complete the sample QC using methods established during the genome-wide association study (GWAS) era. Whole-exome and whole-genome sequencing produce plenty of common variants from which established protocols used to identify and potentially exclude samples can be used. Rare-variant genotyping chips, such as the exome chip, include a set of common variants specifically chosen to identify ethnic outliers that can also be used for other checks such as relatedness.

The ratio of the number of heterozygote calls over the number of homozygote calls in total (het/hom ratio), or the number of alternate homozygous calls (het/alt hom ratio) is a commonly used measure for detecting sample contamination. Given contamination caused by mixed samples, we expect the number of heterozygote calls to increase as differences in genotypes would most often mix to be called as a heterozygote genotype. Thus, a relatively high het/hom ratio compared to the rest of the samples can indicate sample contamination. Often a threshold of 3SD away from the sample mean is used to identify outliers and possibly contaminated subjects.

However, other reasons might exist for a sample to have a relatively high het/hom ratio. For instance, an ethnically different or admixed subject may have a higher proportion of heterozygote genotypes compared to the rest of the sample. Thus, researchers should also look at whether subjects with a high het/hom ratio are also ethnic outliers.

Recently, researchers have developed likelihood methods for identifying sample contamination using next-generation sequencing data. ContEst was developed in 2011 by Cibulskis et al. (2011), and verifyBamID (Jun et al. 2012) was developed by Jun et al. in 2012. VerifyBamID can also use available array-based genotyping alone or with sequencing data to identify DNA sample contamination. VerifyBamID provides a sequence only or a sequence + array estimate of contamination (called freemix and chipmix, respectively) and has several recommendations including using a chipmix or freemix value >0.02 to suggest further follow-up of a subject for possible contamination (see http://genome.sph.umich.edu/wiki/VerifyBamID for further details).

Many samples now being sequenced or genotyped for rare variants have already been genotyped for other studies. When possible, the concordance of the genotype calls from previous sources should be compared with the new genotype calls for each individual. In addition to helping to identify contaminated samples, a concordance check can identify sample or ID swaps.

After contaminated samples are identified and removed from the analysis, samples should be checked for ethnic outliers and cryptic relatedness. Using a set of high-quality common variants available from sequencing or variant chips, established methods such as principal component (PC) analysis and IBD estimation can be used to detect ethnic outliers and cryptic relatedness, respectively. For instance, EIGENSTRAT (Price et al. 2006) can be used to create PCs from an unrelated reference set such as HapMap Phase III (International HapMap Consortium et al. 2010), 1000 Genomes (Genomes Project Consortium et al. 2012), or another set of unrelated samples with known ethnicities. The case–control data can then be mapped onto the PCs to detect subjects outside of the expected ethnic group. For the purpose of detecting ethnic outliers, it is important to only use the reference set to calculate the PCs and then to project those PCs back onto the case–control samples. Including the case–control samples in the original estimation can produce PCs driven not by ethnicity but instead by technology differences or relatedness within the cases or controls.

Detecting ethnic outliers is especially important when completing case–control analysis on rare variants where rare variants may be seen more often or perhaps only in a particular ethnic group. Later, we will discuss controlling for more fine-scale ethnic variation within the case–control sample, and details for rare structural variants are described in detail in Chap. 6.

Programs, such as PLINK (Purcell et al. 2007), can be used to estimate IBD in order to find cryptic relatedness within cases or controls. An estimated IBD greater than or equal to 0.125 to identify third degree or closer relatives is often used as a threshold to exclude related individuals from further analyses. Given a large number of related individuals, researchers may prefer to use a statistical model, such as

mixed models, to control for relatedness instead of excluding samples. Recently, these methods have been developed or extended to incorporate related samples in gene- or region-based tests (Chen et al. 2013).

As when identifying ethnic outliers, using a set of high-quality variants is essential for estimating IBD as well. Variant filters should be applied both within the reference set and the case–control sets independently. Then, IBD can be estimated directly from the high-quality variants from the case–control set, and the overlapping variants from the reference and case–control set can be used to identify ethnic outliers. We recommend several filters including limiting to common variants (MAF>5 %), HWE *p*-value >0.0001, imputation quality >0.95 or 0.99, strict VQSLOD threshold, limiting to the bait regions for exome sequencing, and limiting to regions easily accessible to short-read sequencing, to name a few. The specific filters and thresholds will depend on the type of data (e.g., genotype chip, whole-genome or whole-exome sequencing, etc.).

In addition to detecting contamination, ethnic outliers, and relatedness, other sample QC measures such as median sequence depth, transition vs. transversion ratio (ti/tv ratio), number of variants in total and by MAF group, and sample call rate can be calculated to check for batch effects or sequencing/genotyping errors. Once sample level measures are attained, problem samples or changes to the sequencing protocol or chemistry can often be identified by plotting the measures. Using a simple plot of the variable of interest on the *y*-axis and the sample by date sequenced or genotype plate on the *y*-axis, such as in Fig. 1, is a simple way to identify batch effects or potential inconsistencies in the data. We have provided a couple references for more thoughts on general quality control (Turner et al. 2011; Do et al. 2012). In addition, we provide a brief flow chart of basic QC and analysis steps (Fig. 2).

Variant Quality Control

Limiting the set of variants used for analysis to high-quality variants in both cases and controls is particularly important. Variants should first pass the minimum genotyping or sequencing calling thresholds as discussed in Chaps. 3 and 4, respectively. Researchers may then want to apply more stringent variant filters or apply further per subject variant filters (e.g., genotype quality).

