Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers

Carlo Sidore1–3,19, Fabio Busonero1,2,4,19, Andrea Maschio1,2,4,19, Eleonora Porcu1–3,19, Silvia Naitza1,19, Magdalena Zoledziewska1, Antonella Mulas1,3, Giorgio Pistis1–3, Maristella Steri1, Fabrice Danjou1, Alan Kwong2, Vicente Diego Ortega del Vecchio5, Charleston W K Chiang6, Jennifer Bragg-Gresham2, Maristella Pitzalis1, Ramaiah Nagaraja7, Brendan Tarrier4, Christine Brennan4, Sergio Uzzau8, Christian Fuchsberger2, Rossano Atzeni9, Frederic Reinier9, Riccardo Berutti3,9, Jie Huang10, Nicholas J Timpson11, Daniela Toniolo12, Paolo Gasparini13,14, Giovanni Malerba15, George Dedoussis16, Eleftheria Zeggini10, Nicole Soranzo10,17, Chris Jones9, Robert Lyons4, Andrea Angius1,9, Hyun M Kang2, John Novembre18, Serena Sanna1,20, David Schlessinger7,20, Francesco Cucca1,3,20 & Gonçalo R Abecasis2,20

We report ~17.6 million genetic variants from whole-genome sequencing of 2,120 Sardinians; 22% are absent from previous sequencing-based compilations and are enriched for predicted functional consequences. Furthermore, ~76,000 variants common in our sample (frequency >5%) are rare elsewhere (<0.5% in the 1000 Genomes Project). We assessed the impact of these variants on circulating lipid levels and five inflammatory biomarkers. We observe 14 signals, including 2 major new loci, for inflammatory markers. The new associations would have been missed in analyses based on 1000 Genomes Project data, underlining the advantages of large-scale sequencing in this founder population.

Studies of common genetic variants have provided entry points to analyze the mechanisms underlying many complex traits and diseases1–4. Extension of these studies to the large reservoir of rare and population-specific variants could accelerate the translation of genetic information into biological understanding but has not thus far been systematically applied5,6. Rare variants can be discovered and genotyped with rapidly improving DNA sequencing techniques, but designing studies in which enough copies of each variant can be observed to detect genetic associations is challenging5–7. Studies of families and founder populations, where variants that are rare or absent elsewhere can occur at moderate frequencies, help overcome these limitations8. Here we use genome sequencing in the Sardinian founder population to systematically assess the contribution of genetic variation to quantitative traits, using as examples the levels of blood lipids and inflammatory markers. Discovery of variants associated with these traits could further elucidate causal mechanisms and pathways for cardiovascular diseases and other complex disorders9–11. In addition to confirming signals from studies of common variants12,13, our results identify new genetic variants and associations that would be missed using sequence-based reference panels derived from more cosmopolitan populations.

RESULTS
Sequencing and rare variant yield
We generated whole-genome shotgun sequence data for 2,120 Sardinian individuals, either living in the Lanusei valley and participating in a cohort study of quantitative traits (the SardiNIA study14; 1,122 individuals; 52.8% female and average age of 49.4 years) or from across the island and participating in case-control studies of multiple sclerosis15 and type 1 diabetes16 (referred to here as the...
Table 1 Summary of the discovered variants

| Coding variation | Genome | Regulatory | Silent | Splice site | Essential splice site |
|------------------|--------|------------|--------|-------------|-----------------------|
| Missense         | 84,312 | 2,013      |        |             |                       |
| Nonsense         | 2,013  |            |        |             |                       |

The table provides an overview of the sequencing data, including summary statistics on the data generated, a breakdown by frequency and biological function of all variants discovered, and novelty rates based on comparison to public databases. Finally, we show the distribution of the variants discovered for each sequenced individual. In total, we mapped 22.684 Gb with an average depth of 4.16x.

For each sequenced individual, we identified an average of 3.4 million variants (17.6 million variants overall; Table 1). To assess quality, we sequenced one parent-offspring trio to >65× coverage for each individual. Comparing our initial low-coverage analysis with the results of deep sequencing for this family trio, we estimate an average genotyping error rate of <0.7% at heterozygous sites. As expected, this error rate was lower at sites with minor allele frequency (MAF) >5%, averaging 0.5%, and higher at sites with MAF <5%, averaging ~2% (Supplementary Table 1). Comparing sequence and array genotyping results for 1,068 individuals, we estimate that we have discovered and genotyped >99% of the variants with frequency >0.5% in our sample and ~70% of the variants with frequency <0.5% (Supplementary Table 2).

Of the 17.6 million variants discovered, 172,988 (0.98%) overlapped protein-coding sequences (Table 1). Of these variants, 84,312 were nonsynonymous coding changes.

Figure 1 Geographical differentiation based on common and rare sites. The figure shows allele sharing by Sardinians and the 1000 Genomes Project European populations. (a) Differentiation is represented for three different frequency intervals over the geographical map of Europe. The thickness and color of the lines connecting the dots are proportional to the allele-sharing statistic. (b) The relationship between frequency (evaluated on 360 chromosomes) (x axis) and sharing ratio (y axis) is represented for different 1000 Genomes Project populations (continuous lines). Results are plotted separately for the Sardinia Lanusei valley sample (left) and the case-control (CSCT) samples (right). The dashed lines are used for comparison to show the sharing ratio between TSI and other 1000 Genomes Project populations. IBS, Iberian population in Spain; TSI, Tuscans in Italy; GBR, British from England and Scotland; CEU, Utah residents with Northern and Western European ancestry; FIN, Finnish in Finland; YRI, Yoruba in Ibadan, Nigeria; CHB, Han Chinese in Beijing, China.
haplotypes surrounding $f_2$ variants in Sardinians and populations in the 1000 Genomes Project. We determined the lengths of the shared haplotypes surrounding $f_2$ variants shared by one of our sequenced individuals and one of 100 randomly selected individuals sampled from our study or from a particular 1000 Genomes Project population. (a) The lengths of the shared haplotypes in comparisons between Sardinians and several 1000 Genomes Project populations. (b) The number of $f_2$ variants in each comparison. (c) The number of $f_2$ variants in comparisons within Sardinia (note the wider y-axis range). SNIA, Sardinians from the Lanusei valley.

