Chapter 9
Analytical Approaches for Exome Sequence Data

Andrew Collins

Abstract Sequencing the 1% of the genome coding for proteins (the exome) offers a powerful and often cost-effective route to identifying genetic mutations underlying Mendelian disease. It is possible that exome sequencing in a relatively small number of individuals showing 'extreme' phenotypes or more familial subtypes of complex disease may also be productive. Larger-scale exome and whole genome sequencing studies offer the potential to interrogate the cumulative impact of the numerous rare variants presumed to underlie a substantial proportion of complex disease susceptibility. Exome and, particularly, whole genome sequencing studies yield enormous amounts of data and pose many analytical challenges. Aside from issues concerning the production of high-quality sequence reads and the management and manipulation of huge databases, a major concern, in the early stages of analysis, is the reliable alignment of the short sequence reads against a reference genome. A wide range of algorithms and software tools for alignment have been developed and implemented for this most critical step in every analysis 'pipeline'. A similarly rich set of platforms and analytical tools are available to facilitate the reliable calling of DNA variants. Given the excellent resources now available, the production of a well-characterised database cataloguing novel and known variants in an individual exome is achievable. However, the difficulty of teasing out causal variants from the vast amount of neutral or irrelevant variation presents the greatest challenge. I review here the techniques and tools that have been developed and applied for the analysis of exome data. Exome mapping of genes involved in Mendelian disease has met with considerable success thus far, while applications to complex traits look promising given analysis of sufficiently large numbers of case and control exomes.

Keywords Complex disease · Exome sequencing · Mendelian disease · Sequence alignment · Variant annotation

A. Collins (✉)
Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, University of Southampton, Southampton, UK
e-mail: a.r.collins@soton.ac.uk

Y. Yao (ed.), *Applied Computational Genomics*, Translational Bioinformatics 13,
https://doi.org/10.1007/978-981-13-1071-3_9

9.1 Introduction

Thousands of genetic variants for both Mendelian diseases and complex traits have been identified as causal or associated with disease phenotypes in recent years. These have usually been identified through linkage mapping, in the case of Mendelian disease, and candidate gene studies or genome-wide association studies (GWAS), in the case of complex traits. For complex diseases the majority of the implicated single nucleotide polymorphism (SNP) variants are associated indirectly with disease, usually to a genomic region. Because these regions can be large and/or inter-genic, GWAS associations may or may not indicate whether a specific gene is compromised and involved in disease. In contrast, sequencing enables the identification of all variants in a genome or genomic region such that an individual variant can, in favourable circumstances, be firmly identified as causal. For this reason exome sequencing and whole genome sequencing are already revolutionising the way genetic studies are undertaken.

Recent years have seen dramatic changes in the development and application of DNA sequencing technology. The traditional Sanger sequencing method employing capillary electrophoresis remains the 'gold standard' in terms of the length of the reads and the accuracy of the sequence (Harismendy et al. 2009). However, 'next-generation sequencing' (NGS) methods generate 3 or 4 orders of magnitude more sequence at greatly reduced cost compared to the Sanger approach. These methods sequence DNA molecules spatially separated in flow cell and attached to a solid surface. The process employs optical imaging to record the sequential addition of nucleotides in the sequencing reaction. This enables millions of sequencing reactions to take place in parallel. The first massively parallel NGS platform was launched in 2005 (Majewski et al. 2011). NGS radically overcomes the problem of limited scalability of the Sanger approach (Reis-Filho 2009; Lander 2011) and is capable of generating hundreds of mega- to giga-base pairs (bp) of nucleotide sequence in a single run. Millions of overlapping sequence reads are then aligned and compared to a reference genome to identify differences (polymorphisms). Targeted sequencing of genomic regions of particular interest, of which the most important is undoubtedly the entire exome (the protein-coding exons of all genes), has benefits with respect to reduced cost, data management and increased sequence coverage (for a given quantity of DNA). Exome sequencing typically involves sequencing the ends of fragments from the sheared sample DNA – either one end (single-end sequencing) or both ends (paired-end sequencing) of the fragments. The sequence read lengths are typically in the range of 35–150 bp for Illumina platforms (http://www.illumina.com/applications/sequencing/targeted_resequencing.ilmn) and ~400 base pairs for the Roche 454 sequencer (http://www.roche.com/products/product-list.htm?type=researchers&id=4). The exome comprises only ~1% of the genome (~30 Mb), so an average 'depth' of coverage of the exome of 75 can be

achieved with 3 Gbp of sequence, whereas 90 Gbp would be required for 30-fold-depth coverage of the whole genome (Majewski et al. 2011; Bainbridge et al. 2010).

