1. Boycott, K.M., Vanstone, M.R., Bulman, D.E. & MacKenzie, A.E. *Nat. Rev. Genet.* **14**, 681–691 (2013).
2. de Ligt, J. *et al. N. Engl. J. Med.* **367**, 1921–1929 (2012).
3. Yang, Y. *et al. J. Am. Med. Assoc.* **312**, 1870–1879 (2014).
4. Deciphering Developmental Disorders Study. *Nature* **519**, 223–228 (2015).
5. MacArthur, D.G. *et al. Nature* **508**, 469–476 (2014).
6. Akawi, N. *et al. Nat. Genet.* **47**, 1363–1369 (2015).
7. Guimier, A. *et al. Nat. Genet.* **47**, 1260–1263 (2015).
8. Robinson, P.N. *et al. Am. J. Hum. Genet.* **83**, 610–615 (2008).
9. Stessman, H.A., Bernier, R. & Eichler, E.E. *Cell* **156**, 872–877 (2014).
10. Samocha, K.E. *et al. Nat. Genet.* **46**, 944–950 (2014).
11. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E.E. *Genome Res.* **25**, 142–154 (2015).
12. Groza, T. *et al. Am. J. Hum. Genet.* **97**, 111–124 (2015).

Small island, big genetic discoveries

Guillaume Lettre & Joel N Hirschhorn

Three new studies have identified new genes and sequence variants implicated in blood lipids, inflammatory markers, hemoglobin levels and adult height variation in Sardinia. These reports highlight the usefulness of large-scale genotype imputation based on whole-genome sequencing, particularly in isolated populations, in studying the genetics of complex human phenotypes.

Conceptually, finding genetic variants that influence a heritable trait or disease in humans is simple: it requires phenotypes, genotypes and a test to measure the strength of their correlation. The difficulty resides in the fact that phenotype descriptions are imperfect, and comprehensive, sequence-level genetic data are difficult or expensive to obtain. One particularly promising solution has been to work with genetic isolates, such as French Canadians from Quebec, Finns, Ashkenazi Jews and Icelanders. Because of their isolation, these populations tend to have reduced genetic diversity and a more uniform environment (**Fig. 1**) (ref. 1). These advantages have proven immensely useful in studying the genetics of monogenic disorders and polygenic diseases and traits. In this issue of *Nature Genetics*, work from Sidore *et al.*[2], Danjou *et al.*[3] and Zoledziewska *et al.*[4] in the Sardinian population demonstrates once again the value of well-characterized human founder populations in gene discovery efforts.

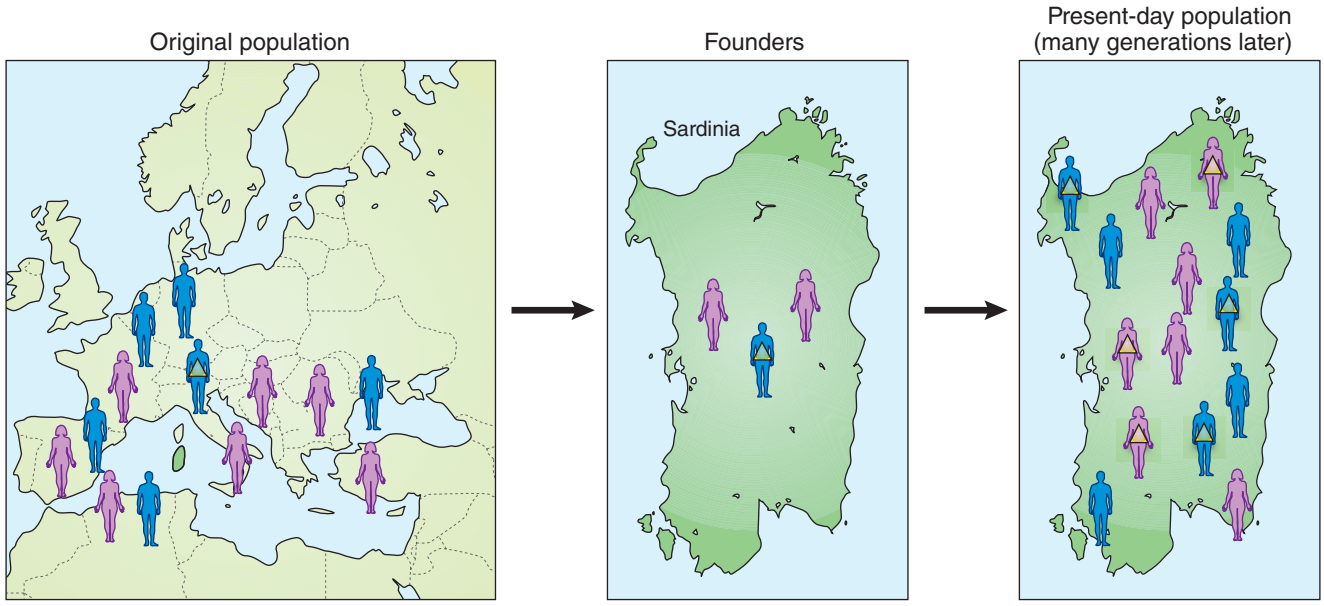
The teams report results from whole-genome DNA sequencing in 2,120 Sardinians: they identify ~17.6 million DNA sequence variants, of which 22% are absent from publicly available databases. This number also includes ~76,000 markers that are common in Sardinia (minor allele frequency (MAF) >5%) but rare elsewhere in the world (MAF <0.5%). This shift toward higher allele frequencies—a cardinal feature of founder populations—is important because statistical power to find genetic associations increases with allele frequency. The large sequence data set and genetic homogeneity of Sardinia also enabled highly effective imputation using array-based genotype information that was available for >6,300 other Sardinians[5]. Notably, imputation accuracy for both common and rare variants was better when using Sardinian haplotypes than haplotypes from the 1000 Genomes Project[6].