2,504 were variants altering essential splice sites and 2,013 were nonsense variants. Consistent with the hypothesis that natural selection makes variants with strong biological impact more likely to be rare and/or geographically restricted, we observed that 59% of the nonsynonymous variants, 53% of the splice-site altering variants and 70% of the nonsense variants had frequency <0.5% (in comparison to 48% of variants across the genome). We also observed that 12% of the nonsynonymous variants, 22% of the splice-site variants and 22% of the nonsense variants were absent from previous sequencing studies (as compared to 22% of all variants, using dbSNP 142 and the Exome Aggregation Consortium (ExAC) databases (see URLs) as surrogates for the results of previous studies).

Genetic differentiation

Because of genetic drift and, to a lesser extent, natural selection following the settlement of Sardinia, many genetic variants that are rare elsewhere in Europe have now reached higher frequencies in Sardinia.22,23. The consequences of this genetic differentiation include a relatively large fraction of population-specific low-frequency variants and long haplotypes shared among present-day carriers of these variants.24. For example, 98% of the variants present at a frequency of ~1.0% (and 99.7% of the variants present at a frequency of ~5.0%) in a sample of ~2,500 individuals from the UK are also present in Phase 1 of the 1000 Genomes Project.25. By contrast, only 77% of the variants with a frequency of ~1.0% (and 99.3% of the variants with a frequency of ~5.0%) in our sample are present in Phase 1 of the 1000 Genomes Project.25. Overall, we estimate that 76,286 variants that are very rare (frequency <0.5%) or absent in the 1000 Genomes Project Phase 3 panel reach frequencies >5% in our sample. We used a machine learning–based scoring algorithm to summarize the deleteriousness of each variant in a Combined Annotation-Dependent Depletion (CADD) score26. Coding variants that are unique to Sardinia appear to be significantly more deleterious than variants of the same frequency that are also observed in the 1000 Genomes Project Phase 3 panel ($P = 0.02$) (Supplementary Fig. 1).

The differentiation of allele frequencies in the Sardinian sample from those in other European populations is also evident in assessments using the $F_{ST}$ differentiation statistic as well as in a principal-component analysis (PCA) of common variants.27,28 (Supplementary Figs. 2 and 3). Whereas $F_{ST}$ between non-Sardinian European populations in the POPRES reference sample averaged 0.001 (range of 0.000–0.004), $F_{ST}$ between the island-wide sample of Sardinians and POPRES European populations averaged 0.006 (range of 0.003–0.010), and the difference was even greater between the Lanusei valley cohort and the POPRES European populations (database of Genotypes and Phenotypes [dbGaP], phs000145.v4.p2) (average of 0.009, range of 0.006–0.013) (Supplementary Fig. 2). The geographical structure is even more evident when considering less frequent alleles: for rare sites, allele sharing by mainland populations and Sardinians is particularly depressed relative to sharing within mainland populations (such as the 1000 Genomes Project CEU (Utah residents of Northern and Western European ancestry) and TSI (Tuscans from Italy) populations)29,30 (Fig. 1). The patterns of differentiation are again clear in the long identical haplotypes surrounding rare $f_2$ variants (variants that are observed on exactly two chromosomes from distinct individuals)25,31 (Fig. 2). Of note, both Sardinian samples showed similar lengths for haplotypes flanking $f_2$ variants that are shared with populations outside Sardinia, consistent with a common ancient demography. The greater relative isolation of the two Sardinian samples was evident when we examined the lengths of the haplotypes flanking the $f_2$ variants present within each sample. For variants shared by individuals in the Lanusei valley, the flanking haplotypes averaged 3,570 kb, with the length dropping to 735 kb when first- and second-degree relatives were excluded from consideration. The haplotypes averaged 580 kb when the variant was shared by a Lanusei valley resident and an individual elsewhere in Sardinia; ~382 kb when the variant was shared with a European sequenced in the 1000 Genomes Project Phase 3 panel; and ~264 kb when the variant was shared with an individual elsewhere in the world in the full set of 1000 Genomes Project data (Fig. 2). These differences in haplotype length were less marked around variants with higher frequencies and hence shared by more than two heterozygous individuals ($f_3$, $f_4$, etc.). This was evident even when comparing samples from the Lanusei valley and elsewhere on the island (Supplementary Table 3).

Relatedness and imputation for the Lanusei valley samples

Participants in the SardiNIA study all live in four towns in the Lanusei valley. The population in this region is relatively stable: all four grandparents were born in the Lanusei valley for at least three-quarters of the study participants.14. A total of 6,602 individuals from the SardiNIA study were genotyped with 4 Illumina arrays (OmniExpress, ExomeChip, Metabochip and Immunochip), providing a scaffold of 890,542 unique SNPs across the genome. Because participants share long stretches of DNA, the genetic information obtained for any individual can be propagated (‘imputed’) to close relatives genotyped with the scaffold of markers32,33. To increase the
Table 2 Summary of lipid association results

| Candidate gene | Chromosome: position | rsID | Alleles (effect/other) | Freq. | Effect (SE) | \( R^2 \) (%) | RSQR | Variant consequence | \( P \) value | \( R^2 \) with previous hit |
|----------------|----------------------|------|------------------------|-------|-------------|---------------|-------|---------------------|-------------|--------------------------|
| LDL            | 1:55505647           | rs11591147 | T/G                  | 0.038 | –0.406 (0.053) | 1.73 × 10^{-14} | 1.0 | Genotyped            | Missense, p.Arg46Leu | Same SNP |
| PCSK9          | 1:59823107           | rs583104 | G/T                  | 0.180 | 0.156 (0.027)  | 1.87 × 10^{-8}  | 0.5 | Genotyped            | Downstream   | 0.821 |
| SORT1          | 11:5248004           | rs11549407 | A/G                | 0.048 | –0.473 (0.051) | 1.17 × 10^{-20} | 1.5 | Stop gain, p.Gln40*  | –           | – |
| HBB            | 19:45419914          | rs429358b | C/T                 | 0.074 | 0.264 (0.039)  | 1.21 × 10^{-11} | 0.8 | 0.999               | Missense, p.Cys130Arg | Same SNP |
| Total cholesterol | 4:41980435  | –    | G/A                  | 0.013 | 0.520 (0.091)  | 6.94 × 10^{-9}  | 0.6 | 0.91                | Intergenic   | – |
| PCSK9          | 1:55505647           | rs11591147 | T/G                  | 0.038 | –0.390 (0.053) | 1.69 × 10^{-13} | 1.0 | Genotyped            | Missense, p.Arg46Leu | Same SNP |
| HBB            | 11:5248004           | rs11549407 | A/G                | 0.048 | –0.490 (0.05)  | 6.88 × 10^{-22} | 1.5 | Stop gain, p.Gln40*  | –           | – |
| APOE           | 19:45412079          | rs7412  | T/C                  | 0.036 | –0.645 (0.053) | 2.47 × 10^{-33} | 2.4 | Genotyped            | Missense, p.Arg176Cys | Same SNP |
| APOE           | 19:45411914          | rs429358b | C/T                 | 0.074 | 0.264 (0.039)  | 1.21 × 10^{-11} | 0.8 | 0.999               | Missense, p.Cys130Arg | Same SNP |
| HDL            | 4:41980435           | –    | G/A                  | 0.013 | 0.520 (0.091)  | 6.94 × 10^{-9}  | 0.6 | 0.91                | Intergenic   | – |
| LPL            | 18:3412386           | rs8092903 | T/C                 | 0.026 | –0.448 (0.082) | 4.49 × 10^{-8}  | 0.8 | 0.945               | Intergenic   | – |

