Published in final edited form as:
Nat Genet. 2015 November ; 47(11): 1272–1281. doi:10.1038/ng.3368.

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

Carlo Sidore[#1,2,3], **Fabio Busonero**[#1,2,4], **Andrea Maschio**[#1,2,4], **Eleonora Porcu**[#1,2,3], **Silvia Naitza**[#1], **Magdalena Zoledziewska**[1], **Antonella Mulas**[1,3], **Giorgio Pistis**[1,2,3], **Maristella Steri**[1], **Fabrice Danjou**[1], **Alan Kwong**[2], **Vicente Diego Ortega del Vecchyo**[5], **Charleston W. K. Chiang**[6], **Jennifer Bragg-Gresham**[2], **Maristella Pitzalis**[1], **Ramaiah Nagaraja**[7], **Brendan Tarrier**[4], **Christine Brennan**[4], **Sergio Uzzau**[8], **Christian Fuchsberger**[2], **Rossano Atzeni**[9], **Frederic Reinier**[9], **Riccardo Berutti**[3,9], **Jie Huang**[10], **Nicholas J Timpson**[11], **Daniela Toniolo**[12], **Paolo Gasparini**[13,14], **Giovanni Malerba**[15], **George Dedoussis**[16], **Eleftheria Zeggini**[10], **Nicole Soranzo**[10,17], **Chris Jones**[9], **Robert Lyons**[4], **Andrea Angius**[1,9], **Hyun M. Kang**[2], **John Novembre**[18], **Serena Sanna**[1,20], **David Schlessinger**[7,20], **Francesco Cucca**[1,3,20], and **Gonçalo R Abecasis**[2,20]

[1]Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, Cagliari, Italy

[2]Center for Statistical Genetics, Ann Arbor, University of Michigan, MI, USA

[3]Università degli Studi di Sassari, Sassari, Italy

[4]University of Michigan, DNA Sequencing Core, Ann Arbor, MI, USA

[5]Interdepartmental Program in Bioinformatics, University of California - Los Angeles, CA,USA

[6]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA

[7]Laboratory of Genetics, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA

[8]Porto Conte Ricerche srl, Tramariglio, Alghero, 07041 Italy

Correspondence should be addressed to: Francesco Cucca, fcucca@uniss.it, Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale di Ricerca, Monserrato, Cagliari, Italy; Goncalo R. Abecasis, goncalo@umich.edu, Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan, USA.

[20]These authors jointly supervised this work.

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., J.N. revised the manuscript and wrote specific sections of it. F.B., A.Ma., A.A., C.J. and R.L. supervised sequencing experiments. F.B., A.Ma., B.T. and C.B. performed sequencing experiments. C.S., E.P., G.P., M.S., F.D. and S.S. carried out genetic association analyses. C.S., A.K., R.A., F.R., R.B., C.J., R.L., H.M.K. were responsible for sequence data processing. C.S., E.P. and G.P. analyzed DNA sequence data. M.Z., A.Mu., F.B., S.U., R.N. carried out SNP array genotyping. M.Z. designed the validation strategy, and M.Z., F.B. and A.Mu. verified genotypes by Sanger sequencing and Taqman genotyping. C.S., J.B.G., M.P., C.F. and S.S. were responsible for selection of samples for sequencing, J.N., C.W.K.C. and V.D.O.V. performed the allele sharing, PCA and F_{ST} analyses, J.H., P.G., G.M., N.J.T., E.Z., D.T., G.D. and N.S. provided replication results. All authors reviewed and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

[9]Center for Advanced Studies, Research, and Development in Sardinia (CRS4), AGCT Program, Parco Scientifico e tecnologico della Sardegna, Pula, Italy

[10]Human Genetics, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1HH

[11]MRC Integrative Epidemiology Unit at the University of Bristol, University of Bristol, Bristol, United Kingdom

[12]Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milano, Italy

[13]DSM-University of Trieste and IRCCS-Burlo Garofolo Children Hospital (Trieste, Italy)

[14]Experimental Genetics Division, Sidra, (Doha, Qatar)

[15]Department of Life and Reproduction Sciences, University of Verona, Verona, Italy

[16]Harokopio University Athens

[17]Department of Haematology, University of Cambridge, Hills Rd, Cambridge CB2 0AH

[18]Department of Human Genetics, University of Chicago, IL, USA

[#] These authors contributed equally to this work.

Abstract

We report ~17.6M genetic variants from whole-genome sequencing of 2,120 Sardinians; 22% are absent from prior sequencing-based compilations and enriched for predicted functional consequence. Furthermore, ~76K 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. Fourteen signals, including two major new loci, were observed for lipid levels, and 19, including two novel loci, for inflammatory markers. New associations would be missed in analyses based on 1000 Genomes data, underlining the advantages of large-scale sequencing in this founder population.

INTRODUCTION

Studies of common genetic variants have provided entry points to analyse the mechanisms underlying many complex traits and diseases ([1–4]). Extension of these studies to the large reservoir of rare and population-specific variants could accelerate translation of genetic information into biological understanding, but has not thus far been systematically applied ([5,6]). 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 challenging ([5–7]). Studies of families and founder populations, where variants that are rare or absent elsewhere can occur at moderate frequencies, help overcome these limitations ([8]). 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 disorders ([9–11]). Besides

Europe PMC Funders Author Manuscripts

confirming signals from studies of common variants ([12,13]), our results reveal novel 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 study ([14]); 1,122 individuals, 52.8% of them female, average age 49.4], or from across the island and participating in case-control studies of Multiple Sclerosis ([15]) and Type 1 Diabetes ([16]) [referred to here as “island-wide sample”; 998 individuals, 48.5% of them female, average age 41.6]. Among these individuals, we sequenced 1,190 parent-offspring pairs distributed across 695 nuclear families in order to facilitate high quality estimation of haplotypes and genotypes ([17]). For each individual we generated an average of 10.7×10^9 mapped bases of high quality sequence (~4-fold coverage of the genome), corresponding to a total of 22.7×10^{12} bases across all individuals. We implemented quality control, alignment, variant calling and genotyping protocols that efficiently handled a sample of this size ([18], **URLs**) (**see Methods**).

In each sequenced individual, we identified an average of 3.4 million variants (17.6 million variants overall; Table 1). To assess quality, we sequenced two parents and a child to >65× coverage per individual. Comparing our initial low-coverage analysis with the results of deep sequencing for these individuals, we estimate an average genotyping error rate of <0.7% at heterozygous sites. As expected ([19]) 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 a frequency >0.5% in our sample and ~70% of variants with frequency <0.5% (Supplementary Table 2).

Among the 17.6M variants discovered, 172,988 (0.98%) overlap protein coding sequences ([20]) (Table 1). Of these variants 84,312 are non-synonymous coding changes; 2,504, essential splice-site altering; and 2,013, nonsense. Consistent with the hypothesis that natural selection makes variants with strong biological impact more likely to be rare and/or geographically restricted, we observe that 59% of non-synonymous, 53% of splice-altering, and 70% of nonsense variants have frequency <0.5% (compared to 48% of variants genome-wide). We also observe that 12% of non-synonymous, 22% of splice-altering and 22% of nonsense variants are absent from prior sequencing studies [compared to 22% of all variants, using dbSNP142 and the Exome Aggregation Consortium (see **URLs**) as surrogates for the results of prior studies ([21])].

