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).

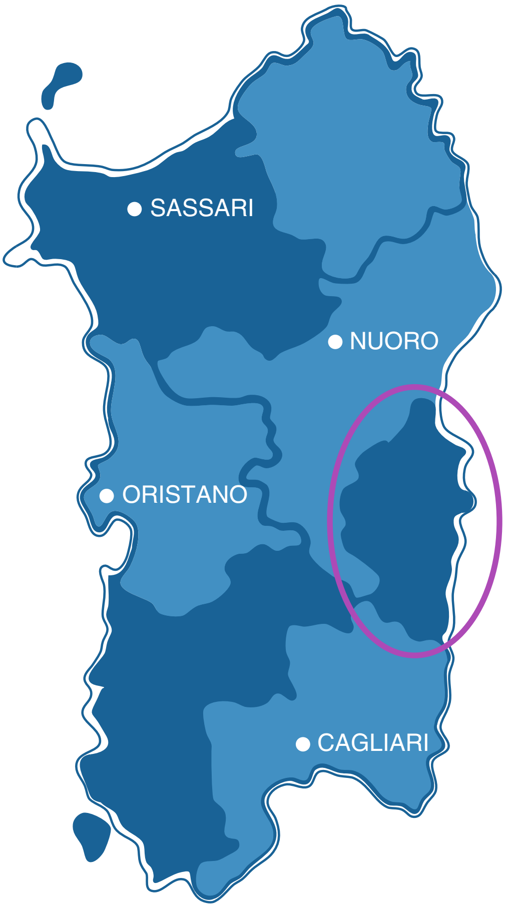
Use of Population Isolates in Genetic Studies

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 AAKKOSJÄ RJESTYKSESSÄ 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

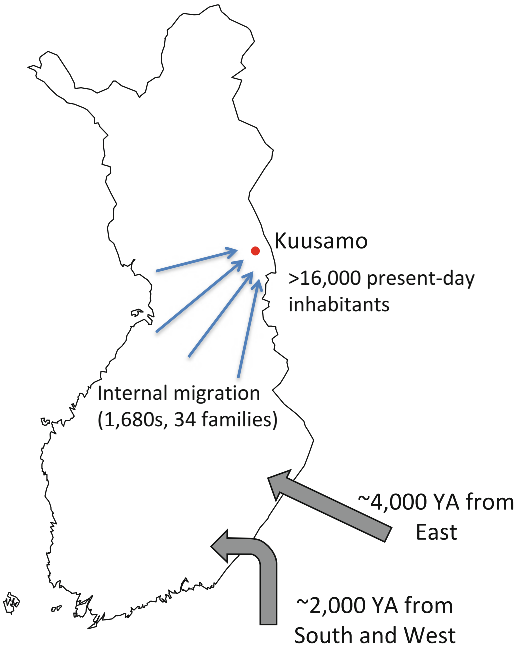

Kuusamo
>16,000 present-day inhabitants
Internal migration (1,680s, 34 families)
~4,000 YA from East
~2,000 YA from South and West

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

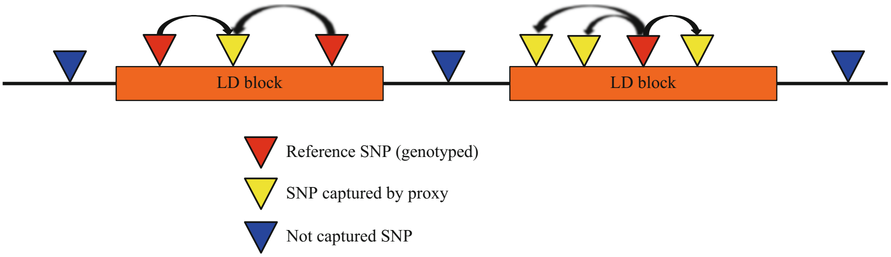


Fig. 3 Schematic representation of a genomic region to be tested for association with a phenotype. Genotyped SNPs (in *red*) are tested directly. Other associations are captured through linkage disequilibrium (by proxy) with the reference SNPs. The three SNPs indicated by *blue triangles* are neither 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