The table lists association signals that reached \( P < 5 \times 10^{-8} \) for association with lipid levels in our study. At each new locus, we indicate the gene(s) likely to be modulated by the lead SNP, the location of the lead variant (human genome build GRCh37), the variant identifier rsID, the nearest gene, the effect allele and the other allele, the frequency of the

**Impact on association of lipid and inflammatory markers**

We focused on the levels of four blood lipids (low-density lipoprotein (LDL) cholesterol, total cholesterol, triglycerides and high-density lipoprotein (HDL) cholesterol) to assess how sequence information might be used to discover the effects of population-specific and low-frequency variation on extensively studied traits (Supplementary Table 4)\(^{12}\). Imputing variants from the sequencing effort on the scaffold of genotyped SNPs expanded the spectrum of variants for association testing in the sample from the Lanusei valley to \( \sim 13.6 \) million (selected with high imputation quality; Online Methods)\(^{34}\). Overall, we identified 14 independently associated variants distributed across 11 loci meeting the classical genome-wide significance threshold of \( P = 5 \times 10^{-8} \) associated with lipid levels in analysis including all individuals or in sex-restricted analysis including only males or only females (Table 2 and Supplementary Figs. 6 and 7).

These included ten variants with moderate effect tagging signals in \( \text{LIPC}, \text{SORT1}, \text{PCSK9}, \text{CILP2}, \text{CEPT} \) and \( \text{APOA5} \) (one signal each) and \( \text{LPL} \) and \( \text{APOE} \) (two signals each)—loci that have been extensively described in previous genome-wide association studies (GWAS) and other association studies. Other signals at known loci were detected at lower levels of significance (Supplementary Table 5). To declare new genome-wide signals, we used a threshold of \( P = 6.9 \times 10^{-8} \), which was calculated by empirically estimating the number of independent tests in a Sardinian genome (Online Methods and Supplementary Table 6).

The results implicate three variants that are rare or absent elsewhere in the world and were missed in studies of European-ancestry samples that included >100,000 individuals\(^{12}\). We previously identified one of these loci through a Sanger sequencing–based effort\(^{35}\); the variant in \( \text{LDLR} \) (frequency = 0.5%) encoding p.Val578Ala (frequency = 0.5%) encoding p.Val578Ala (Supplementary Table 5) was associated with LDL cholesterol and total cholesterol levels and was independent of the known variant, rs73015013 (frequency = 14%; effect = –5.2 mg/dl; \( P = 6.4 \times 10^{-4}, r^2 < 0.001 \)). Here we report a new
The variant, encoding a p.Arg282Ser substitution, was genotyped and included on the ExomeChip array after it was discovered in our sequencing effort, and, thus far, it has been found on only 2 chromosomes in >30,000 Europeans characterized by the Exome Aggregation Consortium. Of note, this variant was the strongest variant modulating triglyceride levels in Sardinia, explaining almost 1% of the phenotypic variance, and was also independent of the known common variant in the locus, rs10750097 (frequency = 17%; effect = 11.9 mg/dl; \( P = 4.6 \times 10^{-9}; r^2 = 0.002 \) (Table 2)). These two examples illustrate the coexistence in the same locus of population-specific low-frequency variants and previously detected and independently associated common variants from cosmopolitan populations (Fig. 3). The third genetic variant was a stop-gain mutation in the HBB gene encoding p.Gln40*, better known as the B039 mutation because the corresponding codon was numbered 39 before the last update of standard protein nomenclature.

Figure 3 Regional association plots for new lipid-associated loci. (a,b) Regional association plots are shown at the HBB locus for LDL cholesterol levels (a) and at the APOA5 locus for triglyceride levels (b); for each locus, results are shown with imputation performed using the Sardinian (left) and 1000 Genomes Project (right) reference panels. At each locus, we plotted the association strength (−log\(_10\) \( P \) value; y axis) versus the genomic positions (on the hg19/GRCh37 genomic build; x axis) around the most significant SNP, which is indicated with a purple dot. Other SNPs in each region are colored to reflect their LD with the top SNP (taken from pairwise \( r^2 \) values calculated on Sardinian and 1000 Genomes Project haplotypes for the left and right panels, respectively). Symbols reflect genomic functional annotations. Genes and the positions of exons, as well as the direction of transcription, are noted below. The plots were drawn using the standalone version of the LocusZoom package."
variants that are unusually frequent in Sardinia can provide insights about biology. In Sardinia, this mutation is the common cause of autosomal recessive β-thalassemia. In our sample, in agreement with earlier epidemiological findings, the heterozygous state was associated with 13.9 mg/dl lower cholesterol levels (P = 1.2 × 10−26) and 16.9 mg/dl lower total cholesterol levels (P = 1.2 × 10−22). Of note, the analysis after 1000 Genomes Project Phase 3 imputation pointed only to an intergenic marker (rs7605388) 122 kb away from the β^+ variant corresponding to the second most associated SNP using the Sardinian reference panel, with a much weaker association signal (P = 1.4 × 10−13) (Fig. 3). Finally, two additional signals were observed for total cholesterol levels at rs115048493 near the TMEM33 and DCAF4L1 genes (P = 9.4 × 10−14) for HDL cholesterol levels at rs8092903 near TGFBI in females (P = 4.4 × 10−9) (Table 2 and Supplementary Table 7), although the biological bases for these associations are presently unclear. Because these signals are below our adjusted genome-wide significance threshold of P = 6.9 × 10−9, these findings remain tentative.