Guillaume Lettre is at the Montreal Heart Institute, Montreal, Quebec, Canada, and the Faculté de Médecine, Université de Montréal, Montreal, Quebec, Canada. Joel N. Hirschhorn is at the Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA, the Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA, and the Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: guillaume.lettre@umontreal.ca **or** joelh@broadinstitute.org

Lipids, inflammatory markers and hemoglobins

The investigators sought to identify novel or fine map existing genetic associations with blood lipid levels[2]. For total and low-density lipoprotein (LDL) cholesterol, the strongest new association they found was with a nonsense variant in the *HBB* gene (encoding β-globin), a well-known mutation (termed $\beta^0$39; encoding p.Gln40*) that causes β-thalassemia in homozygosity. Previously, the same group had reported an association between cholesterol levels and SNPs located 122 kb from the β-globin gene. In that analysis, the investigators used 1000 Genomes Project haplotypes for imputation, in which the $\beta^0$39 mutation is too rare to be properly imputed. Using haplotypes from the Sardinian whole-genome sequencing project, where $\beta^0$39 is common (MAF = 4.8%), they could accurately impute this mutation and show that the association with reduced cholesterol levels at this locus is explained by $\beta^0$39, illustrating the importance of comprehensive ascertainment of variation for fine mapping. The group also identified an *APOA5* missense variant (encoding a p.Arg282Ser substitution) strongly associated with triglyceride levels. This variant has MAF = 2.5% in Sardinians but is essentially absent in the rest of the world and would not have been found using a different reference haplotype panel for imputation. This variant has a strong phenotypic effect and explains nearly 1% of the variation in triglyceride levels in Sardinians. Interestingly, a different but equally rare mutation at the same site in *APOA5* (encoding a p.Arg282Cys substitution) was recently discovered by whole-exome sequencing in much larger numbers of European-ancestry individuals and implicated in myocardial infarction risk[7]. In the same study[2], the researchers also analyzed genetic associations with five inflammatory markers. The most striking new finding is a strong association of two markers (C-reactive protein (CRP) levels and erythrocyte sedimentation rate (ESR)) with a collection of 22 rare non-coding variants that span 5.4 Mb on chromosome 12. These rare variants appear to be specific to Sardinia, and common markers at the locus cannot explain the association signal. It is likely that identifying the causal gene(s) and elucidating the molecular mechanism(s) at play at this locus will require extensive functional experiments.