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 frequency ([22,23]). The consequences of this genetic differentiation are a relatively large fraction of population-specific low-frequency variants and long haplotypes shared among present day carriers of those 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 ~5.0%) in a sample of ~2500 individuals from the United Kingdom 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 frequency ~5.0%) in our sample are present in Phase 1 of the 1000 Genomes Project ([25]). Overall, we estimate that 76,286 variants very rare (frequency <0.5%) or absent in the 1000 Genome Project Phase 3 reach frequencies >5% in our sample. We used a machine learning-based scoring algorithm to summarize the deleteriousness of each variant in a CADD score ([26]). 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 (p=0.02) (Supplementary Figure 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 of common variants ([27,28]) (Supplementary Figure 2 and 3). Whereas F_{ST} between non-Sardinian European populations in the POPRES reference sample averages 0.001 (range 0.000 – 0.004), F_{ST} between the island-wide sample of Sardinians and POPRES European populations averaged 0.006 (range 0.003 - 0.010), and the difference was even greater between the Lanusei valley and POPRES European populations (average 0.009, range 0.006 – 0.013) (Supplementary Figure 2). The geographical structure is even more evident when considering less frequent alleles: sharing between mainland populations and Sardinia is particularly depressed relative to sharing within mainland populations for rare sites (such at 1000 Genomes CEU and TSI) ([29,30]) (Figure 1). The patterns of differentiation are again clear in the long identical haplotypes surrounding rare f_2 variants (variants that are observed in exactly two chromosomes) ([25,31]) (Figure 2). Of note, both Sardinian samples show similar haplotype lengths flanking f_2 variants they share with populations outside Sardinia, consistent with a common ancient demography. The greater relative isolation of the two samples is evident when we examine the length of haplotypes flanking f_2 variants present within each sample. For variants shared between individuals in the valley, flanking haplotypes averaged 3,570 kb, dropping to 735 kb when first and second degree relatives were excluded. These haplotypes averaged 580 kb when shared by a valley resident and an individual elsewhere in Sardinia; ~382 kb when shared with an European sequenced in the 1000 Genomes Project Phase 3; and ~264 kb when shared with an individual elsewhere in the world in the full set of the 1000 Genome Project (Figure 2). These haplotype lengths differences decline around variants with higher frequencies, which typically have more ancient origins, consistent with a period of shared demography (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 study participants ([14]). All 6,602 individuals from the SardiNIA study were genotyped with four 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 (see above), genetic information obtained for any individual can be propagated ("imputed") to close relatives genotyped with the scaffold of markers ([32,33]). To increase the power of genetic association analyses and sample genetic diversity in the valley, we sequenced individuals distributed across different families (Supplementary Table 4). We then searched for shared chromosome stretches between the sequenced individuals and the remaining study participants, allowing us to impute both common and rare variants exceedingly well. Imputation accuracy, measured as the squared correlation between imputed and laboratory genotypes was $r^2 = 0.98$ for variants with frequency >5% and 0.89 for variants with frequency of 0.5 – 1.0% (Supplementary Figure 4). This improved markedly upon imputation results based on haplotypes shared between our samples and 1000 Genomes Project participants ([25]), who include individuals representing genetic diversity across Europe and elsewhere in the world ($r^2 = 0.92$ and 0.62 for variants with MAF >5% and 0.5 – 1.0%, respectively; Supplementary Figure 4). Shared stretches of chromosome used to fill in missing data within each SardiNIA individual originated in other individuals from the valley ~87% of the time, and also strongly correlated with the number of their grandparents born in the area ($r^2 = 0.67$; Supplementary Figure 5).

Impact on association of lipid and inflammatory markers

We focused on 4 blood lipid levels [low-density lipoprotein cholesterol (LDL-c), total cholesterol (TC), triglycerides (TG) and high-density lipoprotein cholesterol (HDL)] to assess how sequence information might reveal effects of population-specific and low frequency variation for 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 ~13.6 million (selected with high imputation quality; see **Methods**) ([34]). Overall, we identified fourteen independently associated variants distributed across eleven loci at the classical genome-wide significant threshold of 5×10^{-8} associated with lipid levels in analysis including all individuals or in sex-restricted analysis including only males or females (Table 2, Supplementary Figures 6 and 7). These include ten variants with moderate effect tagging signals in *LIPC, SORT1, PCSK9, CILP2, CEPT, APOA5* (one signal each), and *LPL, APOE* (two signals) -- loci that have been extensively described in prior GWAS and other association studies. Other signals at known loci were detected at lower association levels (Supplementary Table 5). To declare novel genome-wide signals we used a threshold of 6.9×10^{-9}, which was calculated by empirically estimating the number of independent tests in a Sardinian genome (see **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 them through a Sanger-sequencing based effort ([35]): V578A (frequency 0.5%) in the *LDLR* gene (Supplementary Table 5) is associated with LDL-c and total cholesterol and independent of the known variant rs73015013 (frequency of 14%, effect −5.2 mg/dl, $p=6.4\times10^{-8}$, r2<0.001). Here we report a novel association for triglyceride levels with a missense variant in *APOA5* (frequency 3% in Sardinia, effect −20.7 mg/dl, $p=1.2\times10^{-12}$) (Table 2). This variant, R282S, was genotyped and included in

the ExomeChip array after it was discovered in our sequencing effort, and to date it has been found on only two chromosomes in >30,000 Europeans characterized in the Exome Aggregation Consortium. Of note, this is the strongest variant modulating triglycerides levels in Sardinia -- explaining almost 1% of the phenotypic variance -- and is also independent of the known common variant at the locus, rs10750097 (frequency of 17%, effect +11.9 mg/dl, $p=4.6\times10^{-9}$, r2=0.002) (Table 2). These two examples illustrate co-existence in the same locus of population-specific low frequency variants along with previously detected and independently associated cosmopolitan common variants (Figure 3). The third genetic variant is the stop codon mutation Q40X in the *HBB* gene, better known as *beta(0)39* because the corresponding codon was numbered 39 prior to the last update on standard protein nomenclature. It illustrates how variants that are unusually frequent in Sardinia can provide insights about biology. In Sardinia, this mutation is the common cause of autosomal recessive beta-thalassemia ([36]). In our sample, in agreement with earlier epidemiological findings ([37,38]), the heterozygous state is associated with 13.9 mg/dl lower LDL-c levels ($p=1.2\times10^{-20}$) and 16.9 mg/dl lower total cholesterol levels ($p=1.2\times10^{-22}$). Of note, the analysis after 1000 Genomes Phase 3 imputation points only to an intergenic marker (rs76053862) 122 kb away, the second most associated SNP using the Sardinian reference panel, with a much lower association signal ($p = 1.4\times10^{-13}$) (Figure 3). Finally, two additional signals were observed for total cholesterol levels at SNP rs115048493 near genes *TMEM33* and *DCAF4L1* ($p=6.94\times10^{-9}$) and with HDL-c at SNP rs8092903 near *TGIF1* in females ($p=4.49\times10^{-8}$) (Table 2 and Supplementary Table 7), although the biological bases for these associations are presently unclear. Since these signals are below our adjusted genome-wide threshold of 6.9×10^{-9} these findings remain tentative.

We were interested to see whether 1000 Genomes and HapMap based analysis would also miss important loci for other traits. As 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 had found 16 variants associated with at least 1 of 4 inflammatory markers measured: Interleukin-6 (IL-6), erythrocyte sedimentation rate (ESR), monocyte chemotactic protein-1 (MCP-1) and high-sensitivity C-reactive protein (hsCRP) ([13]). A fifth inflammatory marker, adiponectin (ADPN), showed no significant association in our previous analyses (unpublished results). Nevertheless, with the extended spectrum of variants assessed here we identify another 7 variants associated with MCP-1, hsCRP, ESR or ADPN, at the classical 5×10^{-8} threshold, 5 in 4 previously undetected loci as well as 2 signals at coding variants in known loci (Table 3, Supplementary Figure 6, 7 and 8). Among the newly identified signals, 3 remained significant even with the more stringent threshold of 6.9×10^{-9}. Compared to analyses based on HapMap or 1000 Genomes imputation, we also identified more strongly associated lead variants at 3 known loci (*APOE, HBB* and *RHCE*). These may point to causative variants, as supported by biological evidence, eQTL data and ENCODE annotation (see following paragraphs and Table 3).

In detail, we found a striking novel signal associated with both hsCRP (rs183233091, $p=1.1\times10^{-28}$) and ESR (12:125406340, $p=4.4\times10^{-23}$) on chromosome 12, in a stretch of rare variants encompassing several genes (Figure 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 lead variants increase the levels of both inflammatory traits, with effects that appear to be stronger in males (Supplementary Table 8). The extended associated region spans 5.4 Mb and includes 22 non-coding variants with association p-value $<1\times10^{-15}$ (Supplementary Figure 9). The majority, to our knowledge, are Sardinian specific, as only 10 were found in either the 1000 Genomes Project Phase 3 or in the GoNL project databases([39]) (4 with MAF between 0.1% and 1%, and the other 6 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), while nominal association was detected in a subset of 3,505 Southern European individuals for the top variant ($P_{onetail}$ =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 extremely rare or absent outside Sardinia.

For hsCRP, we detected additional candidate signals. One near *PDGFRL*, a gene previously implicated in inflammatory/autoimmune processes ([40,41]) (Supplementary Figure 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 5×10^{-8} threshold, but not the more stringent threshold for novel findings: one for ADPN, at 13:108884835 near the gene *ABDH13* p=3.3×10^{-8}, and another for MCP-1, at rs76135610, p=1.8×10^{-8}, in a region encompassing the *CBLN1* and *N4BP1* genes, which is associated in females only (see Table 3 and Supplementary Figure 8).

We uncovered two novel independent variants for MCP-1 that cause non-conservative, likely functional, amino acid changes. R89C substitution (rs34599082) in *DARC* causes the FYB-weak phenotype of reduced antigen expression and less ability to bind chemokines ([42]), and M249K in the transmembrane domain of *CCR2* is expected to affect molecular interactions and thereby alter downstream signal transduction of bound ligand ([43]).