We were interested in determining whether analyses based on the 1000 Genomes Project and HapMap reference panels would also miss important loci for other traits. In a second example of a class of especially interesting traits, we focused on the levels of five inflammatory markers. In a previous study, assessing ~2 million genotyped and HapMap-imputed SNPs in the SardiNIA cohort, we found 16 variants associated with at least one of the four inflammatory markers measured: interleukin-6 (IL-6) concentration, erythrocyte sedimentation rate (ESR), monocyte chemotactic protein-1 (MCP-1) levels and high-sensitivity C-reactive protein (hsCRP) levels. A fifth inflammatory marker, adiponectin (ADPN), showed no significant association in our previous analyses (E.P., S.N., S.S., D.S. and E.C., unpublished data). Nevertheless, with the extended spectrum of variants assessed here, we identified another seven variants associated with MCP-1, hsCRP, ESR or ADPN, at the classical significance threshold of P = 5 × 10−8; with five variants in four previously undetected loci as well as two signals at coding variants in known variants (Table 3 and Supplementary Figs. 6–8).

Among the newly identified signals, three remained significant even with the more stringent threshold of P = 6.9 × 10−9. In comparison to analyses based on HapMap or 1000 Genomes Project imputation, we also identified more strongly associated lead variants at three known loci (APOE, HBB and RHCE) (Table 3). These signals may point to causative variants, as supported by biological evidence, expression quantitative trait locus (eQTL) data and Encyclopedia of DNA Elements (ENCODE) annotation.

In detail, we found a striking new signal associated with both hscRP (rs183233091, P = 1.1 × 10−28) and ESR (12:125406340, P = 4.4 × 10−23) on chromosome 12, in a stretch of rare variants encompassing several genes (Fig. 4). The lead variants were not the same but were partially in linkage disequilibrium (LD) (r^2 = 0.19, D′ = 0.79), and the association with hscRP disappeared when conditioning for the lead variant for ESR and vice versa. This implies that the two signals are likely due to the same variant(s), an inference that is also consistent with the biological correlation of these two traits. The rare alleles at the lead variants increased the levels of...
both inflammatory traits, with effects that appeared to be stronger in males (Supplementary Table 8). The extended associated region spanned 5.4 Mb and included 22 noncoding variants with association $P$ value $<1 \times 10^{-15}$ (Supplementary Fig. 9). The majority, to our knowledge, are specific to Sardinians, as only ten were found in either the 1000 Genomes Project Phase 3 panel or the Genomes of the Netherlands (GoNL) project database39 (four with MAF 0.1–1% and the other six with MAF >1% in Europeans). The association of the latter 6 variants with hsCRP was tested for replication in 7,689 European individuals from 8 GWAS cohorts, but no signal was seen (Supplementary Table 9), whereas nominal association was detected in a subset of 3,505 southern European individuals for the top variant (one-tailed $P$ value = 0.04). These results allow us to exclude these SNPs as causal and indicate that the association is instead primarily driven by a variant among those that are extremely rare or absent outside Sardinia.

For hsCRP, we detected additional candidate signals. There was one signal near PDGFRL, a gene previously implicated in inflammatory and autoimmune processes40,41 (Supplementary Fig. 8), which we again failed to confirm in the replication sample set. Currently, there is no other evidence that this signal is genuine, and further studies will be required to assess it. Two additional new signals reached the classical significance threshold of $P = 5 \times 10^{-8}$ but not the more stringent threshold for new findings: one for ADPN at 13:10884835 near the ABHD13 gene ($P = 3.3 \times 10^{-8}$) and another for MCP-1 at rs76135610 ($P = 1.8 \times 10^{-8}$) in a region encompassing

---

**Figure 4** Regional association plots at chromosome 12 for hSCRP and ESR. (a,b) Regional association plots are shown at the chromosome 12 locus for hSCRP (a) and ESR (b); results are shown with imputation using the Sardinian (left) and 1000 Genomes Project (right) reference panels. For further details on plot components, see the legend for Figure 3.
We uncovered two new independent functional variants for MCP-1, that result in non-conservative Cys130Arg (rs34599082) in DARC, causes the ability to bind chemokines, and the p.Arg249Cys substitution in the transmembrane domain of CCR2 is expected to affect molecular interactions and thereby alter the downstream signal transduction of neuronal signaling.

| Gene     | Chromosome start-end | n SNPs | n passing | Burden count | Effect (SE) | P value | R² | Adjusted P value | Direction | P value | R² | Adjusted P value |
|----------|-----------------------|--------|-----------|--------------|-------------|---------|----|------------------|-----------|---------|----|------------------|
| ADPN     | 3:52535766-52558237   | 25     | 23        | 752          | 0.12886     | 0.245 (0.039) | 4.71 × 10⁻¹⁰ | 0.007 | 1.92 × 10⁻⁹       | +         | 1.00 × 10⁻⁷ | 0.007 | 1.00 × 10⁻⁷       |
| MCP-1    | 4:6399158-46401290    | 4      | 3         | 105          | 0.01797     | 0.541 (0.104) | 1.84 × 10⁻⁷ | 0.005 | 0.7092           | +         | 1.00 × 10⁻⁶ | 0.005 | 0.92             |
| IFI16    | 1:158979940-159024686 | 10     | 8         | 567          | 0.09702     | 0.218 (0.046) | 2.50 × 10⁻⁵ | 0.004 | 0.1564           | +         | 1.00 × 10⁻⁷ | 0.003788 | 0.115           |
| ESR      | 2:15349906-556766     | 5      | 3         | 529          | 0.08904     | -0.205 (0.042) | 1.23 × 10⁻⁶ | 0.004 | 0.345            | -         | 1.00 × 10⁻⁷ | 0.013 | 0.025            |
| STAB1    | 11:5247914-5248004    | 2      | 2         | 613          | 0.10318     | -0.345 (0.039) | 9.77 × 10⁻¹⁹ | 0.013 | 0.015            | +         | 1.00 × 10⁻⁶ | 0.013 | 0.025            |
| HBB      | 11:5247914-5248004    | 2      | 2         | 613          | 0.10318     | -0.345 (0.039) | 9.77 × 10⁻¹⁹ | 0.013 | 0.015            | +         | 1.00 × 10⁻⁶ | 0.013 | 0.025            |
| OR52H1   | 11:5247914-5248004    | 2      | 2         | 613          | 0.10318     | -0.345 (0.039) | 9.77 × 10⁻¹⁹ | 0.013 | 0.015            | +         | 1.00 × 10⁻⁶ | 0.013 | 0.025            |
| PTPRH    | 19:55693244-55716713  | 22     | 15        | 1,152        | 0.19931     | -0.146 (0.029) | 8.31 × 10⁻⁷ | 0.004 | 0.1564           | +         | 1.00 × 10⁻⁷ | 0.003788 | 0.115           |