The exome is the best understood component of the genome for relating sequence to function and, similarly, to directly link genetic variants with disease causality (Kumar et al. 2011). For Mendelian disorders, exome sequencing offers a powerful route to identifying the underlying allelic variants since the majority of this class of disease genes are known to disrupt protein-coding sequences. Kryukov et al. (2007) have shown that most rare non-synonymous (missense) alleles are likely to be deleterious, unlike the majority of noncoding sequences. The exome is therefore particularly enriched for variants underlying Mendelian traits. There is also increasing evidence that exome sequencing offers a route to understanding complex disease. For example, it has been shown that rare variants are over-represented in genes already identified (usually by GWAS) as containing common variants involved in complex disease. Johansen et al. (2010) determined a significant burden ('mutation skew') of 154 rare missense or nonsense variants in 438 individuals with hypertriglyceridemia, compared to a significantly lower burden in controls, within four genes known to contain common variants for this condition. Support for the observation of rarer alleles with potentially higher disease penetrance residing within genes implicated by GWAS comes from the study by Rivas et al. (2011). Working on the inflammatory bowel disease (IBD) phenotypes, the authors identified novel rare variants which contribute a greater component to the population risk variance than the known common IBD variants in the *CARD9*, *NOD2*, *CUL2* and *IL18RAP* genes. Lehne et al. (2011) questioned whether missing the regulatory elements that may impact disease phenotype, but are situated outside the exome sequence regions, would reduce the value of applying exome sequencing to complex disease. For most of complex diseases examined, the authors found that most of the association signal from 'suggestive' common variants was found within the coding regions rather than introns. Although they did not consider rare variation directly, the work supports exome sequencing as a strategy to search for genetic variation associated with complex disease.

Despite its evident advantages and early successes, exome sequencing has a number of disadvantages and problems, aside from the obvious lack of information from the bulk of the noncoding genome. Exon capture requires the use of complementary nucleic acid 'baits' to trawl sequence reads from specific exons. Since these are 'small' targets, this can result in uneven coverage of exonic regions, and the baits themselves are only as complete as the information derived from gene annotation and other reference databases. There is also a degree of low-depth hybridisation away from the targets in non-exonic regions although the overlap of sequence reads extending a short distance either side of the bait probes provides some information on adjacent regions. There is a trend towards increasing the coverage of exonic and adjacent regions in the newer products. Perhaps more important than concerns about coverage are a wide range of data analytical considerations, reviewed here.

9.2 Strategies for Exome Projects

The strategy chosen for an exome sequencing study depends on the known, expected or hypothesised genetic mode of inheritance. The costs and analytical challenges of sequencing hundreds of exomes to pursue the complete spectrum of rare variation underlying complex disease are likely to be prohibitive for all but large consortia for the foreseeable future. At the other end of the spectrum, highly successful studies focussed on a small number of related individuals have been achieved for Mendelian diseases. Between these two extremes is perhaps the most intriguing prospect: sequencing a small number of affected relatives showing relatively strong familial patterns for a complex trait and/or focussing on a distinct disease subtype or individuals showing an 'extreme' phenotype of a common disease might identify important rare variation. Success depends on the existence of forms of complex disease closer to the Mendelian end of the disease spectrum, and strategies include focus on individuals with particularly severe forms of a disease and/or markedly early onset. For complex diseases there remains a substantial degree of uncertainty about how best to design such studies, but I consider here some of the findings to date.

9.2.1 Mendelian Disorders

Fewer than half of the allelic variants underlying monogenic diseases showing a Mendelian pattern of inheritance have been identified. The difficulty with finding many of these genes arises from the rarity of affected cases or case families, the existence of similar phenotypes determined by independent mutations (locus heterogeneity) and the reduced reproductive fitness limiting the further analysis of key pedigrees. Many of these more difficult diseases arise as de novo mutations and are not therefore amenable to linkage analysis. However, exome sequencing offers a route to progress and initial applications, focussed on a number of Mendelian disorders, have identified high-penetrance genes through sequencing a very small number of affected family members. Ng et al. (2009) were the first to demonstrate the utility of exome sequencing to identify Mendelian disease variants. As proof of principle, the authors sequenced the exomes of four unrelated cases with Freeman-Sheldon syndrome, a disease for which the causal variant was known, and eight control samples. The authors filtered out common and presumed unimportant variation identified in HapMap and dbSNP and demonstrated that disease variants could be mapped solely by exome sequencing of a few cases. The gene for Miller syndrome (Ng et al. 2010a) was the first example of a gene found for a disease of unknown cause. The DHODH gene was mapped using four affected cases in three independent pedigrees, data filtered against public SNP variant databases, and verified by Sanger sequencing in three additional Miller families. To maximise the chance of identifying the gene, the authors considered a dominant model with at