Some filters, such as a genotype call rate, imputation quality thresholds, or MAF thresholds, can be used for both sequencing and genotyping data. Other variant quality filters are specific to only sequencing or genotyping data. For sequence data, regions that are less accessible to short-read sequencing (e.g., low or high depth of coverage compared to the average, too many reads with zero mapping quality, low average mapping quality) should be removed. These regions can be identified within each dataset separately, or an externally defined set, such as the regions defined by 1000 Genomes (ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/phase1/analysis_results/supporting/accessible_genome_masks), can be used. For single-variant tests, the particular thresholds used to filter variants can be explored post-analysis

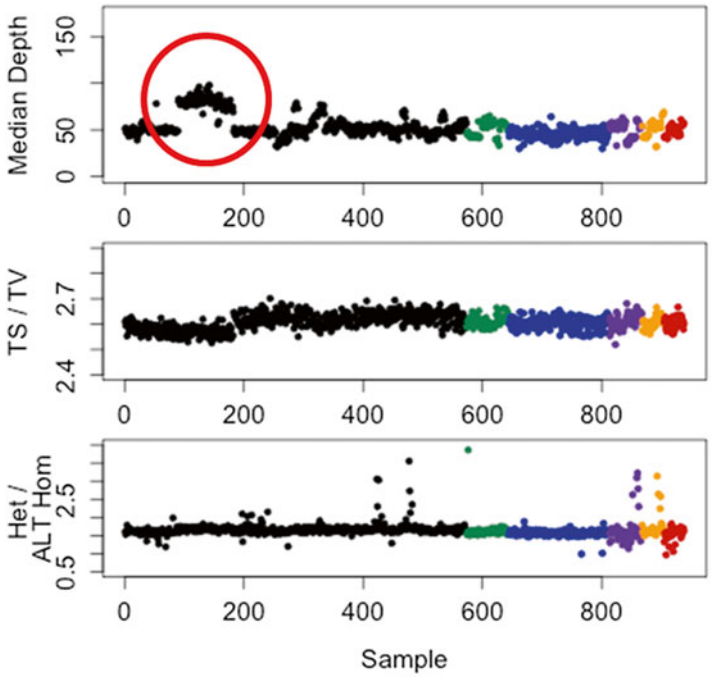

Fig. 1 Example plots of QC measure (*y*-axis) by sample ID (*x*-axis). *Red circle* indicates a sequencing chemistry change identified by the QC plots

using QQ plots as discussed in Sect. "Post-analysis Quality Control." For gene-/region-based tests, variant filters must be applied before analysis.

Variant quality, in general, can often be greatly improved using imputation. Imputation for low-depth whole-genome sequencing is widely established and used (DePristo et al. 2011). Although using imputation for high-depth whole-exome sequencing is less common, it is beneficial especially outside of the target regions where the depth is no longer high (Pasaniuc et al. 2012). We discuss imputation further in the next section.

Imputation

Imputation within the existing set of sequences without any reference panel, called genotype refinement, is commonly used to improve genotype calling within sequence data (DePristo et al. 2011). This is most useful for low-depth sequencing as raw calls often have a relatively high degree of uncertainty. Although not needed for most of the high-depth whole-exome sequencing regions covered by the sequencing baits, genotype refinement can help to refine regions just outside of the baits where the read depth is lower increasing the quality of the genotype calls in these regions. Further, it has been shown to be possible to impute whole-exome

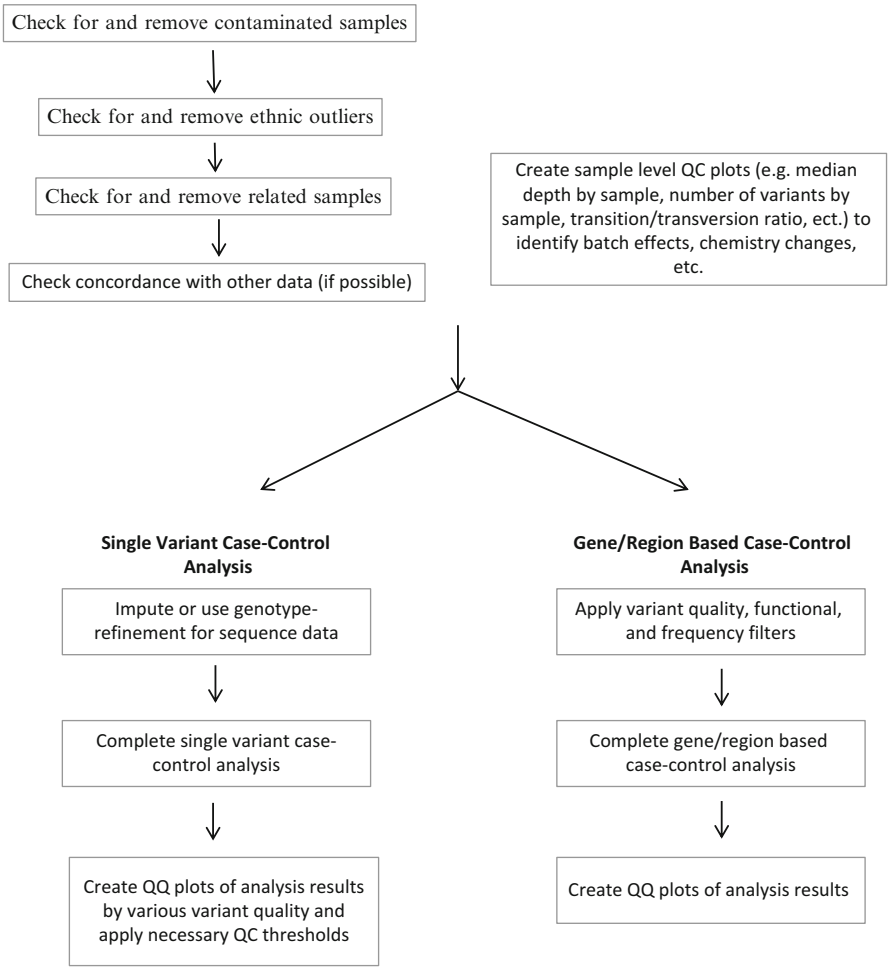

Fig. 2 Flow chart of basic QC and analysis steps

sequencing genome-wide by using a reference panel such as 1000 Genomes or UK10K cohorts and the low-coverage off-target reads produced by whole-exome-sequenced samples (Pasaniuc et al. 2012). This enables analysis of low-frequency and common variants genome-wide from whole-exome sequencing. To ensure the best quality, we recommend removing contaminated samples and using per subject variant filters on genotype quality prior to genotype refinement.

Single-variant case–control analyses can be limited to variants with good imputation quality within the cases and controls ensuring that the genotypes are fairly accurate in each group. This helps to limit false-positive results due to differences in sequencing coverage between cases and controls and is especially important when using external controls.