The table shows the results for the rare variant association tests at genes passing the significance threshold for at least one of the two statistical tests (CMC and VT). Of note, no significant results were observed for LDL cholesterol, hscRP and IL-6. For each gene, we indicate the genomic location assessed for the analysis (in the hg19 genomic build), the number of available SNPs considered, the number of SNPs passing the test-specific criteria for inclusion, and the number and fraction of individuals carrying a rare allele. For the CMC test, the effect size and its standard error (SE), along with the P value and phenotypic variance explained, are reported. For the VT test, the impact on the phenotype (+, increase; −, decrease) of rare variants, the P value and the phenotypic variance explained are reported. We also report the P value observed after adjusting for the lead variant at the same or a nearby gene. Specifically, STAB1 was adjusted for rs7639267; CCR2 was adjusted for rs113403743 and rs200491743; HBB was adjusted for the best lead SNP in the region (rs76728603); and PTPRH was adjusted for the best lead SNP in the region (rs77258314). P values that remained significant after adjustment are shown in bold.
the 7 new signals reflects advantages of studying rare or population-specific variation.

Finally, we used variance-component methods to estimate the combined contribution to lipid levels and inflammatory markers of all the variants we discovered by sequencing\textsuperscript{48}. Together, the variants identified in our sequencing study and successfully imputed explain about half of the heritability for the traits under analysis, with the sole exception of hsCRP levels, for which they explain almost all of the observed trait heritability (Supplementary Fig. 10 and Supplementary Tables 16 and 17). The missing heritability that could not be explained by sequenced variants might be attributable to variants not assessed here, including very rare variants that were not discovered or were poorly imputed or structural variants that were not considered in the present study.

**DISCUSSION**

Our findings, besides elucidating at an unprecedentedly deep level of resolution the genetic structure and substructure of Sardinians, demonstrate the value of whole-genome sequencing–based association studies in this founder population. In Sardinia, variants that are extremely rare in the rest of the world can reach high enough frequencies to provide clear and, in some cases, unexpected biological insights\textsuperscript{49}. For example, we found that the \textsuperscript{β}039 stop-gain mutation causing autosomal recessive \textsuperscript{β}-thalassemia\textsuperscript{36} accounts for a large fraction of variability in LDL cholesterol levels in Sardinia, second only to the APOE variants. The variant is known to be associated with enhanced erythropoiesis (ref. 36 and the companion paper\textsuperscript{50}), with heterozygous carriers having red blood cell counts 23% higher on average ($P < 1 \times 10^{-500}$). This provides a likely explanation for the decreased lipid levels in carriers: large amounts of cholesterol are required for the replenishment and regeneration of cell membranes and intracellular structures in circulating cells and their bone marrow precursors. Although this stop-gain mutation reaches a frequency of 5.0% in our sample, it is not included on standard genotyping arrays and cannot be easily imputed from the HapMap or 1000 Genomes Project panels because it is very rare outside Sardinia (1000 Genomes Project frequency <0.1%). Likewise, by cross-population exclusion mapping, we show that the new strong association signal with both hsCRP and ESR on chromosome 12 is most likely driven by a variant that is extremely rare or absent outside Sardinia.

Furthermore, coding variants that are unique to Sardinia appear to be significantly more deleterious than variants of the same frequency that are also observed in more cosmopolitan collections of samples (Supplementary Fig. 1). This suggests that part of the reservoir of variants that have drifted to higher frequencies in Sardinia and were lost or are extremely rare elsewhere could be especially informative for genetic association and functional studies. The results presented here show a few clear examples.

At the same time, our observations also illustrate the difficulties that will be encountered when attempting to replicate founder variant association results, with the new signals we identified being typically due to variants that are extremely rare or absent elsewhere in the world. In our view, when the variant is present in other populations, evidence for association there could be used to confirm the signal and lack of association could be used to exclude variants as being causal. However, when rare and founder variants are not shared, as will often be the case, confirming the validity of results will require accumulating additional samples in the population initially studied or may depend increasingly on additional criteria such as examination of association at other variants in the same genomic region or the use of more stringent significance levels. Our study demonstrates the benefits of combining high-throughput sequencing and genotyping technologies with imputation methods and customized study designs; we obtained high-quality information on the genomes of >6,000 individuals for an investment that, using conventional deep whole-genome sequencing strategies, would have allowed deep sequencing of only 160–180 genomes. This cost-effective approach increases power in genetic analysis (ref. 19 and the companion papers\textsuperscript{50,51}) and creates the bases for larger research and personalized medicine programs.

**URLs.** Exome Aggregation Consortium (ExAC) Browser, http://exac.broadinstitute.org/; GotCloud, http://genome.sph.umich.edu/wiki/GotCloud; EPACTS, http://genome.sph.umich.edu/wiki/EPACTS; GCTA, http://www.complextraitgenomics.com/software/gcta/; GEMMA, http://www.xzlab.org/software.html; SNPTEST, http://mathgen.stats.ox.ac.uk/geNatistics_software/snpext/snpext.html; UCSC Genome Browser, http://genome.ucsc.edu/; Genevar eQTL browser, http://www.sanger.ac.uk/resources/software/genevar/; NCBI eQTL browser, http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi; Gilad/Pritchard groups eQTL browser, http://eqtl.uchicago.edu/; Exome Picks, http://genome.sph.umich.edu/wiki/ExomePicks; ExomeChip, http://genome.sph.umich.edu/wiki/Exome_Chip_Design; 1000 Genomes Project data repository, ftp://ftp.1000genomes.ebi.ac.uk/; Beagle, http://faculty.washington.edu/browning/beagle/beagle.html.