least one novel non-synonymous SNP, splice variant or coding indel. Their recessive model required genes with at least two novel variants which were either in the same position (homozygous) or in different positions (as a possible compound heterozygote but conditional on, unknown, phase). The success of this enterprise depended to a large extent on the choice of disease. Miller syndrome is a very rare Mendelian disease, and so causal variants were unlikely to be present in reference databases or control exomes. Mapping a rare recessive gene is easier than a dominant gene because fewer genes within the affected individual's exome will have two novel or rare non-synonymous variants. The lack of genetic heterogeneity in the sample of individuals studied was also advantageous, and the authors emphasise the importance of ethnic uniformity in the ancestry of affected cases (Europeans in this case) reducing the likelihood of genetic heterogeneity.

Strategies that might accelerate the mapping of Mendelian disorders in the future include, for recessive models, identifying genes within shared tracts of homozygosity to reduce the pool of potential candidate variants for further consideration. Krawitz et al. (2010) introduced identify-by-descent filtering to map the recessive gene for hyperphosphatasia mental retardation syndrome (HPMRS or Mabry syndrome) in a family with three affected siblings. They developed a hidden Markov model to identify regions with shared identical, maternal and paternal haplotypes but not necessarily derived from a common ancestor. They were then able to identify whether each sibling had the same (identity by descent = 2) homozygous or heterozygous genotype. This process reduced the pool of candidate genes with mutations in all three sibs from 14 to 2 and led to the identification of the PIGV gene as causal.

9.2.2 De Novo Variants

F or 'sporadic' disease sequencing of unaffected parents may facilitate rapid identification of important de novo mutations involved in disease. Girard et al. (2011) sequenced exomes and parents of 14 schizophrenia probands with no previous family history and identified 15 de novo mutations in eight probands. This is a higher de novo mutational burden than the 'background' mutation rate as indicated by the 1000 Genomes Project. Four of the 15 mutations were predicted to lead to a premature stop codon in genes hypothesised to have a role in the disease.

9.2.3 Cancer Germline and Tumour Studies

A route to further understand the genetic basis of cancer is offered by the exome sequencing in both germline and tumour DNA from the same patient and searching (by subtraction of the germline variants) for novel somatic mutations. An early success for this approach is described by Tiacci et al. (2011) who exome-sequenced

germline and tumour DNA from an index patient with hairy-cell leukaemia (HCL). The findings included a somatic heterozygous mutation in the BRAF gene which was known to produce an oncogenic protein. Remarkably, the same variant was identified by Sanger sequencing as present in all 47 additional HCL patients they were screening but in none of their 195 patients with other forms of peripheral B-cell lymphoma or leukaemia. The power of this approach to identify recurrent somatic mutations driving further downstream somatic changes was clearly demonstrated. The findings also support BRAF mutation screening as a diagnostic tool to distinguish HCL from other B-cell lymphomas and identify HCL as a clinically distinct entity from other 'HCL-like' disorders.

9.2.4 Rare Variants in Families: Extreme Phenotypes

Feng et al. (2011) consider strategies for mapping rare variants in complex disease in the context of family data. The authors recognise the critical issues which reduce power, namely, locus heterogeneity (McClellan and King 2010), allelic heterogeneity (2000 pathogenic mutations have been reported in BRCA2), problem of phenocopies (affected individuals in a family that do not share the predisposing mutations) and apparent oligogenic patterns of inheritance due to segregation of many common moderate-risk loci. Nevertheless, Cirulli and Goldstein (2010) argue that family-based designs, particularly for families showing phenotypes from the extremes of a trait distribution, are most likely to achieve success for complex traits until the costs of sequencing reduce sufficiently to favour very large case-control designs. Simulations support a two-stage design with sequencing of two affected individuals per pedigree that are not too closely related to generate an excessive number of false-positive genes or too distantly related to increase the risk of including a phenocopy in the comparison.

9.2.5 Rare Variants in Large Cohorts: Mutational Load

Cooper and Shendure (2011) consider the interpretive challenge of the 'multiple hypothesis testing' problem presented by the enormous number of variants identified in genome sequences and the abundance of false discoveries. They argue that experimental or computational approaches to assess variant function can provide estimates of the prior probability that a given variant is phenotypically important, thereby boosting discovery power. Such empowering classifiers include SIFT scores that use 'evolution as the best measure of deleteriousness', the observation that sequences not removed by natural selection are likely to be important. Application of a comprehensive range of functional and predictive tools is likely to be required for complete characterisation of important low-frequency variation identified in large cohorts of patients with common forms of disease. Evolutionary models

predict that rare deleterious mutations spread across a large number of genes may have a cumulative effect (mutational load) to increase susceptibility to complex disease.