Finally, better leads were found at three known loci. For hsCRP the known association signal near the *APOE* gene was mapped to the known non-synonymous causal variant, C130R. That SNP has been associated with Alzheimer disease, and directly with CRP levels both by candidate gene studies and very recently by exome sequencing-based GWAS ([44,45]); it also coincides with the independent signal for LDL-c levels, linking lipid levels to inflammatory marker regulation. Two new lead variants were found for ESR. One again points to the Q40X mutation in *HBB* (Supplementary Figure 8), consistent with its effect on red cell counts (as shown above for LDL-c), which are in turn inversely correlated with ESR values. This association is thus relevant when interpreting ESR values in these individuals. Finally, a previously reported association on chromosome 1 in an intron of the *TMEM57* gene ([13,46]) is refined to intron 3 of the nearby *RHCE* gene. That gene encodes the Rh blood group antigens, and ESR levels are higher in Rh-positive than in Rh-negative healthy adults, making *RHCE* a plausible candidate (Table 3). The lead SNP at this locus alters several regulatory motifs (ENCODE annotation at UCSC genome browser, see **URLs**) and is

strongly correlated (r^2=0.80) with a nearby eQTL variant (rs11802413 in *TMEM57*) that affects expression of *TMEM57* as well as *RHCE* in liver ([47]).

We also performed gene-based rare variant tests using CMC and VT tests. Six loci passed the Bonferroni threshold of 5×10^{-6} for significance (see **Methods**), but after conditional analysis only two were not driven by nearby associations detected in our single-variant GWAS analysis. Particularly strong associations were observed for *STAB1* ($p=4.7\times10^{-10}$) and adiponectin levels, and another for *PTPRH* ($p=8.3\times10^{-7}$) and ESR levels. These signals, however, were not further investigated (Table 4, Supplementary Table 11), as those traits were not available in the replication cohorts.

All newly associated variants for both blood lipid levels and inflammatory markers were validated by Sanger sequencing (Supplementary Table 10, **Methods**). Using 1000 Genomes imputation, no other signals were identified and all these new signals were either misplaced (as in the Q40X signal, which pointed to other nearby variants) or completely missed (Figure 3 and 4, Supplementary Figure 8, and Supplementary Tables 12 and 13).

Further illustrating the high resolving power of the sequence-based association analyses, CADD assessment showed that all 5 novel genome-wide signals as well as the 2 new independent signals have the highest CADD scores in their regions compared to those in high or moderate LD (r^2>0.5), supporting their potential causative role in trait variation (Supplementary Table 14 and 15). By contrast, only 6 signals among 23 at known loci for the lipids and inflammatory markers – typically driven by common variants -- had top CADD scores, suggesting that the observation for 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 ([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, for which they explain almost all of observed trait heritability(Supplementary Table 16 and 17 and Supplementary Figure 10). 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 poorly imputed, or to 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 the 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 ([49]). For example, we found that the *beta(0)39* stop codon mutation causing autosomal recessive beta-thalassemia ([36]) accounts for a large fraction of LDL-c variability in Sardinia, second only to the *APOE* variants. The variant is known to be associated with enhanced erythropoiesis ([36], companion paper) – the

heterozygous carriers have red blood cell counts 23% greater on average ($p<10^{-300}$). This provides a likely explanation for decreased lipid levels in the 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 codon mutation reaches a frequency of 5.0% in our sample, it is not included in standard genotyping arrays and cannot be easily imputed from HapMap or 1000 Genomes because it is very rare outside Sardinia (1000 Genomes frequency <0.1%). Likewise, by cross-population exclusion mapping, we show that the novel 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 Figure 1). This suggests that part of the reservoir of variants that have drifted to higher frequency 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: the new signals we identified were 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/founder variants are not shared, as will often be the case, confirming the validity of results will require either accumulating additional samples in the population initially being 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 ([19],companion paper) and creates the bases for larger research and personalised medicine programs.

METHODS

Study Samples

To survey genetic variation across Sardinia, we selected individuals participating in the SardiNIA longitudinal study of aging ([14]) or in case-control studies of Multiple Sclerosis ([15]) and Type 1 Diabetes ([16]). All participants gave informed consent, with protocols approved by institutional review boards for the University of Cagliari, the National Institute on Aging, and the University of Michigan.

The SardiNIA project includes 6,921 individuals, representing >60% of the adult population of four villages in the Lanusei valley in Sardinia. Details of phenotype assessments for these

samples have been published previously ([13,14]). In particular, LDL-c levels were estimated using the Friedewald formula. Individuals with triglycerides >400 mg/dl or those taking lipid lowering medications were excluded from the LDL-c, and those on medication were also excluded from analyses of other lipids. Summary statistics for individuals considered for 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 ([32,51]). We used known family relationships among SardiNIA study participants and the ExomePicks program (**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 one affected child and two 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 favoured sequencing of parent-offspring trios to improve the accuracy of resulting haplotypes ([17]). Part of the sequencing data used in this study are available through dbGap, under “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 provide improved coverage of regions associated with cardiovascular and metabolic disease - CardioMetaboChip ([52]), immune disorders - ImmunoChip ([53]), and coding variation – ExomeChip, (**URLs**). Genotyping was carried out according to manufacturer protocols at the SardiNIA Project Laboratory (Lanusei, Italy), at the Technological Centre - Porto Conte Ricerche (Alghero, Italy) and at the National Institute on Aging Intramural Research Program Laboratory (Baltimore, MD). Genotypes were called using GenomeStudio (version 1.9.4) and refined using Zcall (version 3) ([54]). We applied standard per sample quality control filters to remove samples with low call rates or where reported relationships and/or sex disagreed with genetic data ([55]). We also applied per marker quality control filters to remove markers with low call rates, deviations from Hardy-Weinberg equilibrium, excess discordance among duplicates or identical twin genotypes, excess Mendelian inconsistencies or MAF=zero. Altogether, unique 890,542 autosomal markers and 16,325 X-linked markers were genotyped across SardiNIA study samples. Among the autosomal QCed markers, 809,193 are array specific (60,966 from ExomeChip, 112,717 from ImmunoChip, 100,554 from MetaboChip and 534,956 from OmniExpress) and 972 SNPs were typed in all the 4 arrays. The remaining 80,377 SNPs were typed in 2 or 3 arrays. For 870,108,399 genotypes assayed in >1 array, genotype concordance rate was

>99.99%. Our analyses include the 6,602 individuals that were successfully genotyped with all four arrays.

Sequencing

Sequence data were generated at the Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna (CRS4) and at the University of Michigan Medical School Core Sequencing Lab. Libraries were generated from 3-5 μg of genomic DNA using sample prep kits from Illumina and New England Biolabs. Paired-end sequence reads (typically, 100 to 120-bp in length) were generated with Illumina Genome Analyzer IIx, Illumina HiSeq 2000 and Illumina HiSeq 2500 instruments. Samples were sequenced to an average depth of 4.16×. A single nuclear family (two parents and one child) was sequenced to average depth >65× per individual to facilitate assessment of genotyping error rates.

Reads were aligned to the human reference genome (GRCh37 assembly with decoy sequences, as available in the 1000 Genomes Project ftp site, **URLs**) using BWA-0.5.9 ([56]), trimming read tails with average base quality <15. After alignment, base qualities were recalibrated and duplicate reads were flagged and excluded from analysis. We reviewed summary metrics generated using QPLOT([57]) and verifyBamId ([58]) for each aligned sample, to remove samples with low sequencing depth, poor coverage of regions with high or low GC content, or evidence for sample contamination.

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 NHLBI Exome Sequencing Project. The approach examines all samples jointly to identify an initial variant list, improving our 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 reference and alternate alleles at heterozygous sites, information on the evidence for alternate alleles by strand and read position, excess of 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, 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 ([26]) and Ensembl Variant Effect Predictor ([20]).

Europe PMC Funders Author Manuscripts

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 4 arrays provided us with an incomplete but high quality catalogue 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 using MACH software ([59]), using 30 iterations of the haplotyping Markov chain and 400 states per iteration. Imputation used minimac software ([60]) and a reference panel including haplotypes estimated by sequencing. To reduce the number of duplicated haplotypes, whenever a parent-offspring trio was sequenced, only 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 only markers with an imputation quality (RSQR) >0.3 or >0.6 if the estimated MAF was >=1% or <1% respectively([34]). For comparison, we repeated imputation using the 1000 Genomes Project Phase 3 haplotype set (using all 2,504 available samples, from November 2014 release) and used RSQR >0.3 for all variants as a filter for imputation accuracy, as suggested by ([34]). This strategy led to 13.6 million and 12.7 million markers useful for analyses on the Sardinian-based and 1000 Genomes-based datasets, respectively.