**ACKNOWLEDGMENTS**

We thank all the volunteers who generously participated in this study and made this research possible. This research was supported by National Human Genome Research Institute grants HG006581, HG005552, HG006513, HG007022 and HG007089; by National Heart, Lung, and Blood Institute grant HL117626; by the Intramural Research Program of the US National Institutes of Health, National Institute on Aging, contracts N01-AG-1-2109 and HHSN271201100656C; by Sardinian Autonomous Region (L.R. 7/2009) grant CRP5-154; by the PB550 InterOmics MIUR Flagship Project; by grant FaReBio2011 ‘Farmaci e Reti Biotecnologiche di Qualità; by a US National Institutes of Health National Health Research Service Award (NRSA) postdoctoral fellowship (F32GM106656) to C.W.K.C; and by the UC MEXUS/CONOCYT Fellowship to V.D.O.d.V.

**AUTHOR CONTRIBUTIONS**

D.S., F.C. and G.R.A. conceived and supervised the study. C.S., S.N., S.S., D.S., F.C. and G.R.A. drafted the manuscript. E.P., M.Z., C.W.K.C. and J.N. revised the manuscript and wrote specific sections of it. F.B., A. Maschio, \textsuperscript{52}
Parkes, M. et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet. 39, 830–832 (2007).

Willer, C.J. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat. Genet. 41, 25–34 (2009).

Chen, W. et al. Genetic variants near TIMP3 and high-density lipoprotein–associated loci influence susceptibility to age-related macular degeneration. Proc. Natl. Acad. Sci. USA 107, 7401–7406 (2010).

Do, R. et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat. Genet. 45, 1345–1352 (2013).

Rathore, K., Kaiteris, S. & Abecasis, G.R. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum. Mol. Genet. 21, R1–R9 (2012).

Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA 111, E455–E464 (2014).

Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon sequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA 106, 3871–3876 (2009).

Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. Nat. Rev. Genet. 1, 182–190 (2000).

Clarke, R. et al. Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men. Arch. Intern. Med. 167, 1373–1378 (2007).

Pai, J.K. et al. Inflammatory markers and the risk of coronary heart disease in men and women. N. Engl. J. Med. 351, 2999–3010 (2004).

Orr, V. et al. Genetic variants regulating immune cell levels in health and disease. Cell 155, 242–256 (2013).

Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).

Naitza, S. et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet. 6, e1001288 (2010).

Piilia, G. et al. Heritability of cardiovascular and personality traits in 6,146 Sardinians. PLoS Genet. 2, e132 (2006).

Sanna, S. et al. Variants within the immunoregulatory CEBB gene are associated with multiple sclerosis. Nat. Genet. 42, 495–497 (2010).

Zoledziewska, M. et al. Variation within the OLECG164 gene shows consistent disease association with both multiple sclerosis and type 1 diabetes in Sardinia. Genes Immun. 10, 15–17 (2009).

Chen, W. et al. Genotype calling and haplotyping in parent-offspring trios. Genome Res. 23, 142–151 (2013).

Jun, G., Wing, M.K., Abecasis, G.R. & Kang, H.M. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. Genome Res. 25, 918–925 (2015).

Li, Y., Sidore, C., Kang, H.M., Boehnke, M. & Abecasis, G.R. Low-coverage sequencing; implications for design of complex trait association studies. Genome Res. 21, 940–951 (2011).

McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070 (2010).

Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001).
**ONLINE METHODS**

**Study samples.** To survey genetic variation across Sardinia, we selected individuals participating in the SardiNIA longitudinal study of aging or in case-control studies of multiple sclerosis or type 1 diabetes. All participants gave informed consent to study protocols, which were approved by the Sardinian local research ethic committees: Comitato Etico di Azienda Sanitaria Locale 8, Lanusei (2009/0016600) and Comitato Etico di Azienda Sanitaria Locale 1, Sassari (2171/CE) and by the NIH Office of Human Subject Research as governed by Italian institutional review board approval.

The SardiNIA project includes 6,921 individuals, representing >60% of the adult population of 4 villages in the Lanusei valley on Sardinia. Details of phenotype assessments for these samples have been published previously. In particular, LDL cholesterol levels were estimated using the Friedewald formula. Individuals with triglyceride concentrations >400 mg/dl or those taking lipid-lowering medications were excluded from the LDL cholesterol analysis, and those on medication were also excluded from analyses of all lipid traits. Summary statistics for individuals considered for the GWAS analyses are reported in Supplementary Table 3.

When array genotype data are available, sequencing a subset of individuals in a family allows for missing genotypes to be imputed in the remaining individuals by tracking haplotype segregation through the family. We used known family relationships among SardiNIA study participants and the ExomePicks program (see URLs) to prioritize individuals for sequencing. For each family, the program identifies subsets of individuals whose haplotypes can be estimated very accurately (for example, parent-offspring trios) and estimates the fraction of the genome for each additional family member that can be imputed using these haplotypes.

Our ongoing case-control studies of type 1 diabetes and multiple sclerosis include 10,106 individuals and 1,109 nuclear families, each with an affected child and 2 unaffected parents. Participants were recruited through regional clinics and hospitals distributed throughout Sardinia, with the majority of participants recruited in Cagliari (in the south of Sardinia) or Sassari (in the north). Again, we favored sequencing of parent-offspring trios to improve the accuracy of the resulting haplotypes. Part of the sequencing data used in this study are available through dbGaP, under the SardiNIA Medical Sequencing Discovery Project, study accession phs000313.v3.p2.