In this scenario a given mutation may be present in only a few individuals and have a negligible effect on trait variation, but, in combination with many similar variants, the burden of mutation may underlie causality (Howrigan et al. 2011). Pooled association tests and collapsing methods (Price et al. 2010; Dering et al. 2011) provide routes to testing mutational burden in large-scale genetic studies.

9.3 Exome Data

Data from a sequencer are typically presented in FASTQ format in which there are four lines per read comprising sequence identification labels, raw sequence and quality scores for each of the bases in the sequence (http://en.wikipedia.org/wiki/FASTQ_format). The quality score represents, as a single ASCII character, the probability (p) that the base call it refers to is incorrect. The Sanger version of the Phred quality score is $Q\text{sanger} = -10\log 10p$. Two such FASTQ files are generated for paired-end sequencing with sequential entries corresponding to the sequenced ends of each DNA fragment. Li et al. (2009a) describe the now standard 'sequence alignment/map' (SAM) format for storing short read alignments and mapping coordinates against a reference sequence. A software package (SAMtools) is used for processing such files and has options for positional sorting, indexing, format conversion and calling and viewing variants. The standardised format allows for efficient capture of read and alignment information by defining codes that characterise aligned sequences and identified variations from the reference sequence. These include, for example, codes to represent matches and mismatches, insertions, deletions and sequences with 'soft' and 'hard' clipping to represent non-matched sequences which are either present or missing from the alignment. Their CIGAR format provides a compact way of storing good alignments and also representing bases misaligned to the reference genome. The SAM format has a binary equivalent file (BAM file) which improves processing performance by supporting more rapid retrieval of aligned sequences in specific genomic regions.

9.3.1 Sequence Alignment

Accurate alignment of short read sequences against a reference genome is the most critical step towards cataloguing the polymorphisms represented in a sample. The process requires a reliable reference genome with known sequence and millions of short reads from the sample genome. Many algorithms have been developed to align sequence reads against the reference genome. Li and Homer (2010) and Ruffalo et al. (2011) survey the range of sequence alignment packages. Short read alignment

packages include Bowtie (Langmead et al. 2009), BWA (Li and Durbin 2009), MAQ (Li et al. 2008), mrsFAST (Alkan et al. 2009), Novoalign (http://www.novocraft.com/main/index.php), SHRiMP (Rumble et al. 2009) and SOAPv2 (Li et al. 2009b). Of these, BWA is one of the most frequently used aligners. It exploits indexing built using the Burrows-Wheeler transformation (Burrows and Wheeler 1994) which enables fast searching and generates a quality score that can be used to reject poorly supported alignments. Ruffalo et al. undertook a simulation-based comparison and noted that the different approaches trade off speed and accuracy to optimise detection of different variant classes. Some algorithms were more efficient at different stages in the alignment process. For example, BWA and SOAP were found to align genomes quickly but required significant time to index the genome, whereas Novoalign required less time for indexing time but performance showed greater dependence on the number of reads. Novoalign offers high sensitivity and specificity with respect to accuracy of alignments and uses information on base qualities at all stages in the alignment (Li and Homer 2010) although this impacts on speed of the alignment. However, higher performance can be achieved by running the message passing interface (MPI) version on a computer cluster and exploiting multithreading.

9.3.2 Variant Calling

Given an aligned set of reads, it is essential to identify and 'mark' duplicate reads so that they do not influence variant calling. Tools to achieve this include PICARD (http://sourceforge.net/apps/mediawiki/picard/index.php?title=Main_Page) and SEAL (Pireddu et al. 2011), an alignment tool which combines BWA with the detection and removal of duplicate reads. Duplicates are likely to be PCR artefacts from the library preparation stage or optical duplicates from the sequencer. Duplicates are most simply defined as those reads that map to exactly the same locations. Other quality control preprocessing includes base quality score recalibration (applied to a BAM file) (http://www.broadinstitute.org/gsa/wiki/index.php/Base_quality_score_recalibration). This procedure recalibrates the scores to more accurately reflect the probability of mismatching the reference genome. The Genome Analysis Tool Kit (GATK) provides quality score recalibration which targets not only overall base quality inaccuracy but identifies higher quality subsets of bases by accounting for decline in base quality known to occur towards the ends of sequence reads.