Estimates of Imputation Accuracy

To further evaluate imputation accuracy, we carried out imputation using CardioMetaboChip, ImmunoChip and OmniExpress as a scaffold, and compared imputed genotypes with ExomeChip genotypes. This comparison excluded any markers that overlap between the 3 scaffold arrays and the ExomeChip (Supplementary Figure 4). To track the origin of haplotypes used as templates during imputation, we interspersed dummy markers in the haplotypes, arbitrarily labelled with allele ‘1’ for individuals recruited from the Lanusei valley and labelled 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 from the case-control control cohort study, and all POPRES European populations with sample sizes greater than 15. To obtain unrelated Sardinian individuals we removed a random individual from each pair of putative relateds until no pairs of individuals had an estimated proportion of IBD sharing ≥0.05 (as measured using PLINK based on variants with MAF>5%). We calculated the Weir & Cockerham F_{ST} values between all pairs of populations. Significance is assessed by 1000 permutations of individual labels between a given pair of populations (Supplementary Figure 2).PCA analysis was performed using EIGENSTRAT version 5.0 after removing one SNP of each pair of SNPs with r^2 ≥0.8 (in

windows of 50 SNPs and steps of 5 SNPs) as well as SNPs in regions of known to exhibit extended long-range LD ([61]). 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 PCA coordinate space, one a time (Supplementary Figure 3). This analysis requires a small adjustment to the placement of each sample, which otherwise would be shifted towards the origin ([62]). To address this, we devised a regression-based empirical correction scheme (J. Novembre and colleagues, unpublished). The approach uses a leave-one-out procedure to learn how the shift effect depends on the PC values, and then applies this correction to all projected values. This procedure is not sensitive to the inclusion of related and thus we are able to project the full Sardinian sample. To display levels of allele sharing between populations at different allele frequencies we used a metric previously described([29,30]).

Association Testing

We searched for evidence of association using EPACTS ([63]), 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 out of the 890,542). The advantage of this model is that the kinship matrix encodes a wide range of sample structures, including both cryptic relatedness than population stratification. As a proof of appropriate adjustment of all confounders, the genomic control was 0.97, 0.99, 0.97, 1.01, 1.01, 1, 1.01, 1 and 1 for LDL-c, HDL-c, TC, TG, ADPN, hsCRP, IL-6, MCP-1 and ESR respectively. Only additive effects of each allele were considered and age, age-squared and sex were included as covariates in all analyses. Traits were normalized with quantile transformation, prior analyses. For the inflammatory traits, we also included smoke and BMI as covariates ([13]).

To identify sex-specific effects, we firstly performed GWAS analysis separately for males and females using the same transformation and same covariates (excluding gender) as in the primary GWAS. We then assessed significance to observed differences by testing heterogeneity of effect sizes with a chi-square test implemented in METAL ([64]).

Rare variant analysis

We performed two regional-based tests: the Combined Multivariate and Collapsing (CMC) ([65]) and the variable thresholds method (VT) ([66]). Both tests were implemented in EPACTS (see **URLs**) to account for familiar relationships in our GWAS. To perform these rare variants tests we used all non-synonymous SNPs and variants altering splicing, with MAF <5%. In each test, we assessed 10,000 regions and thus considered a Bonferroni threshold of 5×10^{-6} to declare significance.

Calculation of variance explained

The variance explained by the strongest associated SNPs was calculated for each trait as the difference of R^2-adjusted observed in the full and the basic model, where the basic model only includes phenotypic covariates (age, age^2 and sex for lipid levels traits, age, age^2, sex, BMI and smoke for the inflammatory markers) and the full model also includes all the independent SNPs associated with a specific trait. Variance for all available SNPs was

calculated using GCTA software ([48]) taking account of both closely and distantly related pairs of individuals ([67]). The set of all available SNPs included all quality checked SNPs after removing those which were monomorphic in the subset of phenotyped individuals (this set is also called as “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 for rs72658864 which did not reach the threshold but was supported by previous reports.

Estimate of 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 available at cran.r-project.org)([68]). 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 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 consists in simulating phenotypes under the 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.

For the SardiNIA samples we simulated three sets of 300 normally distributed phenotypes assuming three different heritability (20%, 40% and 70%) using Merlin (--simul option)([69]). We assumed no underlying QTLs among the genotyped and imputed variants. For the CaseControl study, we simulated 300 normally distributed phenotypes under the null hypothesis of no association. 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 6.9×10^{-09} and of 1.4×10^{-08} for GWAS with all variants and with only variants with MAF>0.5%, respectively.

Variant Replication

We searched for replication of the two novel signals associated with hsCRP in 7,689 individuals from 8 European cohorts (TwinsUK, FVG, VBI, HA, HP, ALSPAC, INCIPE1, INCIPE2)([70–73]); ESR, MCP-1 and ADPN values were not available in those samples. In TwinsUK and ALSPAC we analysed genotypes from whole-genome sequence data([74]), while for FVG, VBI, HA, HP, INCIPE1 and INCIPE2 cohorts we used genotypes imputed

using the 1000 Genomes Phase I sequencing panel. Specific details on each cohort are provided in Supplementary Note. Association was evaluated by fitting a linear regression model that included age and gender as covariates, using as software GEMMA (TwinsUK, FVG, VBI, HA, HP) and SNPTEST (ALSPAC, INCIPE1, INCIPE2)(see **URLs**). Normalization was not applied to the trait.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

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 HG005581, HG005552, HG006513, HG007022, and HG007089; by National Heart Lung and Blood Institute grant HL117626; by the Intramural Research Program of the NIH, National Institute on Aging, with contracts N01-AG-1-2109 and HHSN271201100005C; by Sardinian Autonomous Region (L.R. no. 7/2009) grant cRP3-154; by PB05 InterOmics MIUR Flagship Project; by grant FaReBio2011 “Farmaci e Reti Biotecnologiche di Qualità”; by NIH NRSA postdoctoral fellowship (F32GM106656) to C.W. K. C.; and by UC MEXUS / CONOCYT fellowship to V.D.O.V. The replication cohorts acknowledge the use of data generated by the UK10K Consortium, supported by the Wellcome Trust award WT091310. The UK10K research was specifically funded by a Wellcome Trust award: 10,000 UK genome sequences: accessing the role of rare genetic variants in health and disease (WT091310/C/10/Z). Nicole Soranzo’s research is supported by the Wellcome Trust (Grant Codes WT098051 and WT091310), the EU FP7 (EPIGENESYS Grant Code 257082 and BLUEPRINT Grant Code HEALTH-F5-2011-282510) and the NIHR BRC. ING-FVG cohort was supported by grant: Ministero della Salute - Ricerca Finalizzata PE-2011-02347500 (to PG); ING-VB study thanks the inhabitants of the Val Borbera for participating in the study, Michela Traglia, Cinzia Sala and Corrado Masciullo for data management, and funding sources Fondazione Cariplo (Italy), Ministry of Health, Ricerca Finalizzata (Italy) 2008 and 2011-2012, Public Health Genomics Project 2010. HELIC cohorts are thankful to the residents of the Pomak villages and of the Mylopotamos villages for participating, and funding sources Wellcome Trust (098051) and the European Research Council (ERC-2011-StG 280559-SEPI).