While genotype refinement and imputation work well with low-frequency and common variants, the accuracy of the refinement decreases substantially as the MAF decreases (DePristo et al. 2011; Abecasis et al. 2012; Li et al. 2011). Further, genotype refinement is not possible, by definition, for singleton variants. Thus, it may be better to use other variant quality thresholds and measures other than imputation quality for gene-based tests that are often performed using primarily rare variants. Imputation of rare variants is discussed further in Chap. 10.

Population Stratification

As discussed in Chap. 19, population stratification exists across variants of all MAF. To control for population stratification in association tests for low-frequency or common variants, we can use existing and proven methods including a selection of principal components within the regression model (Price et al. 2006). This will help to correct for population stratification on a moderate to large ancestry scale, for instance, the north to south or east to west gradients within Europe. However, rare variants can show different, perhaps more focused, population stratification patterns that may not be well captured using traditional methods such as principal components (Mathieson and McVean 2012).

As such, controls (both internal and external) should be closely matched by ethnicity to cases. This is especially important for case–control analyses that include rare variants where methods to correct for population stratification are still being developed.

Post-analysis Quality Control

After running case–control association analysis, researchers should again check the data looking for any patterns or indication that the results are biased or incorrect. Quantile–quantile plots, or QQ plots, are probably the most commonly used method for checking the results of hundreds or thousands of association tests. A QQ plot compares the expected distribution of test statistics to the observed distribution. Since the majority of genome- or exome-wide association tests will have no true association, most of the observed test statistics should match the expected distribution under the null hypothesis of no association.

Often, the χ^2 distribution is used to calculate the expected distribution. Knowing this, it is important to know and understand the assumptions and properties of the distribution on which the association test is based. Many case–control tests do use a chi-squared distribution. The chi-squared test assumes that the sample size with relation to the variant frequency is large enough to have an accurate approximation. Too small of a sample size or variant frequency will produce inaccurate test statistics (Larntz 1978). Fisher's exact test is another commonly used case–control test for rare variants. As the name implies, Fisher's exact test uses an exact calculation instead of approximation to calculate the p-value eliminating the assumption of a

large sample size to variant frequency. However, Fisher's exact test is known to be overly conservative, producing a distribution of *p*-values that are larger than expected under the null hypothesis of no association (Berkson 1978). This conservative property is due to using a discrete test statistic with a fixed significance level and is most pronounced under small sample sizes or variants with rare frequencies. Permutation is another commonly used method for calculating the appropriate test statistic distribution when model assumptions may not be met or the exact null distribution is unknown (Hirschhorn and Daly 2005).

Preventing Bias When Using Internal Controls

When using internal controls, simple planning can help prevent future bias or confounding. While being chosen, controls should be ethnically matched to cases whether through self-reported status or, when possible, using ethnicity defined from previous genetic data. Cases and controls should be equally balanced throughout sequencing dates and lanes or throughout genotyping plates to help alleviate bias caused by batch effects or chemistry changes.

Controlling for Bias When Using External Controls

While many of the QC steps discussed above are applicable and important whether using internal or external controls, special care must be given when using external controls. It is important to remember that the ultimate goal is to arrive at a set of variants and subjects over which the cases and controls are well matched and possible sources of bias are controlled for or alleviated.

Sequencing Studies

When using external controls, it is likely that the samples were not only sequenced at a different time and place but also using a different technology or platform. For sequencing, this means that regions of the genome were targeted and sequenced to different depths in controls and cases. For instance, the cases may have been whole-exome sequenced at a high depth, whereas the controls (perhaps from the 1000 Genomes or the cohort arm of the UK10K project) may have been whole-genome sequenced at a low depth. This was the study design for the UK10K project where 6,000 extreme phenotype samples were whole-exome sequenced at a depth of ~50× and 4,000 cohort samples were whole-genome sequenced at a depth of ~6×. Alternatively, the cases and controls could both be whole-exome sequenced using different exome target sets. For instance, the whole-exome sequencing of the UK10K samples used Agilent SureSelect Human All Exon 50 Mb array, whereas

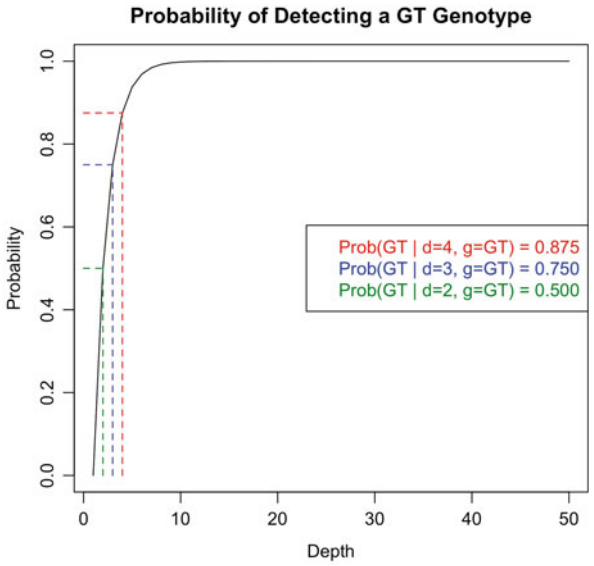

Fig. 3 Probability of detecting both alleles given a depth and a heterozygous genotype assuming that both alleles are equally attainable through sequencing

the NHLBI Exome Sequencing Project used one of three target solutions: Agilent SureSelect Human All Exon Kit v2 [31 Mb], NimbleGen-designed custom RefSeq/CCDS design [28 Mb], and NimbleGen SeqCap EZ v1 [32 Mb] (Tennessen et al. 2012; Futema et al. 2012). Differences in depth or regions targeted will cause bias in the numbers of variants detected and, more importantly for single-variant- and region-based association tests, will cause bias in whether an alternate allele is detected at all in the regions where the coverage is drastically different between cases and controls. An example of this is shown in Fig. 3 where we show the probability of detecting both alleles at least once given a certain depth and a heterozygous genotype with the assumption that both alleles are equally captured by the sequencing technology. We can see that while the probability quickly increases to one, the probability is much lower for low-depth sequencing.