Tools such as GATK and SAMtools are capable of identifying short indels in exome data, but accurate characterisation of indels in exome data is challenging. For example, short indels tend to occur in the vicinity of tandem repeats, but accurate alignment in these regions is difficult. Furthermore, where an indel is present, it may create local misalignments against the reference sequence which can generate false SNP calls. Therefore, local realignment around indels is required to minimise the number of mismatching bases (http://www.broadinstitute.org/gsa/gatkdocs/

release/org_broadinstitute_sting_gatk_walkers_indels_IndelRealigner.html). Local realignment aims to resolve regions with misalignments caused by indels into clean reads, prior to applying tools to identify the variant content of the exome. Calling variants while using the information from more than one exome simultaneously increases the quality of variant calls. GATK's UnifiedGenotyper module employs a Bayesian genotype likelihood model to derive the most likely genotypes as applied to multiple samples simultaneously. The program also generates a posterior probability for a segregating variant allele as well as genotype at each locus.

VarScan (Koboldt et al. 2009, http://varscan.sourceforge.net/) is designed for identifying SNPs and indels in NGS data and is particularly suited to filtering in tumour-normal (tumour-germline) paired samples. Given such paired data, VarScan tests the somatic status of each variant and classifies them as germline, somatic or loss of heterozygosity by comparing the read counts between samples. VarScan uses the 'pileup' files of variant output from the SAMtools program from the germline and tumour DNAs simultaneously. Variant positions shared between both files meeting the minimum read depth coverage are compared and variants classified accordingly. Filtering against a germline sample of variants has obvious benefits in terms of reducing variant volume and complexity in the expectation of identifying recurrent 'driver' mutations that underlie the disease.

9.3.3 Filtering and Identifying Disease Susceptibility Genes

Sets of variant calls from an exome sequence include a large number of false positives. Suggested quality control filters, as implemented, for example, in the GATK program, include removal of variants at sites with low mapping quality scores and removal of apparent heterozygotes in which one allele is supported by less than 30% of sequence reads, variants not supported by reads mapping to both strands (strand bias). A significant difference of NGS from traditional Sanger sequencing is that the error rates for the called bases are markedly higher. This underlies the importance of obtaining high coverage 'depth' (the number of independent sequence reads aligned at one location). For this reason the removal of variants supported by only low read depth (e.g. 10 reads or less) is an important QC step.

Even given robust quality control throughout the analytical pipeline, the resulting file of SNPs and indels will contain many thousands of variants. The most pressing issue is how to determine the relationship (if any) of specific variants identified to the disease phenotype(s). Annotation of variants and filtering to identify and remove 'unimportant' variation can be achieved by tools such as Annovar (Wang et al. 2010) which enables local download of all variants in genomic databases (1000 genomes, dbSNP, etc.) and provides tools for flexible filtering to remove common variation unlikely to be involved in disease. This is not straightforward since a number of these databases, such as recent versions of dbSNP, contain known rare and disease-causing variants which might be relevant to the phenotype under investigation. However, reduction in complexity of voluminous data at this stage is

essential since an individual exome is likely to carry ~10,000 amino acid altering SNPs (Ng et al. 2010b). A (probably small) proportion of these are likely to negatively impact health, but the majority simply contribute to the large diversity of proteins and have little or no deleterious impact. For Mendelian diseases it is likely that the rare high-penetrance variants involved are private to affected individuals fully supporting the value of filtering out the common variation represented in genomic databases. Efficient filtering reduces the pool of potential disease influencing variants enabling cost-effective follow-up of a much smaller number of genes and/or variants. Studies of Mendelian disorders assume a single highly penetrant coding mutation is sufficient to cause disease and that mutation is very rare and probably restricted to affected individuals. The volume of variation can be much reduced by only considering variants that change the protein sequence (non-synonymous), coding indels and splice acceptor and donor site changes. However, for non-Mendelian traits, it is known, from GWAS studies, that common intronic, regulatory and synonymous variation has an impact on disease, and so filtering is likely to lose information. Even after filtration against common variant databases, and after considering only protein-changing variants, the high number of variants in an individual exome is large enough to challenge further progress. In silico approaches computationally evaluate potential disease severity of variants by making multispecies comparisons and using models of molecular evolution (Kumar et al. 2011). The degree of conservation at individual positions and databases of permitted substitutions indicates the potential impact of a given change. It is known that disease-associated SNPs are over-represented at locations in the genome that have changed to only a limited degree over evolutionary time. Variants at locations conserved throughout vertebrates are more likely to be involved in Mendelian disease, and the same has been found to be true for the locations of somatic variation in cancers. Intense purifying selection against damaging variants at these locations is likely to occur through a reduction in reproductive fitness. For this reason molecular evolutionary predictions are considered less useful for complex disease where later onset has limited impact on fecundity. However, there is a spectrum of genetic disease from single-gene Mendelian disorders to complex traits. Therefore, in silico prediction may be valuable for more 'extreme' forms of complex disease (e.g. early onset, more severe disease subtypes, familial cases). Ranking variants by their predicted or known effect on protein function and their degree of conservation using tools, such as SIFT (Kumar et al. 2009), PolyPhen2 (Adzhubei et al. 2010), LRT (Chun and Fay 2009) and MutationTaster (Schwartz et al. 2010), and composite databases of functional predictions such as dbNSFP (Liu et al. 2011) is an important further step towards reducing data depth and complexity. The various algorithms output scores which quantify the extent to which a non-synonymous variant is likely to be deleterious. Such an approach has already been used with success to prioritise novel variants for follow-up in Mendelian disease studies (Ng et al. 2010a). SIFT ('Sorting Tolerant From Intolerant', http://sift.bii.a-star.edu.sg/) predicts the effect on protein function of single amino acid changes. The SIFT algorithm works by searching for similar sequences that are likely to have matching functions, generates