**Genotyping.** All SardiNIA study samples were genotyped with four different Illumina Infinium arrays: one high-density array, OmniExpress, which surveyed common variation across the genome, and three low-density targeted arrays that provided improved coverage of regions associated with cardiovascular and metabolic disease on the Cardio-Metabochip. Immunome arrays that provided improved coverage of regions associated with immune traits assessed by sequencing. To evaluate our power to discover rare variants through low-pass sequencing, we examined 1,068 samples that were both genotyped and imputed across 1,148 individuals for imputation. To reduce computational effort, we did not attempt to impute singleton variants. After imputation, we retained for association analysis only markers with an imputation quality (RSQ) >0.3 or >0.6 if the estimated MAF was 21% or <1%, respectively. For comparison, we repeated imputation using the 1000 Genomes Project Phase 3 haplotype set (using all 2,504 available samples; November 2014 release) and used RSQ >0.3 for all variants as a filter for imputation accuracy, as suggested in ref. 34. This strategy led to the identification of 13.6 million and 12.7 million markers.

**Variant calling.** Variant calling and genotyping was carried out using our GotCloud pipeline (see URLs). Briefly, GotCloud organizes large sequence analysis jobs into many small jobs that can be distributed across a high-performance computing cluster. GotCloud previously contributed to variant calls for the 1000 Genomes Project and the National Heart, Lung, and Blood Institute (NHBLI) Exome Sequencing Project. The approach examines all samples jointly to identify an initial variant list, improving the ability to detect low-frequency variants with low-coverage data. This initial list of variants is then annotated with information on sequencing depth, mapping quality, the ratio of the reference and alternate alleles at heterozygous sites, information on the evidence for alternate alleles by strand and read position, excess heterozygosity and others. This information was used to build a support vector machine (SVM)-based classifier to distinguish between true variants (such as those seen in HapMap or validated by the 1000 Genomes Project using Omni arrays) and likely false positive variants. The list of likely false positives was seeded with variants that had extreme sequencing depth and unbalanced representation of reference and alternate alleles, both by strand and position. Finally, using the list of likely high-quality sites, genotypes were estimated using the haplotype-aware calling algorithms implemented in BEAGLE (see URLs), to generate initial haplotype estimates, and TrioCaller, to refine this initial haplotype set. The entire computational process required approximately 20 years of computing time (6 CPU years for quality control and alignment and 14 CPU years for variant discovery and genotyping). The likely functional impact of variants was annotated using CADD scores and the Ensembl Variant Effect Predictor.

**Variant discovery power.** To evaluate our power to discover rare variants through low-pass sequencing, we examined 1,068 samples that were both sequenced and genotyped with the 4 genotyping arrays previously described. The four arrays provided us with an incomplete but high-quality catalog of low-frequency variants in these samples. We organized these variants by frequency and tabulated the fraction of variants that were rediscovered in our sequencing-based analysis for each frequency bin. Overall, we estimate that our sequencing effort discovered ~70% of the variants with frequency <0.5%, 98.8% of variants with frequency 0.5–5%, and >99% of variants with frequency >5% (Supplementary Table 2).

**Haplotyping and imputation.** Genotypes were phased with MACH software, using 30 iterations of the haplotyping Markov chain and 400 states per iteration. Imputation used Minimac software and a reference panel including haplotypes estimated by sequencing. To reduce the number of duplicated haplotypes, whenever a parent-offspring trio was sequenced, only the parental haplotypes were included in the imputation reference panel (resulting in 1,488 individuals for imputation). To reduce computational effort, we did not attempt to impute singleton variants. After imputation, we retained for association analysis only markers with an imputation quality (RSQ) >0.3 or >0.6 if the estimated MAF was 21% or <1%, respectively. For comparison, we repeated imputation using the 1000 Genomes Project Phase 3 haplotype set (using all 2,504 available samples; November 2014 release) and used RSQ >0.3 for all variants as a filter for imputation accuracy, as suggested in ref. 34.
Estimates of imputation accuracy. To further evaluate imputation accuracy, we carried out imputation using the Cardio-Metabochip, Immunochip and OmniExpress arrays as a scaffold and compared imputed genotypes with ExomeChip genotypes. This comparison excluded any markers that overlapped among the three scaffold arrays and the ExomeChip (Supplementary Fig. 4). To track the origin of the haplotypes used as templates during imputation, we interspersed dummy markers in the haplotypes, arbitrarily labeled with allele 0 for individuals recruited from the Lanusei valley and labeled with allele 0 for individuals recruited elsewhere in Sardinia.

Population structure analyses. To calculate $F_{ST}$ we used a random sampling of 200 unrelated individuals from the Lanusei valley and 200 individuals from the case-control control cohort study together with all POPRES European populations with sample sizes greater than 15. To obtain unrelated Sardinian individuals, we removed a random individual from each putative related pair until no pairs of individuals had an estimated proportion of identity-by-descent (IBD) sharing $\geq 0.05$ (as measured using PLINK on the basis of variants with MAF $>5\%$). We calculated Weir and Cockerham $F_{ST}$ values for all pairs of populations. Significance was assessed by 1,000 permutations of individual labels between a given pair of populations (Supplementary Fig. 2).

Principal-component analysis was performed using EIGENSTRAT version 5.0 after removing one SNP from each pair of SNPs with $r^2 > 0.85$ (in windows of 50 SNPs and steps of 5 SNPs) as well as SNPs in regions known to exhibit extended long-range LD65. We first considered a subset of 400 unrelated Sardinians along with all POPRES European populations. We then considered the full set of sequenced genomes and projected samples into an existing principal-component analysis coordinate space, one at a time (Supplementary Fig. 3). This analysis requires a small adjustment to the placement of each sample, which otherwise would be shifted toward the origin44. To address this, we devised a regression-based empirical correction scheme (D.O.d.V. and J.N., unpublished data). The approach uses a leave-one-out procedure to learn how the shift effect depends on the principal-component values and then applies this correction to all projected values. This procedure is not sensitive to the inclusion of related samples, and we were thus able to project the full Sardinian sample. To display levels of allele sharing between populations at different allele frequencies we used a previously described metric29,30.

Association testing. We searched for evidence of association using EPACTS59, a software that performs a linear mixed model adjusted with a genomic-based kinship matrix calculated using all quality-checked genotyped autosomal SNPs with MAF $>1\%$ (599,975 SNPs of the 890,542 total SNPs). The advantage of this model is that the kinship matrix encodes a wide range of sample structures, including both cryptic relatedness and population stratification. As a proof of appropriate adjustment for all confounders, the genomic control was 0.97, 0.99, 0.97, 1.01, 1.01, 1.01, 1 and 1 for the analysis of LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides, ADPN, hsCRP, IL-6, MCP-1 and ESR, respectively. Only additive effects for each allele were considered, and age, age$^2$ and sex were included as covariates in all analyses. Trait measures were normalized by quantile transformation, prior analyses. For the inflammatory traits, we also included smoking status and body mass index (BMI) as covariates13.