Just this crude difference in the probability of detecting a heterozygous genotype can cause an overall inflation in the distribution of test statistics as shown in Fig. 4.

As discussed previously in Sect. "Imputation," genotype refinement can greatly improve the accuracy of genotype calls for regions with low depth. This can, in turn, reduce the inflation in the test statistic distribution as seen in Fig. 4. It may also be possible to control for differences in average depth by including average sample depth as a covariate in the regression (Garner 2011).

Since the accuracy of genotype refinement decreases, in general, as the MAF decreases, more stringent variant quality filters (e.g., depth, mapping quality, VQSR, genotype quality, etc.) may be used instead to ensure that rare variants are of high quality for gene-/region-based methods. In addition, there are methods that can incorporate variant quality directly into gene-based methods (Asimit et al. 2012). More research is needed into the area of gene-based tests especially for situations where

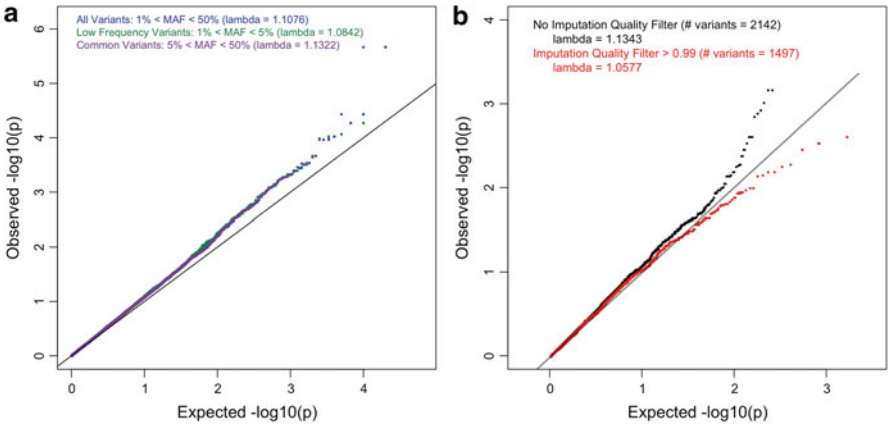

Fig. 4 QQ plots for single-variant associations. (**a**) Logistic regression using a likelihood ratio test on 10,000 simulated variants using a biased probability distribution based on severe differences in depth between cases (50×) and controls (4×). The colors represent different MAF thresholds. Blue: 1 % < MAF < 50 %, green: 1 % < MAF < 5 %, and purple: 5 % < MAF < 50 % and (**b**) results from running a score test in SNP test on real data from the UK10K project using 667 SCOOP subjects as cases whole-exome sequenced at a high depth (~50×) and 2,432 cohort subjects as controls whole-genome sequenced at a low depth (~6×). Variants were filtered to chromosome 20, to have an MAF >1 %, a PASS status from the UK10K pipeline (VQSLOD truth sensitivity of 99.5 %), and to exclude regions in the genome not easily accessible to the sequencing as defined by the 1000 Genomes project

sequencing platform or technology for the external controls differs greatly from that used for the cases. For instance, it is unknown whether current methods could be modified to enable low-depth whole-genome sequences to be used as controls for high-depth whole-exome sequences within the context of gene-/region-based analyses.

Rare-Variant Genotyping Chip Studies

As previously described, when possible, genotyping for cases and controls should be planned out ahead of time to prevent bias due to different technologies, chip versions, or batches that results in differential genotype frequencies or missingness. For genotype chips, this would first manifest as differences in the signal intensity plots that are used to call the genotypes (Ziegler and König 2010). However, it is possible that due to various factors, the cases and controls might be genotyped using different versions of the genotyping chip or at different times or places all of which can introduce bias. To correct for differences in the signal intensity plots due to using different variant genotyping chip versions, the genotypes for each chip version can first be called separately. This will help to prevent biased calls due only to differences in the chip version probes. Then, previously poorly called rare variants can be recalled with zCall (Goldstein et al. 2012). The same process can be applied if differences are seen or suspected due to genotyping in different centers or at

different times. As described above, QC metrics and QQ plots should be used to identify inflation or batch effects as soon as possible in the process.

Diseased Controls

As many studies have focused on sequencing or genotyping sets of diseased samples, an obvious and important question is whether these samples can be used in some capacity as controls for other case sets. The diseased control sets could be another arm of the same study as the case set, such as within the UK10K project, or could be a publically available dataset. Both the UK10K project and the NHLBI Exome Variant Server, which were chosen to be from continuous or disease extremes from population cohort studies, are publically available through application to EGA (https://www.ebi.ac.uk/ega/) or dbGaP (http://www.ncbi.nlm.nih.gov/gap), respectively.

In addition to challenges discussed previously (e.g., population stratification, different technologies, etc.), there may be some overlap of the disease etiology between the cases and diseased controls. This overlap in the disease etiology may diminish or even completely remove the ability to detect variants, genes, or regions associated with cases status that are also associated with the control disease. Further, depending on the hypothesis being tested, it may be unclear whether a significant association indicates that the variant, gene, or region is associated with case status or actually associated with control status instead.

To try to prevent or reduce a decrease in power due to overlapping etiology, researchers can focus on using diseased controls with little known overlap. Further, researchers can exclude samples based on existing covariate information if available. For instance, controls could be excluded on the basis of BMI for use as controls for the SCOOP sample set.

Multiple Control Sets

Researchers may want to combine control sets to potentially increase power. The control sets might consist of some combination of internal controls, diseased external controls, and population external controls. In addition to potentially increasing the power to detect a true association, using multiple control sets can help to clarify that a significant association is due to the case status rather than diseased controls. Several control sets can be combined either before or after association analysis. The similarity of the control sets with regard to sequencing technology or platform will help guide when control sets should best be combined. Drastically different technologies, such as whole-exome sequencing at a high depth and whole-genome sequencing at a low depth, will likely be best combined post-association analysis. Samples with high-depth whole-exome sequencing even using slightly different coverage may be able to be combined prior to variant calling and analysis using just the intersection of high-quality variants. Research is currently being done to determine the best time and method to combine control sets.