URLs

Exome Aggregation Consortium 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: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

UCSC 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

Pritchard’s lab eQTL browser: http://eqtl.uchicago.edu/

Exome Picks: http://genome.sph.umich.edu/wiki/ExomePicks

Europe PMC Funders Author Manuscripts

ExomeChip: http://genome.sph.umich.edu/wiki/Exome_Chip_Design

1000G Data repository: ftp://ftp.1000genomes.ebi.ac.uk

REFERENCES

1. Parkes M, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet. 2007; 39:830–832. [PubMed: 17554261]
2. Willer CJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat. Genet. 2009; 41:25–34. [PubMed: 19079261]
3. 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. U. S. A. 2010; 107:7401–7406. [PubMed: 20385819]
4. Do R, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nat. Genet. 2013; 45:1345–1352. [PubMed: 24097064]
5. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum. Mol. Genet. 2012; 21:R1–9. [PubMed: 22983955]
6. Zuk O, et al. Searching for missing heritability: Designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 2014; 111:E455–E464. [PubMed: 24443550]
7. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:3871–3876. [PubMed: 19202052]
8. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat. Rev. Genet. 2000; 1:182–190. [PubMed: 11252747]
9. Clarke R, et al. Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men. Arch. Intern. Med. 2007; 167:1373–1378. [PubMed: 17620530]
10. Pai JK, et al. Inflammatory markers and the risk of coronary heart disease in men and women. N. Engl. J. Med. 2004; 351:2599–2610. [PubMed: 15602020]
11. Orru V, et al. Genetic variants regulating immune cell levels in health and disease. Cell. 2013; 155:242–56. [PubMed: 24074872]
12. Global Lipids Genetics Consortium. et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 2013; 45:1274–1283. [PubMed: 24097068]
13. 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. 2012; 8:e1002480. [PubMed: 22291609]
14. Pilia G, et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. 2006; 2:e132. [PubMed: 16934002]
15. Sanna S, et al. Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. Nat. Genet. 2010; 42:495–497. [PubMed: 20453840]
16. Zoledziewska M, et al. Variation within the CLEC16A gene shows consistent disease association with both multiple sclerosis and type 1 diabetes in Sardinia. Genes Immun. 2009; 10:15–17. [PubMed: 18946483]
17. Chen W, et al. Genotype calling and haplotyping in parent-offspring trios. Genome Res. 2013; 23:142–151. [PubMed: 23064751]
18. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. Genome Res. 2015 doi:10.1101/gr.176552.114.
19. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011; 21:940–951. [PubMed: 21460063]
20. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinforma. Oxf. Engl. 2010; 26:2069–2070. [PubMed: 20562413]
21. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]

22. Francalacci P, et al. Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. Am. J. Phys. Anthropol. 2003; 121:270–279. [PubMed: 12772214]
23. Francalacci P, et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science. 2013; 341:565–569. [PubMed: 23908240]
24. Zavattari P, et al. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. Hum. Mol. Genet. 2000; 9:2947–2957. [PubMed: 11115838]
25. 1000 Genomes Project Consortium. et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
26. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 2014; 46:310–315. [PubMed: 24487276]
27. Novembre J, et al. Genes mirror geography within Europe. Nature. 2008; 456:98–101. [PubMed: 18758442]
28. Nelson MR, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. 2008; 83:347–358. [PubMed: 18760391]
29. Gravel S, et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. 2011; 108:11983–11988. [PubMed: 21730125]
30. Nelson MR, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337:100–104. [PubMed: 22604722]
31. Mathieson I, McVean G. Demography and the age of rare variants. PLoS Genet. 2014; 10:e1004528. [PubMed: 25101869]
32. Chen W-M, Abecasis GR. Family-based association tests for genomewide association scans. Am. J. Hum. Genet. 2007; 81:913–926. [PubMed: 17924335]
33. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu. Rev. Genomics Hum. Genet. 2009; 10:387–406. [PubMed: 19715440]
34. Pistis G, et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur. J. Hum. Genet. 2014 doi:10.1038/ejhg.2014.216.
35. Sanna S, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. 2011; 7:e1002198. [PubMed: 21829380]
36. Cao A, Galanello R. Beta-thalassemia. Genet. Med. 2010; 12:61–76. [PubMed: 20098328]
37. Maioli M, et al. Plasma lipoprotein composition, apolipoprotein(a) concentration and isoforms in beta-thalassemia. Atherosclerosis. 1997; 131:127–133. [PubMed: 9180253]
38. Maioli M, et al. Plasma lipids in beta-thalassemia minor. Atherosclerosis. 1989; 75:245–248. [PubMed: 2712866]
39. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 2014; 46:818–825. [PubMed: 24974849]
40. Hou S, et al. Genetic variant on PDGFRL associated with Behçet disease in Chinese Han populations. Hum. Mutat. 2013; 34:74–78. [PubMed: 22926996]
41. Xu M, et al. An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics. 2008; 9(Suppl 1):S12. [PubMed: 18366601]
42. Tournamille C, et al. Arg89Cys substitution results in very low membrane expression of the Duffy antigen/receptor for chemokines in Fy(x) individuals. Blood. 1998; 92:2147–2156. [PubMed: 9731074]
43. Shi X-F, et al. Structural analysis of human CCR2b and primate CCR2b by molecular modeling and molecular dynamics simulation. J. Mol. Model. 2002; 8:217–222. [PubMed: 12192431]
44. Schick UM, et al. Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum. Mol. Genet. 2014 doi:10.1093/hmg/ddu450.

45. Golledge J, et al. Apolipoprotein E genotype is associated with serum C-reactive protein but not abdominal aortic aneurysm. Atherosclerosis. 2010; 209:487–491. [PubMed: 19818961]
46. Kullo IJ, et al. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. Am. J. Hum. Genet. 2011; 89:131–138. [PubMed: 21700265]
47. Schadt EE, et al. Mapping the Genetic Architecture of Gene Expression in Human Liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]
48. Yang J, Lee SH, Goddard ME, Visscher PM. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol. Biol. Clifton NJ. 2013; 1019:215–236.
49. Moltke I, et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature. 2014; 512:190–193. [PubMed: 25043022]

Methods references

50. Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–7. [PubMed: 20634204]
51. Burdick JT, Chen W-M, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. Nat. Genet. 2006; 38:1002–1004. [PubMed: 16921375]
52. Voight BF, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012; 8:e1002793. [PubMed: 22876189]
53. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat. Rev. Genet. 2013; 14:661–673. [PubMed: 23917628]
54. Goldstein JI, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinforma. Oxf. Engl. 2012; 28:2543–2545. [PubMed: 22843986]
55. McCarthy MI, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 2008; 9:356–369. [PubMed: 18398418]
56. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 2010; 26:589–595. [PubMed: 20080505]
57. Li B, et al. QPLOT: a quality assessment tool for next generation sequencing data. BioMed Res. Int. 2013; 2013:865181. [PubMed: 24319692]
58. Jun G, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. 2012; 91:839–848. [PubMed: 23103226]
59. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 2010; 34:816–834. [PubMed: 21058334]
60. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 2012; 44:955–959. [PubMed: 22820512]
61. Price AL, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 2008; 4:e236. [PubMed: 18208327]
62. Lee S, Zou F, Wright FA. CONVERGENCE AND PREDICTION OF PRINCIPAL COMPONENT SCORES IN HIGH-DIMENSIONAL SETTINGS. Ann. Stat. 2010; 38:3605–3629. [PubMed: 21442047]
63. Kang HM, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2009; 42:348–54. [PubMed: 20208533]
64. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma. Oxf. Engl. 2010; 26:2190–2191. [PubMed: 20616382]
65. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 2008; 83:311–321. [PubMed: 18691683]
66. Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 2010; 86:832–838. [PubMed: 20471002]

67. Zaitlen N, et al. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. PLoS Genet. 2013; 9:e1003520. [PubMed: 23737753]

68. Xu C, et al. Estimating genome-wide significance for whole-genome sequencing studies. Genet. Epidemiol. 2014; 38:281–290. [PubMed: 24676807]

69. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 2002; 30:97–101. [PubMed: 11731797]

70. Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort Profile: TwinsUK and healthy ageing twin study. Int. J. Epidemiol. 2013; 42:76–85. [PubMed: 22253318]

71. Esko T, et al. Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. Eur. J. Hum. Genet. EJHG. 2013; 21:659–665. [PubMed: 23249956]

72. Traglia M, et al. Heritability and demographic analyses in the large isolated population of Val Borbera suggest advantages in mapping complex traits genes. PloS One. 2009; 4:e7554. [PubMed: 19847309]

73. Winkelmann BR, et al. Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. Pharmacogenomics. 2001; 2:S1–73. [PubMed: 11258203]

74. Taylor PN, et al. Whole-genome sequence-based analysis of thyroid function. Nat. Commun. 2015; 6:5681. [PubMed: 25743335]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