To identify sex-specific effects, we first performed GWAS analysis separately for males and females using the same transformation and same covariates (excluding sex) as in the primary GWAS. We then assessed the significance of observed differences by testing the heterogeneity of effect sizes with a $\chi^2$ test implemented in METAL49.

Rare variant analysis. We performed two region-based tests: CMC67 and VT$^a$s tests. Both tests were implemented in EPACTS (see URLs) to account for family relationships in our GWAS. To perform these rare variant tests, we used all nonsynonymous SNPs and variants altering splicing, with MAF $<5\%$. In each test, we assessed 10,000 regions and thus considered a Bonferroni threshold of $P < 5 \times 10^{-8}$ to declare significance.

Calculation of variance explained. The variance explained by the strongest associated SNPs was calculated for each trait as the difference in $R^2$-adjusted observed in the full and basic models, where the basic model only included phenotypic covariates (age, age$^2$ and sex for lipid level traits; age, age$^2$, sex, BMI and smoking status included for the inflammatory markers) and the full model also included all the independent SNPs associated with a specific trait. Variance for all available SNPs was calculated using GCTA software48, taking account of both closely and distantly related pairs of individuals69. The set of all available SNPs included all quality-checked SNPs after removing those that were monomorphic in the subset of phenotyped individuals (this set is also called the ‘accessible genome’).

Conditional analysis. To identify independent signals, we performed GWAS analysis for each trait by adding the leading SNPs found in the primary GWAS as covariates to the basic model. A SNP reaching the classical genome-wide significance threshold ($P < 5 \times 10^{-8}$) was considered a significant independent signal, with the sole exception of rs72658864, which did not reach the threshold but whose association was supported by previous reports.

Estimate of the genome-wide significance threshold in Sardinians. We defined a threshold for significance that applies to Sardinians when considering whole-genome sequencing data using empirical estimates (R package). We performed analyses in the SardiNIA cohort as well as in a cohort of 2,700 unrelated individuals from the Sardinian case-control study of multiple sclerosis and type 1 diabetes, who have been genotyped using OmniExpress and Immunochip arrays and imputed using the Sardinian reference panel. This additional cohort was used to ensure that there was no bias introduced into the estimation of the threshold by dealing with families in the SardiNIA study. The method consisted in simulating phenotypes under null and running single-marker association tests to calculate the threshold to maintain a family-wide error rate of 5%. Associations were performed for all the SNPs on chromosome 3, and the genome-wide significance threshold was then predicted assuming that the whole genome is approximately 15.6 times longer than chromosome 3 (ref. 70).

For the SardiNIA samples, we simulated 3 sets of 300 normally distributed phenotypes assuming 3 different heritabilities (20%, 40% and 70%) using Merlin (-simul option)71. We assumed no underlying quantitative trait loci (QTLs) among the genotyped and imputed variants. For the case-control study, we simulated 300 normally distributed phenotypes under the null hypothesis of no association. The results were highly comparable among all scenarios (Supplementary Table 6). To obtain a more accurate estimate, we increased the number of simulations up to 1,000 for all the phenotypes (except for the phenotype with 70% of heritability because it is not a typical scenario in GWAS). We then calculated the genome-wide significance thresholds for analyses that aim to test all variants and for those that evaluate only variants with MAF $>0.5\%$. Our estimates led to a significant threshold of $P < 6.9 \times 10^{-9}$ and of $P < 1.4 \times 10^{-8}$ for GWAS with all variants and with only variants with MAF $>0.5\%$, respectively.

Variant replication. We searched for replication of the 2 new signals associated with hsCRP in 7,689 individuals from 8 European cohorts (TwinsUK, FVG, VBI, HA, HP, ALSpac, INCipe1 and INCipe2)72–75. ESR, MCP-1 and ADPN values were not available in these samples. In TwinsUK and ALSpac, we analyzed genotypes from whole-genome sequencing data76, whereas for the FVG, VBI, HA, HP, INCipe1 and INCipe2 cohorts we used genotypes imputed from the 1000 Genomes Project Phase 1 sequencing panel. Specific details on each cohort are provided in the Supplementary Note. Association was evaluated by fitting a linear regression model that included age and sex as covariates, using as software GEMMA (TwinsUK, FVG, VBI, HA and HP) and SNPTEST (ALSPAC, INCipe1 and INCipe2) (see URLs). Normalization was not applied to the traits.

53. Burdick, J.T., Chen, W.-M., Abecasis, G.R. & Cheung, V.G. In silico method for inferring genotypes in pedigrees. Nat. Genet. 38, 1002–1004 (2006).
54. Voight, B.F. et al. The MetaboCHIP, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 8, e1002793 (2012).
55. Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* 14, 661–673 (2013).

56. Goldstein, J.I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* 28, 2543–2545 (2012).

57. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369 (2008).

58. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595 (2010).

59. Li, B. *et al.* QPLOT: a quality assessment tool for next generation sequencing data. *BioMed. Res. Int.* 2013, 865181 (2013).

60. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* 91, 839–848 (2012).

61. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834 (2010).

62. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838 (2010).

63. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* 38, 281–290 (2014).

64. Lee, S., Zou, F. & Wright, F.A. Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Stat.* 38, 3605–3629 (2010).

65. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354 (2010).

66. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191 (2010).

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

68. Price, A.L. *et al.* Pooled association tests for rare variants in exom-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838 (2010).

69. Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* 9, e1003520 (2013).

70. Esco, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur. J. Hum. Genet.* 21, 659–665 (2013).

71. Frigia, M. *et al.* Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. *PLoS ONE* 4, e7554 (2009).

72. Winkelmann, B.R. *et al.* Rationale and design of the LURIC study—a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics* 2, S1–S73 (2001).

73. Taylor, P.N. *et al.* Whole-genome sequence-based analysis of thyroid function. *Nat. Commun.* 6, 5681 (2015).