Validation and Replication

Thorough QC and uninflated QQ plots do not ensure that promising associations are unbiased. In general, but especially when using external controls, it is necessary to validate and then to replicate any findings. Validating signals means to confirm, usually through genotyping the original cases and controls, that the genotype calls and thus subsequent association signal are accurate within the original sample and were not due to biases in sequencing, genotyping, or imputation. Once variants have been validated, association signals should be replicated, when possible, in independent case–control sets. Single-variant associations can often be genotyped, whereas gene-/region-based associations should usually be sequenced. Similar to when using internal controls, cases and controls should be matched and mixed throughout the process as much as possible to alleviate bias by ethnicity, technology, and batches prior to analysis. Successful validation and replication provide additional assurance that the association signal is not likely due to hidden bias. Rare-variant replication is discussed further in Chap. 17.

Conclusion

We have outlined several important steps and considerations necessary when completing case–control analysis using sequence data or rare-variant genotyping chips. When strict and thorough QC and assessment are used throughout the process, valid association results can likely be attained. Like with GWAS, validation and replication of results are necessary for additional confirmation.

References

Abecasis GR, Auton A, Brooks LD et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Asimit JL, Day-Williams AG, Morris AP, Zeggini E (2012) ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73(2):84–94. doi:10.1159/000336982

Berkson J (1978) In dispraise of exact test—do marginal totals of 2×2 table contain relevant information respecting table proportions. J Statist Plan Infer 2(1):27–42. doi:10.1016/0378-3758(78)90019-8

Chen H, Meigs JB, Dupuis J (2013) Sequence kernel association test for quantitative traits in family samples. Genetic Epidemiology 37(2):196–204

Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 27(18):2601–2602. doi:10.1093/bioinformatics/btr446

DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498. doi:10.1038/ng.806

Do R, Kathiresan S, Abecasis GR (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21(R1):R1–R9. doi:10.1093/hmg/dds387

Farooqi S, O'Rahilly S (2006) Genetics of obesity in humans. Endocr Rev 27(7):710–718. doi:10.1210/er.2006-0040

Futema M, Plagnol V, Whittall RA et al (2012) Use of targeted exome sequencing as a diagnostic tool for Familial Hypercholesterolaemia. J Med Genet 49(10):644–649. doi:10.1136/jmedgenet-2012-101189

Garner C (2011) Confounded by sequencing depth in association studies of rare alleles. Genet Epidemiol 35(4):261–268. doi:10.1002/gepi.20574

Genomes Project Consortium, Abecasis GR, Auton A et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632

Goldstein JI, Crenshaw A, Carey J et al (2012) zCall: a rare variant caller for array-based genotyping—genetics and population analysis. Bioinformatics 28(19):2543–2545. doi:10.1093/bioinformatics/bts479

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95–108. doi:10.1038/nrg1521

International HapMap Consortium, Altshuler DM, Gibbs RA et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58. doi:10.1038/nature09298

Jun G, Flickinger M, Hetrick KN et al (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91(5):839–848. doi:10.1016/j.ajhg.2012.09.004

Larntz K (1978) Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. J Am Stat Assoc 73(362):253–63. doi:10.2307/2286650

Li L, Li Y, Browning SR et al (2011) Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One 6(9):e24945. doi:10.1371/journal.pone.0024945

Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44(3):243–246. doi:10.1038/ng.1074

Pasaniuc B, Rohland N, McLaren PJ et al (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44(6):631–635. doi:10.1038/ng.2283

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909. doi:10.1038/ng1847

Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. doi:10.1086/519795

Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69. doi:10.1126/science.1219240

Turner S, Armstrong LL, Bradford Y, et al (2011) Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet Chapter 1:Unit1.19. doi:10.1002/0471142905.hg0119s68

Ziegler A, König IR (2010) A statistical approach to genetic epidemiology: concepts and applications. Wiley-VCH, Germany

Trans-Ethnic Fine-Mapping of Rare Causal Variants

Xu Wang and Yik-Ying Teo

Introduction

GWAS have achieved great success in identifying genetic variants that are associated with complex diseases and human traits. To date, there are more than 4,000 genetic variants reported with genome-wide significant evidence in more than 1,500 publications, according to the US National Human Genome Resource Institute (NHGRI, http://www.genome.gov/gwastudies/). Despite these remarkable successes, the identified variants only explain a small proportion of the trait heritability, such as height, where only 5 % of phenotypic variance has been explained by the identified loci despite a heritability estimate of 80 % (Visscher 2008).

Designed on the basis of the common disease–common variant hypothesis, that complex disease or human trait susceptibility is modestly influenced by genetic variants that are present in the population with minor allele frequency (MAF)

X. Wang (✉)
Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore
e-mail: a0023748@nus.edu.sg

Y.-Y. Teo
Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, Singapore

Life Sciences Institute, National University of Singapore, Singapore, Singapore

Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456, Singapore

Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, DOI 10.1007/978-1-4939-2824-8_18

Table 1 Population genetic characteristics of common and rare variants

Characteristics	Common variants	Rare variants
MAF threshold	>5 %	<1 %
Time of mutation	Ancient	Recent
Expected effect size[a] (Wang et al. 2012)	Moderate	Large
	$1.0 < RR < 2.0$	$RR > 2.0$
	$\Delta < 0.4$	$\Delta < 0.4$
LD structure	High LD ($r^2 > 0.8$) with neighboring common variants	Weak LD ($r^2 < 0.3$) with neighboring variants

[a]RR stands for relative risk, which is relevant to case–control studies, while Δ indicates the standardized difference of a quantitative trait between carriers of the two alleles

exceeding 5 % (Cirulli and Goldstein 2010; Gibson 2011; Sebat et al. 2007), GWAS fundamentally relies on the presence of genetic correlation to survey the human genome in an efficient manner. By focusing on well-defined "tags" that are representative markers of the information content in the neighboring genomic regions, LD allows >80 % of the common variants in the human genome to be summarized by around one million SNPs. Discoveries of genotype–phenotype associations to date have thus been made with these tagging SNPs where the SNPs by themselves are not necessarily functional. The underlying causal variants that are biologically responsible for phenotype variation are seldom assayed directly and, in most situations, still unknown. While several reports have suggested that identifying the causal variants can increase the amount of heritability explained (Sanna et al. 2011; McCarthy and Hirschhorn 2008), it is increasingly clear that the common disease–common variant hypothesis is unlikely to fully explain the genetic etiology to diseases and traits.