an alignment of those sequences and computes probabilities for all possible substitutions from the alignment. Those with $p < 0.05$ are classified as deleterious mutations or, otherwise, tolerated. PhyloP (Pollard et al. 2010) similarly provides a conservation score highlighting locations that are conserved from invertebrates to humans in which substitutions are highly likely to disrupt critical protein function. PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/) also predicts the impact of an amino acid substitution on protein structure and function. The algorithm uses sequence and structural features to evaluate the impact of amino acid replacements within a multiple sequence alignment of homologous proteins, the extent of modification of the resultant protein and whether the substituted allele originated at a particularly mutable site. The alignment process uses the set of homologous sequences and employs clustering to construct and refine their multiple alignment. The functional significance of a substitution is predicted from the set of features by a naive Bayes classifier (Adzhubei et al. 2010). Chun and Fay (2009) develop a likelihood ratio test (LRT, http://www.genetics.wustl.edu/jflab/lrt_query.html) which compares the null model of neutral codon evolution to the alternative model that the codon has evolved under negative selection. Deleterious mutations are considered to be the non-synonymous SNPs that significantly disrupt the constrained codons defined by the LRT. The LRT generates a *p*-value for the likelihood ratio test of codon constraint. The test is developed from data for 32 vertebrate species. Chun and Fay (2009) found, however, a disturbingly low degree of overlap between predictions made by the LRT, SIFT and PolyPhen with 76% of predictions unique to one of the three methods and only 5% of predictions made by all three. With this in mind Liu et al. (2011) argue that, because the various alternative algorithms have their own strengths and weaknesses, it is useful to construct a consensus prediction. This is presented in their dbNSFP database (http://sites.google.com/site/jpopgen/dbNSFP) which contains functional predictions from multiple algorithms compiling predicted scores for non-synonymous variants from SIFT, PolyPhen2, LRT, MutationTaster and PhyloP.

9.3.4 Collapsing Methods for Rare Variants in Large Samples

Rarer variants are likely to be enriched for alleles with functional disease impact and may show larger effect sizes than common alleles as a consequence of purifying selection. However, the penetrance of most of these variants is likely to be comparatively low (Bodmer and Bonilla 2008). Therefore, for most complex disease phenotypes, the cumulative impact of many rare variants is likely to contribute significantly to the disease phenotype. However, the power to detect such alleles is low due the relatively low penetrance, the small number of copies of a given variant present and the need for stringent correction for the number of variants tested. For this reason analytical approaches for large samples have been developed that test for the combined effects of a set of rare variants, thereby greatly reducing the number of

statistical tests while maximising power. Such a 'collapsing' approach requires prior specification of the set of variants to be combined to make the test. Li and Leal (2008) point out that misclassification resulting from the collapsing of nonfunctional variants with functional sites adversely affects the power of the test. Misclassification can arise when non-causal variants are included and when functional variants are excluded because they either have not been sequenced or have incorrectly been classified as nonfunctional by bioinformatics tools. In contrast multiple-marker methods which test several sites for their influence on phenotype simultaneously are more robust to misclassification, but potentially less powerful than collapsing methods. Li and Leal's combined multivariate and collapsing (CMC) method aims to maximise power while being robust to misclassification. This and related tests are reviewed by Dering et al. (2011). The collapsing method defines an indicator variable X for the jth case individual to define whether or not that subject carries any rare variant in the target of interest (e.g. a gene) such that $X\mathrm{j} = 1$ when a rare variant is present and 0 when absent with Y_{j} similarly defined for controls. The test made is for association of multiple rare variants in which the proportion of rare variants in cases and controls differ. This is a fixed allele-frequency threshold approach for which power was investigated by Price et al. (2010). The authors examined different thresholds at which to define a variant as 'rare' (their T1 and T5 models representing 1 and 5% allele-frequency thresholds, respectively). They also describe a version of the test which weights (under the null hypothesis of no association) the contribution of each SNP by the inverse square root of the expected variance, based on allele frequencies computed from controls. This approach gives much higher weights to very rare variants. Price et al. propose a variable threshold approach which assumes an unknown threshold T for which variants with a MAF below the threshold are more likely to be functionally important than those above. The authors compute the maximum test statistic over a wide range of values of T to obtain the maximum of the threshold specific test statistics. The p-values are determined (as in all collapsing methods) by permutation tests.