In Danjou *et al.*[3], the authors tested associations with three different hemoglobin macromolecules (fetal hemoglobin (HbF) and adult hemoglobins A1 (HbA1) and A2 (HbA2)). Determinants of between-individual variation in these macromolecules is important in the context of several hemoglobinopathies, including β-thalassemia and sickle cell disease, because hemoglobin isoform levels influence clinical severity[8]. Genome-wide association



Katie Vicari/Nature Publishing Group

Figure 1 Sardinia was initially populated by people who originated mostly from the Iberian and Italian peninsulas. If one of these founders carried a private or rare mutation (yellow triangle), the frequency of this mutation could increase substantially after many generations owing to the bottleneck effect followed by rapid population growth.

studies (GWAS) have already identified dozens of loci associated with HbA1 levels and a handful of additional associations with HbA2 and HbF levels. In their analysis of the Sardinian data set, the investigators found five new associations. These include the association of HbA2 levels with a rare missense variant in *ZFPM1* (also known as *FOG1*; encoding p.Asn972Ser; MAF = 0.7%) and the association of HbF levels with an intronic variant in *NFIX*. Both of these genes encode transcription factors previously implicated in hematopoiesis. The *NFIX* variant has a frequency of 1% in Sardinians but, apart from Spain (MAF = 0.5%), is absent in Europeans from the 1000 Genomes Project. The variant site is also polymorphic in some African and South American populations. Before the current effort, this marker was difficult to impute and had escaped detection in previous GWAS for HbF levels.

The *NFIX* finding, along with other findings from these new reports, highlights a significant challenge for genetic studies in founder populations. Because such variants are often private to the population in which association is identified or are very rare elsewhere, it is nearly impossible to replicate the detected associations in independent samples unless additional samples from the same population are available. For example, the *NFIX* variant site was monomorphic in the large TwinsUK cohort and thus could not be tested. Replication is the gold standard to validate genetic association results, and, in its absence, additional information should be carefully considered and claims ideally supported with wet-lab experiments. Alternatively, founder populations may provide evidence pointing to a particular gene that could be supported by findings in other populations for different variants within the same gene (as in the *APOA5* example).

Short stature in Sardinia

In Zoledziewska *et al.*[4], the authors investigated the genetics of height for Sardinians, one of the shortest populations in Europe. Height is a classic polygenic trait, and GWAS have already found ~700 variants associated with adult stature[9]. The team identified two variants that are more frequent in Sardinia and have a relatively strong effect on height when compared with the effect of the GWAS findings (2–4 cm versus <0.3 cm)[4]. The first variant is a nonsense mutation in the *GHR* gene (encoding growth hormone receptor); mutations in *GHR* have been associated with short stature (Laron syndrome). The second variant at the *KCNQ1* locus falls within an imprinted region and has a strong association with short stature when maternally inherited. The investigators also found that there is specific enrichment of height-decreasing alleles in Sardinia when considering all ~700 known variants implicated in height genetics. Excitingly, this enrichment suggests that natural selection may have favored short stature in Sardinia, although the precise reason(s) for this selective pressure remain unclear.

In conclusion, these three articles have identified many variants with relatively strong effects on clinically significant traits. They confirm that whole-genome sequencing followed by imputation in large genotyped samples is a viable strategy for the discovery of associated genes and that population isolates can provide an important dimension to this strategy. Furthermore, many phenotypes have not been tested yet, and whole-genome sequencing will allow extension to insertion-deletion variants (indels) and other large structural rearrangements, leading to additional new discoveries in Sardinia and beyond.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Peltonen, L., Palotie, A. & Lange, K. *Nat. Rev. Genet.* **1**, 182–190 (2000).
2. Sidore, C. *et al. Nat. Genet.* **47**, 1272–1281 (2015).
3. Danjou, F. *et al. Nat. Genet.* **47**, 1264–1271 (2015).
4. Zoledziewska, M. *et al. Nat. Genet.* **47**, 1352–1356 (2015).
5. Kong, A. *et al. Nat. Genet.* **40**, 1068–1075 (2008).
6. 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).
7. Do, R. *et al. Nature* **518**, 102–106 (2015).
8. Sankaran, V.G., Lettre, G., Orkin, S.H. & Hirschhorn, J.N. *Ann. NY Acad. Sci.* **1214**, 47–56 (2010).
9. Wood, A.R. *et al. Nat. Genet.* **46**, 1173–1186 (2014).