The focus has since shifted to functional variants that are present at lower frequencies in the population, broadly defined as low frequency (MAF is between 1 and 5 %) or rare (MAF <1 %), although these are discussed together in this chapter. These variants are expected to contribute to common diseases by exerting larger effects on the phenotype, such that these variants contribute to explain a modest degree of phenotypic variance, despite their low frequencies in the population (see Table 1). For example, Tang and colleagues reported a variant rs17863783 with a risk allele frequency of 2.5 % in 5,284 healthy controls and an odds ratio of 0.55 for bladder cancer risk (Bosse et al. 2012), and a report by Nejentsev and colleagues that identified four rare variants with almost a twofold reduction in type 1 diabetes risk through re-sequencing the *IFIH1*gene that was initially implicated by GWAS (Nejentsev et al. 2009). The latter study demonstrates the importance of surveying across the whole allelic spectrum: from common variants with small or modest effects to low frequency or rare variants with moderate to large effects, in order to understand the genetic contributions to complex diseases and common traits.

Fine-Mapping of Causal Variants

Leveraging on the presence of LD has allowed GWAS to survey most of the genome by genotyping a smaller subset of well-chosen tag SNPs. However, the selection of these SNPs prioritizes their ability to summarize the information of their neighboring variants, rather than on the biological significance of the SNPs. GWAS thus identify indirect associations, where the SNPs discovered to be associated with the phenotype are not biologically meaningful by themselves but are simply correlated to the underlying (and often unknown) functional variants. Given that the real aim of a genetic association study is to identify the genomic unit (either the gene or the specific SNP within a gene) that causes a biological change to produce an impact on the phenotype, there is a need to follow up on the discoveries made by GWAS to localize these functional units. This process is known as *fine-mapping* the causal variants, which can either mean to identify the exact functional polymorphisms or to narrow the genomic region where the functional polymorphisms may reside.

There are two general approaches to the process of fine-mapping: (a) through targeted re-sequencing of a candidate region which, with sufficient sequencing coverage, is expected to locate most of the polymorphic positions in the region, and these can subsequently be tested for association with the phenotype, and (b) through in silico genotyping or genotype imputation with well-chosen reference haplotype panels obtained from either targeted or whole-genome sequencing of a set of population samples, which will infer the genotypes for the variants that are present on the haplotype panels for subsequent testing of association with the phenotype (Tang et al. 2012). The expectation in both approaches is that the functional polymorphism will present the strongest signal or be among the top signals. An example of the latter fine-mapping strategy was demonstrated by Jallows and colleagues, where the classic functional variant for sickle-cell anemia (rs334 at 5,248,232 bp on chromosome 11) was successfully localized by imputing 2,500 severe malaria cases and controls off a population-specific reference panel built from targeted re-sequencing of a 111 kb region surrounding the GWAS findings in 62 additional samples (Jallow et al. 2009).

Trans-Ethnic Fine-Mapping of Common Causal Variants

The principle of fine-mapping relies on segregating the causal variant(s) from other SNPs that are not functionally relevant with respect to the phenotype. For common variants, long stretches of LD paradoxically confound the process of fine-mapping. The process of discovering genomic regions that are associated with a phenotype has benefitted from the presence of LD. However, regions exhibiting strong LD mean that the causal variants are highly correlated with neighboring variants and thus present similarly strong evidence of phenotypic association that are virtually indistinguishable from the causal variants (see Fig. 1).

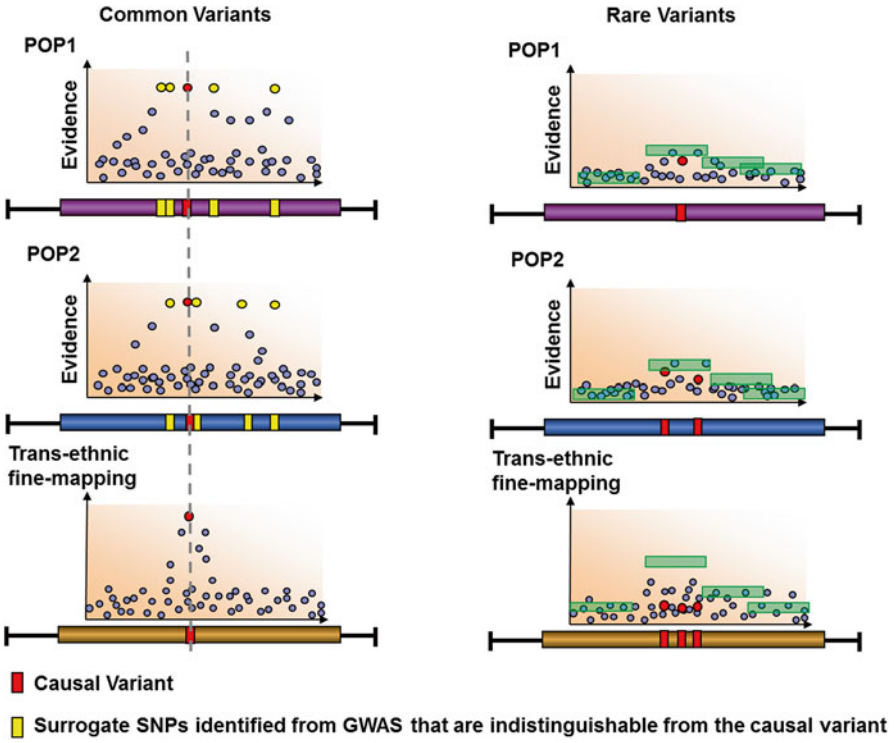

Fig. 1 Trans-ethnic fine-mapping of common and rare causal variants. The panels on the *left column* illustrate the trans-ethnic fine-mapping of common causal variants. The same causal variant (*red vertical bar* and *circle*) is present in two populations of different genetic ancestries (*top* and *middle panels*), but sits on two distinct haplotypes (represented by the purple and blue horizontal bars). Long stretches of high LD imply neighboring markers are near-perfect surrogates of the causal variants (*yellow vertical bars* and *circles*) and thus present a similar degree of evidence as the causal variant. Meta-analyzing the evidence from both populations (*third panel*) allows the causal variant to be distinguished as the SNP with the strongest evidence. The panels on the right column illustrate the corresponding situations in the trans-ethnic fine-mapping of rare causal variants. Due to the likely nature that rare causal variants are ancestry specific, meta-analyzing the evidence from individual SNPs across the two populations is unlikely to boost the statistical evidence if the same causal variants are not present across both populations. However, due to the sparsity of the association signals, a common approach is to aggregate the evidence across multiple SNPs in a contiguous region such as a gene exon to measure genetic burden, and meta-analyzing the region-based evidence across multiple populations can boost statistical power if the region is causally implicated across these populations