An important addition to the range of collapsing approaches incorporates predicted functional information that improves the statistical test. Price et al. (2010) incorporated PolyPhen2 probabilistic scores for neutral and deleterious amino acid changes as weights in the regression. In their simulation study, setting the significance level to $p = 0.05$, power was higher at 60 and 69% for the variable threshold and variable threshold with PolyPhen scores models, respectively, compared to 55, 50 and 54% for the T1, T5 and weighted threshold models, respectively.

Luo et al. (2011) point out some of the limitations of collapsing methods, noting that variants at different genome locations may have different effect sizes which are unlikely to be determined only by their frequencies and collapsing without assigning weights that are functions of variant frequencies cannot fully exploit information of genetic effect sizes; multiple rare variants may be correlated, so grouping them needs to take this into account. They develop functional principal component analysis (FPCA)-based statistics for which they determine higher power to detect association with rare variants and enhanced ability to filter out sequence errors.

9.3.5 Copy Number Variant (CNV) and Loss of Heterozygosity Analysis

Test for structural variation has been typically undertaken using array comparative genome hybridisation (CGH) which tests up to one million probes and can detect variants in the size range of 10–25 kilobases. But much higher resolution can be achieved from sequence data, and Yoon et al. (2009) develop methods for detecting CNVs in whole genome sequences. However, similar application to exome sequence data presents difficulty because the read sequence distribution is not random or unbiased and the read depths do not follow a normal distribution from which deviations suggest the presence of a copy number variant. However, if the biases are controlled, exome sequencing data present the opportunity to detect structural variants at much higher resolution and extend the utility of the data beyond the identification of single nucleotide variants and small indels. The problems presented by the discrete nature of the exome read distribution are considered by Sathirapongsasuti et al. (2011) who describe a method to detect copy number variations (CNVs) and loss of heterozygosity (LOH) in exome data. The approach uses normalised depth ratios in paired samples (such as tumour/germline) that have been processed in a similar way, including library preparation, and share similar average depth of coverage. This approach was shown to identify CNVs as small as 120 pb representing single exons with higher than average coverage. The read depth data can be more flexibly used in non-matched exome samples, for example, by using data from a pool of control exomes to serve as, effectively, a matched control sample (since the average copy number is likely to be two given a sufficiently large number of control exomes).

9.3.6 Strategies for Efficient Analysis and Data Management

The alignment of short sequence reads has been regarded as a major bottleneck in the analysis of NGS data (Li and Homer 2010). However, improving the algorithms and the development of tools which exploit distributed processors has reduced this bottleneck, at least for exome sequence. Important developments include platforms which automate pipelines and provide integration of bioinformatics tools to facilitate exome analysis. An example is Galaxy (Goecks et al. 2010, http://galaxy.psu.edu/) which provides a web-based platform to facilitate accessibility of NGS data analysis, exploiting the latest informatics tools, while tracking data provenance and ensuring reproducibility of analysis pathways undertaken. Galaxy is intended to free users from the necessity to develop computer code and the need to learn the implementation details of individual software packages. Galaxy offers a framework for performing exome studies which enables reconstruction of the analysis pathways undertaken by capturing details of analyses performed through a web

interface. Perhaps most significant, given that that exome sequencing will shortly be superseded by far more challenging whole genome sequencing, is the development of a cloud computing enabled version (http://www.genomeweb.com/informatics/galaxy-joins-host-bioinformatics-projects-embracing-cloud-infrastructure-option). Cloud computing, in which computation is offered as a service, provides access to much greater computational power and storage than is available to an individual lab. Cloud computing is therefore regarded as a route to reducing some of the concerns about the management and analysis from the ongoing and developing NGS 'data deluge'.