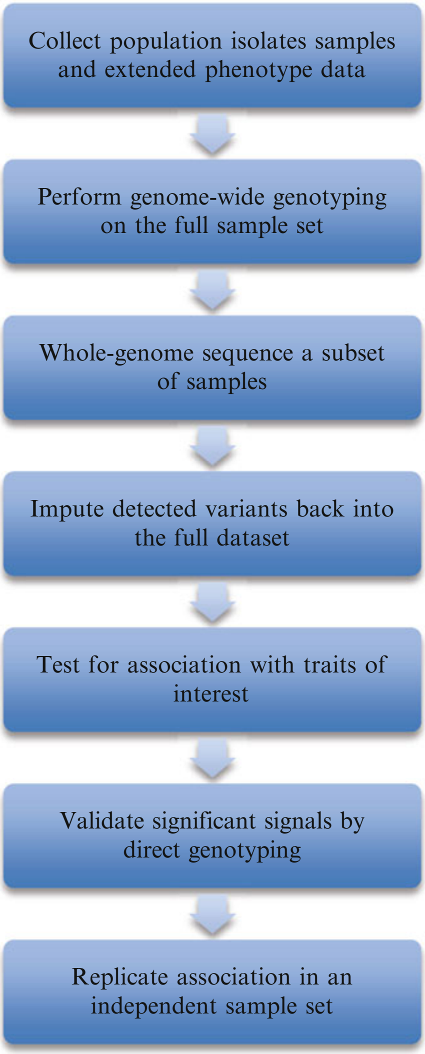

Collect population isolates samples and extended phenotype data
Perform genome-wide genotyping on the full sample set
Whole-genome sequence a subset of samples
Impute detected variants back into the full dataset
Test for association with traits of interest
Validate significant signals by direct genotyping
Replicate association in an independent sample set

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),

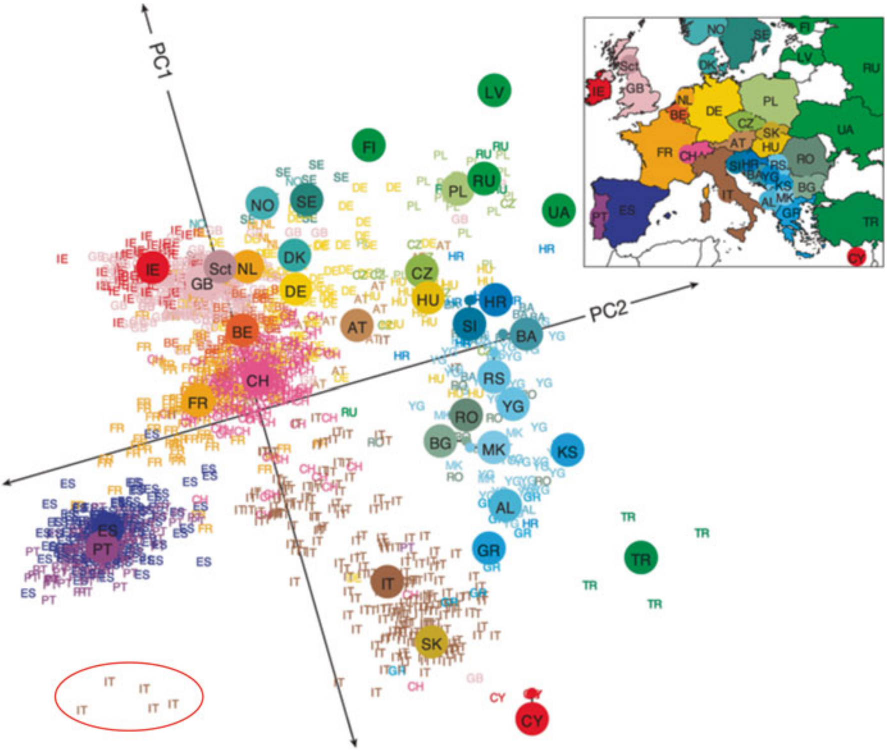


Fig. 5 Population structure within Europe (adapted from Novembre et al. 2008)

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