Due to different evolutionary and migration history, LD structure can vary significantly across populations, particularly between those from different ancestries (Teo et al. 2009a). Assuming that the causal variant is functional and shared across populations of different ancestries, there is the opportunity to leverage on varying patterns of genetic correlation between populations in order to localize the causal

Table 2 Comparisons between trans-ethnic fine-mapping of common and rare causal variants

Conditions for trans-ethnic fine-mapping	Common causal variants	Rare causal variants
1. Presence of a causal variant across populations from different genetic ancestries	Likely to be an older mutation, thus present and functional across populations from different genetic ancestries	Likely to be more recent, thus tend to be ancestry or population specific, where the same SNP may be causal in one population but monomorphic or not functional in other populations
2. Method of discovering and quantifying genetic association	Each SNP is typically the unit of analysis, and association testing measures the evidence of each SNP to be linked to the phenotype of interest	While SNP-based analyses are performed as with common variants, the typical unit of measurement aggregates the allele counts across multiple SNPs in a region to measure genetic burden, thus presenting a region-based evidence
3. Linkage disequilibrium (LD) between a causal variant and neighboring SNPs	Likely to be in LD with neighboring SNPs, and these SNPs present evidence of similar magnitude as the causal variant	Likely to be in weak or impractical level of LD with neighboring SNPs due to low frequency of the functional allele

variant. This can happen in two manners: (a) the functional allele at the causal SNP resides on several distinct haplotypes in different populations, and few SNPs will display consistent evidence of phenotypic association across multiple populations upon harmonizing the fine-mapping evidence from these populations, and (b) the functional allele at the causal SNP resides on one main haplotype that is present across different population, except the strength and extent of LD between the causal variant and the surrogate SNPs on this haplotype differ between populations, and harmonizing the fine-mapping evidence narrows the genomic region to the intersection of the different haplotype lengths (Tang et al. 2012) (Fig. 1).

Trans-Ethnic Fine-Mapping of Rare Causal Variants

There have been numerous reports of success in the use of trans-ethnic strategies to localize the causal variants from GWAS discoveries (Tang et al. 2012; Wu et al. 2013; Hughes and Sawalha 2011; Franceschini et al. 2012). Whether this process can be similarly extended to localize low frequency or rare causal variants remains to be seen. Here, we present an overview of the situations that facilitate the process of trans-ethnic fine-mapping of common causal variants and discuss the parallel situations for rare variants (Table 2).

Presence of a Causal Variant Across Populations of Different Ancestries

The fundamental concept of trans-ethnic analyses assumes that the same genetic unit, whether it is an SNP, a gene exon, or the entire gene itself, is biologically responsible for altering the expression of the phenotype across the different populations that are being jointly analyzed. For common causal variants, this assumption is likely to be valid given that these mutations tend to be older and would have occurred prior to the divergence of these different populations (Raychaudhuri 2011).

In contrast, rare SNPs are more likely to be recent mutations and thus ancestry or even population specific (Raychaudhuri 2011). This presents a significant challenge in attempts to pool the evidence of phenotypic association at a rare SNP, since the SNP may be polymorphic and functional in one population, but may be monomorphic in the remaining populations, and the joint analysis attenuates rather than strengthens the statistical evidence (Teo et al. 2009b).

The 1000 Genomes Project (Durbin et al. 2010) (1KGP, http://www.1000genomes.org) provided vital insights to the distribution of polymorphic SNPs across global populations. Through whole-genome sequencing of more than 2,500 individuals from at least 20 population groups around the world, the 1KGP presents an unbiased survey of genetic variation across diverse populations. One of the crucial findings that is relevant to determine the success of trans-ethnic association analyses is on the specificity of polymorphisms according to MAF. The 1KGP reported that common variants with MAF exceeding 10 % are shared across almost all the populations in Phase I of the project, whereas only 17 % of the rare variants are present in populations within the same ancestry group; and 53 % of the rare variants with MAF <0.5 % are population-specific (Abecasis et al. 2012). This finding suggests that, while trans-ethnic analyses of rare variants may be realistic for populations from the same ancestry, it is unlikely to be feasible to extend this to multiple populations from diverse ancestries.

Method of Discovering and Quantifying Genetic Associations

A GWAS typically analyzes each SNP independently for evidence of phenotypic association. The strength and direction of the association is similarly quantified at the SNP level, measuring the impact of each additional copy of the minor allele in altering phenotype. This relies on standard statistical procedures such as analysis of variance (ANOVAs) or regression analyses or univariate approaches such as chi-square tests or *t*-tests of averages. These approaches have proven to be reasonably successful in locating bona fide associations with common variants.

However, the statistical ability of these methods to successfully detect evidence of phenotypic association depends on observing sufficient number of samples that are carrying particular copies of the two alleles. These approaches are thus poorly powered to measure the evidence at rare variants, where the number of samples

carrying the risk allele may be very small. For example, Asimit and Zeggini illustrated, through a series of simulations, that as the causal allele frequency decreases from 5 % to 1 % to 0.1 %, the sample size required to attain a power of 80 % to detect an allelic odds ratio of 2 at the accepted genome-wide significance level of $P = 5 \times 10^{-8}$ increases from 2,500 to 12,000 to 117,000 (Asimit and Zeggini 2010). As a result, analyses of rare variants for phenotype association typically aggregate the cumulative impact of multiple SNPs located in a contiguous genomic region, for example, by pooling the number of copies of rare alleles within a phenotype stratum.