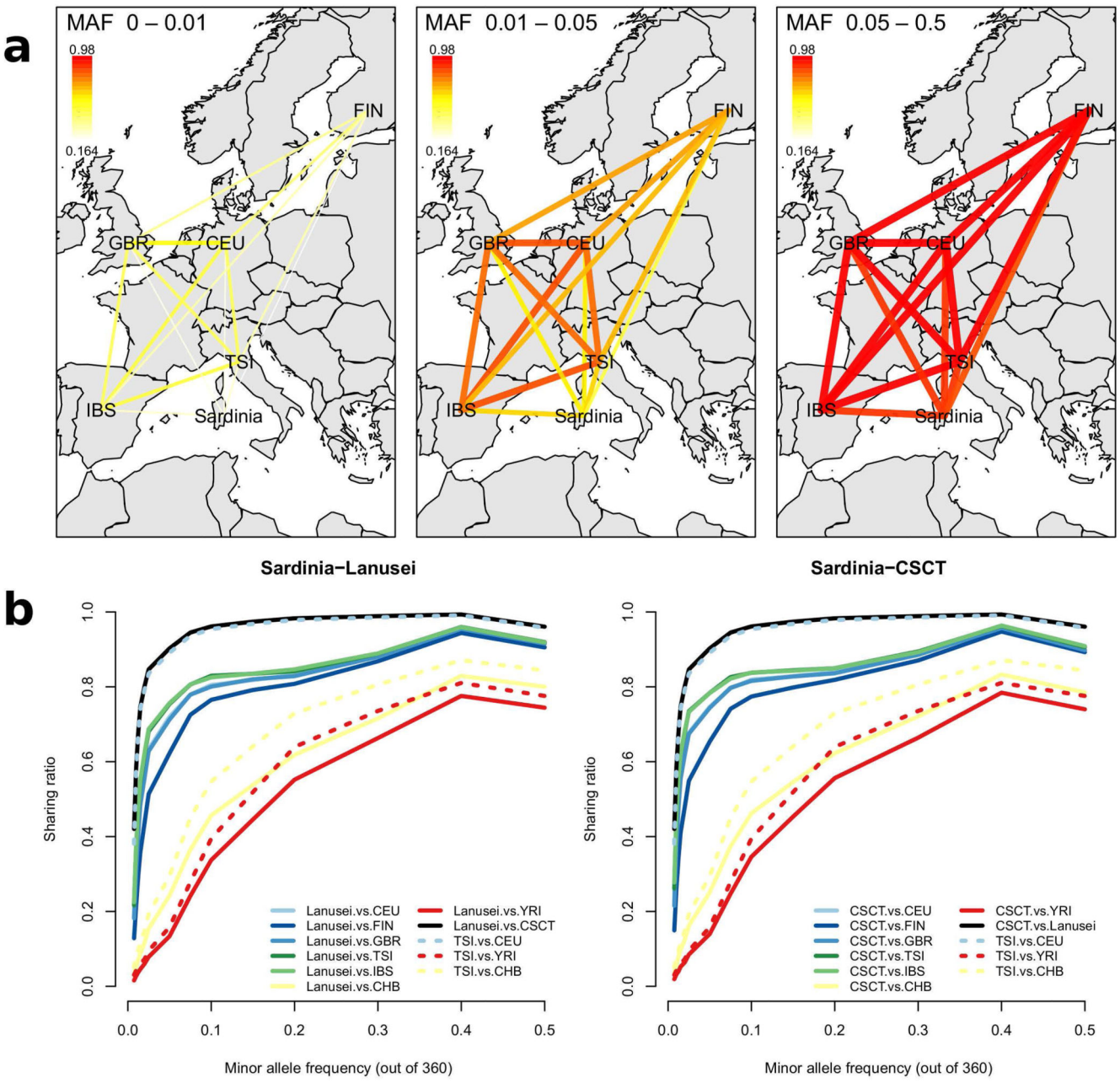

Figure 1. Geographical differentiation based on common and rare sites

The figure show allele sharing among the Sardinian and the 1000 Genomes European populations. In panel a) differentiation is represented for three different frequency intervals over the geographic map of Europe. The thickness and the color of the lines connecting the dots are proportional to the allele sharing statistic as indicated in the color map. In panel b) we instead represent the relationship between the frequency (X axis) and the sharing ratio (on the Y axis) for different 1000 Genomes Project populations (continuous lines). Results are plotted separately for the Lanusei valley sample (left panel) and the case control samples (right panel). The dotted line are used as comparison to show the sharing ratio between the TSI and other 1000 Genomes Project populations.

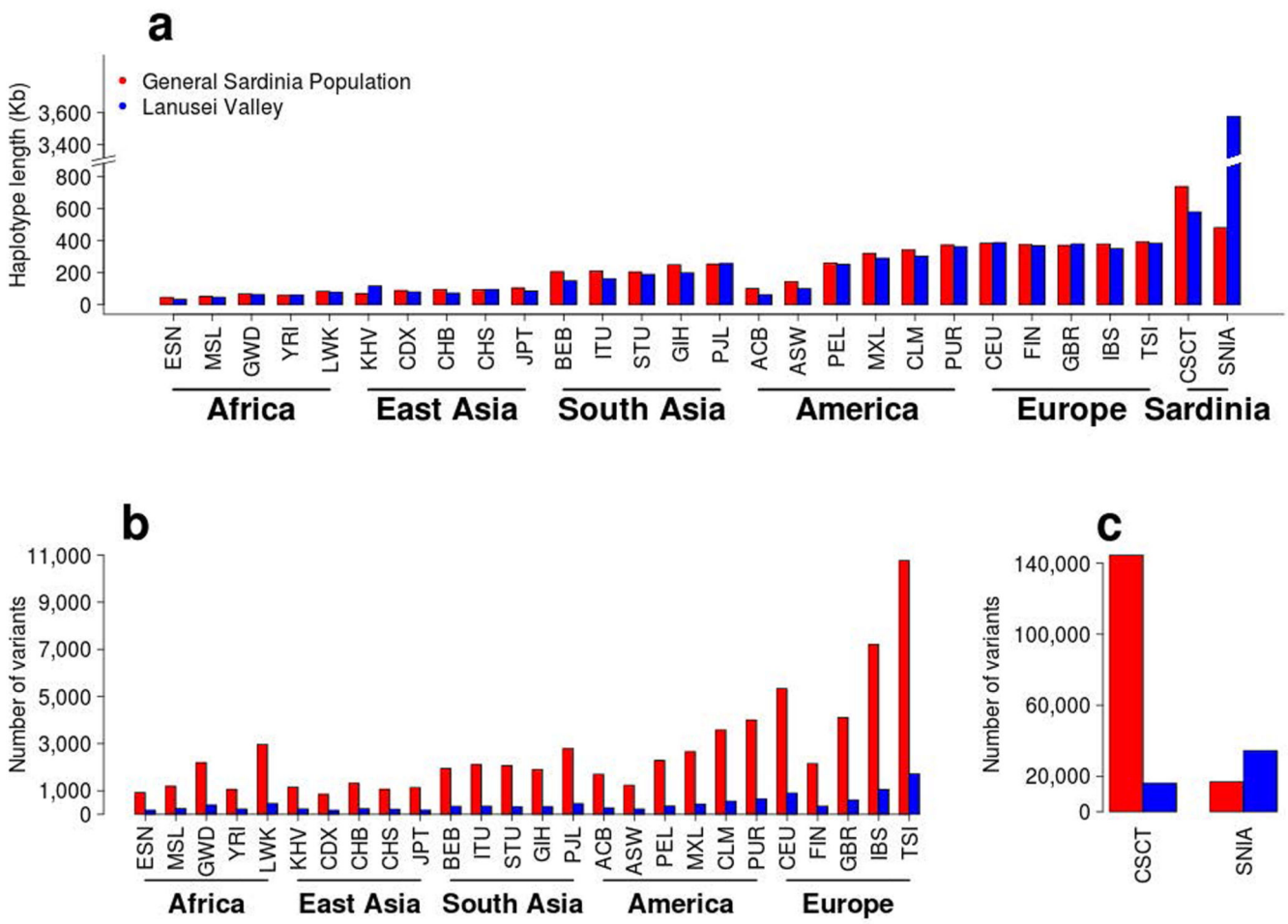

Figure 2. Length of shared haplotypes surrounding f_2 variants within Sardinians and populations in 1000 Genomes

Length of shared haplotypes surrounding f_2 variants shared between one of our sequenced individuals and one of 100 randomly selected individuals sampled from our study or from a particular 1000 Genomes Project population. Panel a) shows the length of these shared haplotypes, in kilobases, in comparisons between Sardinia and several 1000 Genomes Project populations. Panel b) shows the number of f_2 haplotypes in each comparison. Panel c) shows the number of f_2 haplotypes in comparisons within Sardinia (note the wider Y-axis range).