9.4 Conclusions

A range of strategies are being employed to exploit exome sequencing for the identification of rarer variation underlying Mendelian disease and complex traits. Genotyping a small number of affected individuals in families showing strongly Mendelian patterns of inheritance has already proven to be a highly successful strategy with several important genes identified. Such an approach relies on the sharing of underlying causal variant(s) between family members. With higher penetrance variants, it is possible to combine evidence from linkage in these scenarios to reduce the list of potential causal targets. Thus, targeted follow-up can focus on the variants identified in these regions. For more complex phenotypes, strategies include investigating cases with 'extreme' or otherwise unusual phenotypes (e.g. early onset disease, well-defined disease subtypes). Such an approach assumes that a relatively small number of moderate-penetrance variants might emerge as contributory to disease. In this situation family-based designs, where possible, are likely to reduce the overall complexity and number of targets for follow-up. Extensive filtration based on known or predicted gene function further delimits variants for greater consideration. From the study of cancer genomes, novel somatic variation can be identified by filtering out germline variation.

With respect to all studies involving complex disease in unrelated individuals, statistical analysis is plagued by low power and one strategy is to combine rare variants for analysis using some form of 'collapsing' approach.

In the longer term whole genome sequencing will replace exome sequencing and provides a range of new problems. The most obvious of these arises from the fact that it is now possible to produce DNA sequence more quickly and cheaply than the computing infrastructure can be developed to manage it (Stein 2010). Indeed the cost of sequencing is now decreasing much faster than the cost of storage of the data, and storage costs are likely to exceed the cost of production in the near future. Further development of novel strategies including cloud computing, in which hardware, runtime and data storage are effectively rented for specific projects, offers a credible way forwards. The Galaxy package has been implemented successfully on the Elastic Compute Cloud (EC2) web service offered by Amazon and provides a comprehensive range of cloud-enabled tools for NGS analysis. Such developments

are promising although, as Stein (2010) points out, there remain major obstacles with respect to the network bandwidth and the transfer of huge volumes of data on and off networks. It is clear that the future development and application of NGS offers both great promise and major challenges.

References

Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
Alkan C, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7.
Bainbridge MN, et al. Whole exome capture in solution with 3Gbp of data. Genome Biol. 2010;11:R62.
Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695–701.
Burrows M, Wheeler D. A block sorting lossless data compression algorithm, Technical report 124. Palo Alto: Digital Equipment Corporation; 1994.
Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.
Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11:415–25.
Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40.
Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011;35:S12–7.
Feng B-J, et al. Design considerations for massively parallel sequencing studies of complex human disease. PLoS One. 2011;6(8):e23221.
Girard SL, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet. 2011;43(9):860–4.
Goecks J, Nekrutenko A, Taylor J, Team TG. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11:R86.
Harismendy O, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10:R32.
Howrigan DP, et al. Mutational load analysis of unrelated individuals. BMC Proc. 2011;5(Suppl 9):S55.
Johansen CT, et al. Mutation skew in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010;42(8):684–7.
Koboldt DC, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283–5.
Krawitz PM, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphoshatasia mental retardation syndrome. Nat Genet. 2010;42(10):827–9.
Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007;80(4):727–39.
Kumar P, et al. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. Nat Protoc. 2009;4:1073–81.
Kumar S, Dudley JT, Filipski A, Liu L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends Genet. 2011;27(9):377–86.
Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187097.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
Lehne B, Lewis CM, Schlitt T. Exome localization of complex disease association signals. BMC Genomics. 2011;12:92.
Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754–60.
Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473–83.
Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.
Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851–8.
Li H, et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 2009a;25:2078–9.
Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009b;25(15):1966–7.
Liu X, Jian X, Boerwinkle E. DbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894–9.
Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. Genome Res. 2011;21:1099–108.
Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011;48:580–9. https://doi.org/10.1136/jmedgenet-2011-100223.
McClellan J, King MC. Genetic heterogeneity and human disease. Cell. 2010;141:210–7.
Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6.
Ng SB, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet. 2010a;42:30–5.
Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. Hum Mol Genet. 2010b;19:R119–24.
Pireddu L, Leo S, Zanetti G. SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics. 2011;27(15):2159.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.
Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86:832–8.
Reis-Filho JS. Next-generation sequencing. Breast Cancer Res. 2009;11(Suppl 3):S12.
Rivas MA, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066–75.
Ruffalo M, LaFramboise T, Koyuturk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics. 2011;27:2790–6. https://doi.org/10.1093/bioinformatics/btr477.
Rumble SM, et al. SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol. 2009;5(5):e1000386.
Sathirapongsasuti JF, et al. Exome sequencing-based copy number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics. 2011;27:2648–54. https://doi.org/10.1093/bioinformatics/btr462.
Schwartz JM, et al. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575–6.
Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11:207.
Tiacci E, et al. BRAF mutations in hairy-cell leukemia. N Engl J Med. 2011;364(24):2305–15.
Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Res. 2010;38:e164.
Yoon S, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19:1586–92.