As described in Chaps. 12 and 13, the underlying assumption for such tests of genetic burden is that the set of rare variants within a region collectively influence the disease susceptibility, and the statistical evidence is measured according to whether the rare alleles tend to be more specific to subjects in a phenotype classification. However, methods such as the cohort allelic sum test (CAST) (Morgenthaler and Thilly 2007), the weighted sum test (WST) (Madsen and Browning 2009), and the collapsing regression method (Morris and Zeggini 2010) tend to ignore the direction of the effects of the rare alleles, and these tend to lower the power of the aggregated allele counts to correlate with phenotype expression, since rare alleles from different causal variants may be deleterious or beneficial. The sequence kernel association test (SKAT) (Wu et al. 2011) properly accommodates for the direction of the effects of rare alleles and has been shown to possess higher statistical power than most of the collapsing approaches.

For a genomic region that genuinely harbors causal variants across multiple populations, pooling the evidence from individual SNPs is unlikely to improve the strength of the statistical association, since the architecture of rare variants suggests that different rare causal variants in the same region are likely to be present across the different populations. However, given that the unit of analysis for rare variants typically interrogates the entire genomic region, trans-ethnic analyses can boost the ability to locate these associated regions by aggregating the statistical evidence of phenotypic association (Fig. 1). Identifying the rare causal variants in the emerging genomic region will require interrogating which SNPs contribute to the primary association signal within each population and by assessing the annotations—a process of fine-mapping that similarly is unlikely to benefit from trans-ethnic strategies.

Linkage Disequilibrium Between a Causal Variant and Neighboring SNPs

Causal variants with minor allele frequencies that are in excess of 5 % are often in useful levels of LD with neighboring SNPs, and they tend to present similar evidence of phenotypic association as the causal variants. GWAS has relied on such long stretches of high LD in identifying the markers that correlate with phenotype expression. Trans-ethnic fine-mapping of these common causal variants is thus necessary to distinguish the surrogate SNPs from the causal variants.

The situation is notably different for rare causal variants, as these tend to be in weak levels of LD with surrounding markers due to their low minor allele counts.

From this perspective, there is no need to depend on trans-ethnic fine-mapping to localize rare causal variants, and often the causal variants can be identified by interrogating the evidence within a population, as suggested by Zhu and colleagues who developed the "preferential LD" approach (Zhu et al. 2012). They suggested that weak levels of LD are present between a rare causal variant and a small set of markers that may be used to locate the genomic region, but such LD is still considerably stronger than those present between the causal variants and other surrounding SNPs. Based on this assumption, the "preferential LD" approach searches for rare variants with unexpectedly higher LD with the discovery variant, which are subsequently more likely candidates as the causal variants. When applied to a range of diseases, this approach successfully confirmed two well-known rare causal variants for Crohn's disease in the *NOD2* gene (Wang et al. 2010), two non-synonymous *ITPA* variants (rs1127354 and rs7270101) that cause ribavirin-induced hemolytic anemia (Fellay et al. 2010), and rare variants in *UGT1A6* gene for bladder cancer (Tang et al. 2012).

Conclusion

Trans-ethnic fine-mapping has seen remarkable success in disentangling the conundrum of long stretches of high LD to either locate common causal variants or at least narrow the genomic region where these functional variants at MAF >5 % can be found. However, the genetic architecture of rare variants is considerably different from that of common variants without the complication introduced by LD. For common causal variants, it appears existing methods are more than adequate to locate and validate an association signal, and the challenge lies in identifying the genuine causal variants from perfect surrogates. For rare variants, the greater challenge appears to lie in locating and validating an associated genomic region, rather than in fine-mapping the causal variants. Indeed, once a genomic region has been systematically confirmed to be associated with a phenotype, fine-mapping the causal variants is unlikely to require more than the careful interrogation of which rare SNPs contributed to the association signal and their functional annotations within one study cohort.

References

Abecasis GR et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Bosse M et al (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genet 8(11):e1003100

Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415–425

Durbin RM et al (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
Fellay J et al (2010) ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. Nature 464(7287):405–408
Franceschini N et al (2012) Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am J Hum Genet 91(4):744–753
Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145
Hughes T, Sawalha AH (2011) The role of epigenetic variation in the pathogenesis of systemic lupus erythematosus. Arthritis Res Ther 13(5):245
Jallow M et al (2009) Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 41(6):657–665
Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384
McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17(R2):R156–R165
Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615(1–2):28–56
Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188–193
Nejentsev S et al (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925):387–389
Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. Cell 147(1):57–69
Sanna S et al (2011) Fine mapping of five Loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet 7(7):e1002198
Sebat J et al (2007) Strong association of de novo copy number mutations with autism. Science 316(5823):445–449
Tang W et al (2012) Mapping of the UGT1A locus identifies an uncommon coding variant that affects mRNA expression and protects from bladder cancer. Hum Mol Genet 21(8):1918–1930
Teo YY et al (2009a) Genome-wide comparisons of variation in linkage disequilibrium. Genome Res 19(10):1849–1860
Teo YY et al (2009b) Power consequences of linkage disequilibrium variation between populations. Genet Epidemiol 33(2):128–135
Visscher PM (2008) Sizing up human height variation. Nat Genet 40(5):489–490
Wang K et al (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet 86(5):730–742
Wang Y, Chen YH, Yang Q (2012) Joint rare variant association test of the average and individual effects for sequencing studies. PLoS One 7(3):e32485
Wu MC et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93
Wu Y et al (2013) Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genet 9(3):e1003379
Zhu Q et al (2012) Prioritizing genetic variants for causality on the basis of preferential linkage disequilibrium. Am J Hum Genet 91(3):422–434

Erratum to: Functional Annotation of Rare Genetic Variants

Graham R. S. Ritchie and Paul Flicek

Erratum to:
Chapter 5 in: E. Zeggini, A. Morris (eds.),
***Assessing Rare Variation in Complex Traits*,**
https://doi.org/10.1007/978-1-4939-2824-8_5

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-1-4939-2824-8_5

E. Zeggini, A. Morris (eds.), *Assessing Rare Variation in Complex Traits*, https://doi.org/10.1007/978-1-4939-2824-8_19