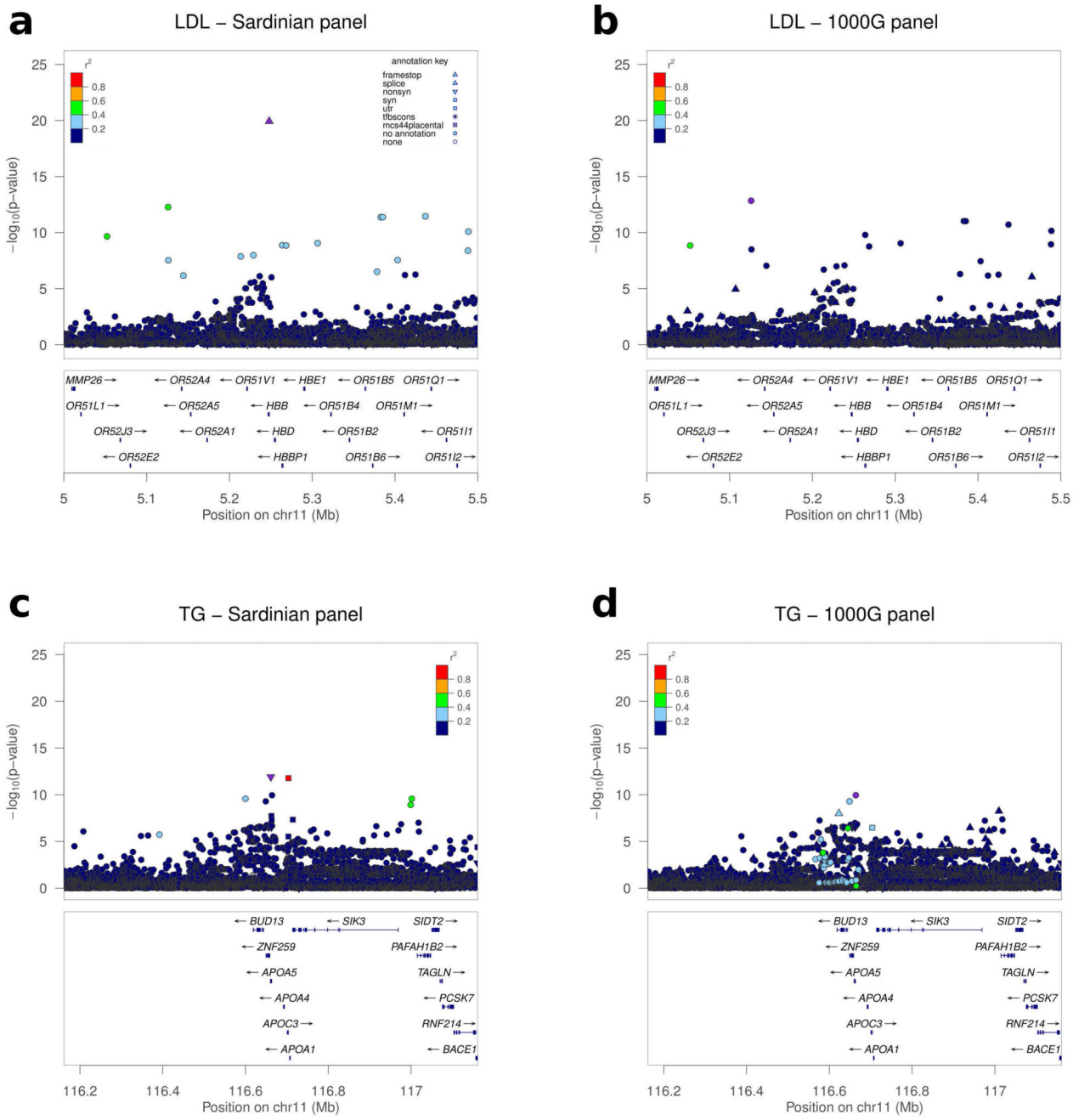

Figure 3. Regional association plots for novel lipids loci

Regional association plots at the *HBB* locus for LDL-c, and at *APOA5* for triglycerides for imputation performed using the Sardinian (panels a and c) and 1000 Genomes (panels b and d) reference panels, respectively. At each locus, we plotted the association strength (Y axis shows the –log 10 pvalue) versus the genomic positions (on the hg19/GRCh37 genomic build) around the most significant SNP, which is indicated with a purple dot. Other SNPs in the region are color-coded to reflect their LD with the top SNP as in the inset (taken from pairwise r^2 values calculated on Sardinian and 1000 Genomes haplotypes for left and right

Europe PMC Funders Author Manuscripts

panels, respectively). Symbols reflect genomic functional annotation, as indicated in the inner box of panel A. Genes and the position of exons, as well as the direction of transcription, are noted in lower boxes. This plot was drawn using the standalone version of the LocusZoom package ([50]).

Europe PMC Funders Author Manuscripts

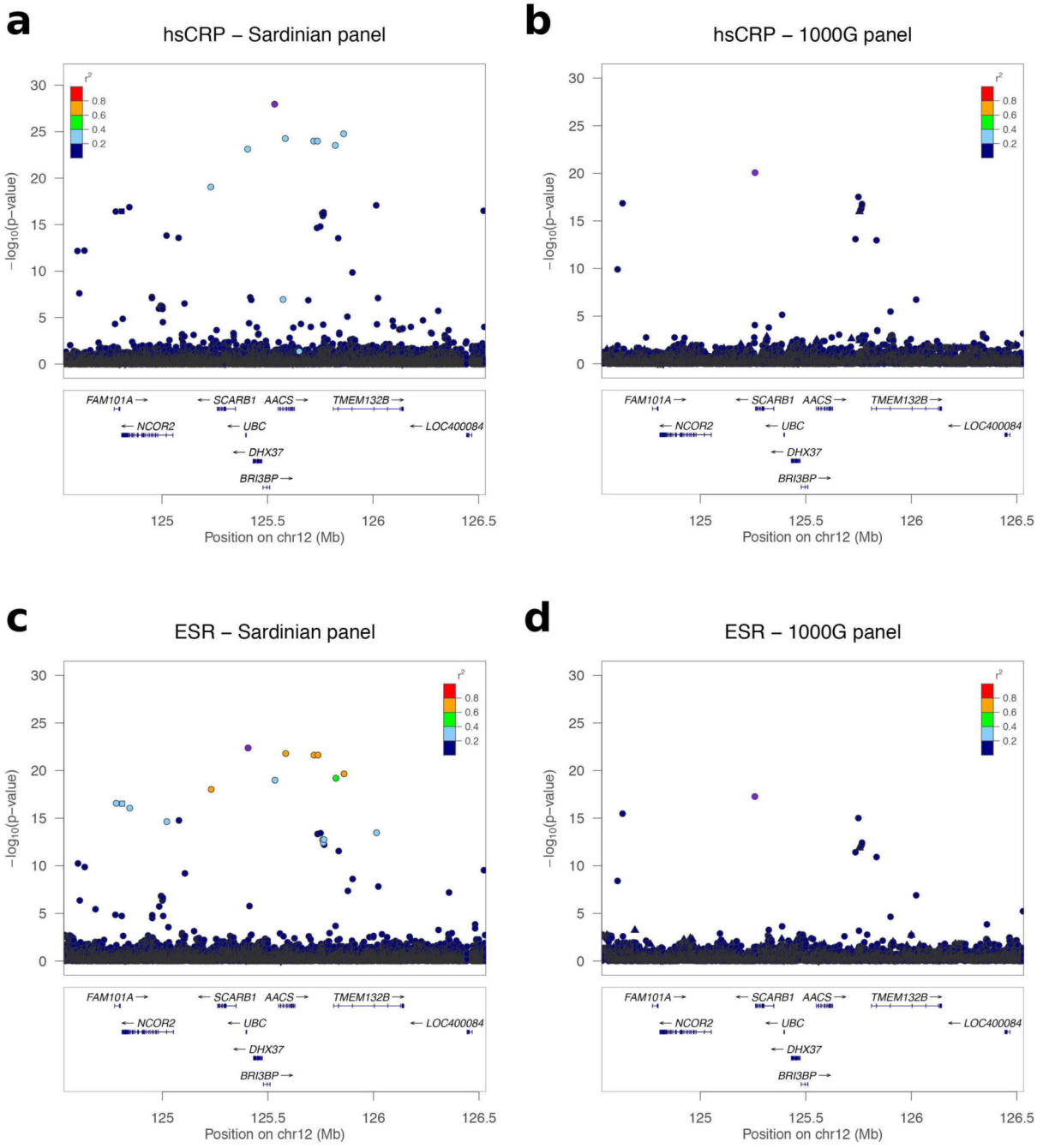

Figure 4. Regional association plot at chromosome 12 for hSCRP and ESR

Regional association plots at the chromosome 12 locus for hsCRP and for ESR, using the Sardinian (panels a and c) and 1000 Genomes (panels b and d) reference panels for imputation, respectively. For the plot style, see Figure 3 legend.

Table 1

Summary of Discovered Variants

The table provides an overview of the sequencing data, including summary statistics on data generated, a breakdown by frequency and biological function of all variants discovered and their novelty rate based on public databases. Finally, we show the distribution of variants discovered per each sequenced individual.

Data Generation							
Total Mapped Bases	***** 22,684 Gb *****						
Average Depth	***** 4.16× *****						
			Coding Variation				
	Genome	**Regulatory**	**Silent**	**Splice**	**Essential Splice**	**Missense**	**Nonsense**
Total Variation							
No. of Variants	17.6M	1,596,737	63,062	21,097	2,504	84,312	2,013
Novelty rate vs dbSNP 135	31.6%	31.7%	24.0%	31.8%	36.2%	34.8%	48.7%
Novelty rate vs dbSNP 142	21.7%	21.6%	15.2%	19.1%	26.5%	22.6%	34.8%
Novelty rate vs dbSNP142 and Exome Aggregation	21.6%	21.5%	7.0%	14.2%	21.8%	11.8%	21.6%
Total Variation by Frequency							
Common (MAF > 5%)	31.8%	31.2%	29.1%	28.7%	26.8%	20.7%	14.5%
Low Frequency (MAF 0.5-5%)	19.8%	21.2%	21.5%	20.7%	20.1%	19.8%	15.8%
Rare (MAF < 0.5%)	47.7%	47.5%	49.4%	50.6%	53.2%	59.5%	69.7%
Singletons	9.0%	8.8%	9.2%	9.6%	9.8%	12.3%	17.9%
Variation per individual							
5th Percentile	3,332,299	293,928	10,619	3,331	361	10,738	158
Average	3,359,655	293,928	10,778	3,396	380	10,920	172
95th Percentile	3,383,736	298,766	10,934	3,465	400	11,100	186

Europe PMC Funders Author Manuscripts

Table 2

Summary of Lipid Association Results

The table lists association signals that reach p <5×10^{-8} for association with lipid levels in our study. At each novel locus, we indicated the genes likely to be modulated by the lead SNP, the location of the lead variant (human genome build GRCh37), the variant identifier rs#, the nearest gene, the effect and other allele, the frequency of the effect allele, the effect size in standard deviation units and the standard error, the pvalue, the proportion of variance explained by the allele (R2%), the imputation accuracy (RSQR) , the functional consequence of the variant and the r2 with hits previously identified in ([12]). When reporting a second signal within a locus, we first controlled for association with the local peak variant, as indicated by an asterisk (*, **) in the corresponding rows. Novel signals are shown in bold.

Candidate Gene	Chr:position	rs name	Effect Allele / Other	Freq	Effect (StdErr)	pvalue	R2(%)	RSQR	Variant Consequence	r^2 with previous hit
LDL										
PCSK9	1:55505647	rs11591147	T/G	0.038	−0.406(0.053)	1.73×10^{-14}	1.0	Genotyped	Missense, R46L	Same SNP
SORT1	1:109821307	rs583104	G/T	0.180	0.156(0.027)	1.87×10^{-08}	0.5	Genotyped	Downstream	0.821
HBB	**11:5248004**	**rs11549407**	**A/G**	**0.048**	**−0.473(0.051)**	**1.17×10^{-20}**	**1.5**	**0.917**	**Stop gained, Q40X**	-
CILP2	19: 19456917	rs58489806	T/C	0.074	−0.232(0.042)	2.58×10^{-08}	0.5	Genotyped	Intronic	0.858
APOE	19:45412079	rs7412	T/C	0.036	−0.645(0.053)	2.47×10^{-33}	2.4	Genotyped	Missense, R176C	Same SNP
APOE	19:45411941	rs429358 [a]	C/T	0.074	0.264(0.039)	1.21×10^{-11}	0.8	0.999	Missense, C130R	Same SNP
TC										
PCSK9	1:55505647	rs11591147	T/G	0.038	−0.390(0.053)	1.69×10^{-13}	1.0	Genotyped	Missense, R46L	Same SNP
TMEM33, DCAF4L1, SLC30A9	4:41980435	-	G/A	0.013	−0.520(0.091)	6.94×10^{-9}	0.6	0.91	Intergenic	-
HBB	**11:5248004**	**rs11549407**	**A/G**	**0.048**	**−0.490(0.05)**	**6.88×10^{-22}**	**1.5**	**0.917**	**Stop gained, Q40X**	-
CILP2	19:19456917	rs58489806	T/C	0.074	−0.260(0.041)	2.15×10^{-10}	0.7	Genotyped	Intronic	0.858
APOE	19:45412079	rs7412	T/C	0.036	−0.544(0.053)	2.06×10^{-24}	1.7	Genotyped	Missense, R176C	Same SNP
APOE	19:45411941	rs429358 [a]	C/T	0.074	−0.210(0.038)	2.18×10^{-08}	0.5	0.999	Missense, C130R	Same SNP
HDL										
LPL *	8:19815256	rs286	T/A	0.125	0.257(0.046)	2.70×10^{-08}	1.2	Genotyped	Intronic	0.315
LIPC	15:58687603	rs174418	T/C	0.467	0.136(0.021)	7.96×10^{-11}	0.7	0.999	Intergenic	0.485
CETP	16:56989590	rs247616	T/C	0.268	0.190(0.023)	2.37×10^{-16}	1.1	Genotyped	Intergenic	0.994

Europe PMC Funders Author Manuscripts

Candidate Gene	Chr:position	rs name	Effect Allele / Other	Freq	Effect (StdErr)	pvalue	R2(%)	RSQR	Variant Consequence	r^2 with previous hit
TGIF1 *	18:3412386	rs8092903	T/C	0.026	−0.448(0.082)	4.49×10^{-08}	0.8	0.954	Intronic	-
TG										
LPL	8:19845376	rs7841189	T/C	0.209	−0.160(0.026)	8.36×10^{-10}	0.6	Genotyped	Intergenic	Same SNP
APOA5	**11:116661101**	-	**T/G**	**0.025**	**−0.450(0.064)**	$\mathbf{1.24 \times 10^{-12}}$	**0.9**	Genotyped	**Missense, R282S**	-
APOA5	11:116664040	rs10750097 [b]	G/A	0.172	0.160(0.027)	4.64×10^{-09}	0.6	Genotyped	Upstream	Same SNP
CILP2	19:19456917	rs58489806	T/C	0.074	−0.260(0.039)	2.14×10^{-11}	0.8	Genotyped	Intronic	0.858

[a] Association parameters reported for this marker refer to a model that includes rs7412 as additional covariate

[b] Association parameters reported for this marker refer to a model that includes 11:116661101 as additional covariate

* Results refer to the sex specific analyses. See Supplementary Table 7 for more details.

Europe PMC Funders Author Manuscripts

Nat Genet. Author manuscript; available in PMC 2016 May 01.

Table 3

Summary of Inflammatory Marker Association Results

The table shows the association results at that reach $p < 5\times10^{-8}$ for ADPN, hsCRP, ESR, MCP-1 and IL-6. At each locus, we indicated the genes likely to be modulated by the lead SNP. For each lead SNP, we also showed the rs ID when available, the effect allele and its frequency, the regression coefficients, , the proportion of variance explained by the allele (R2%), the imputation accuracy (RSQR) for those that were imputed, the biological type of the corresponding nucleotide change, and the r^2 with the hits previously reported in ([13]). Novel signals are shown in bold; independent signals are shown in italics.

Candidate Gene	Chr:position	rs name	Effect Allele / Other	Freq	Effect (StdErr)	pvalue	R2(%)	RSQR	Variant Consequence	r^2 with previous hit
ADPN										
ADIPOQ	3:186559460	rs17300539	A/G	0.156	0.247 (0.025)	1.35×10^{-22}	1.6	Genotyped	Intergenic	--
ABHD13	13:108884835	N/A	A/G	0.001	−1.519 (0.275)	3.35×10^{-08}	0.5	0.921	3′UTR	--
hsCRP										
CRP	1:159684665	rs3091244	A/G	0.428	0.207 (0.019)	5.28×10^{-27}	2.0	Genotyped	Intergenic	0.249
PDGFRL	**8:17450500**	**rs73198138**	**A/G**	**0.004**	**−0.894 (0.151)**	$\mathbf{3.31\times10^{-09}}$	**0.6**	**0.977**	**Intronic**	--
HNF1A	*12:121415293* [a]	*rs7139079*	*G/A*	*0.377*	*−0.118 (0.020)*	2.11×10^{-09}	0.6	*0.998*	*Intergenic*	*0.710*
BRI3BP, AACS	**12:125533106**	**rs183233091**	**A/G**	**0.010**	**1.054 (0.094)**	$\mathbf{1.09\times10^{-28}}$	**2.1**	**0.941**	**Intergenic**	--
APOE	19:45411941	rs429358	C/T	0.073	−0.237 (0.036)	3.78×10^{-11}	0.7	1	Missense, C130R	0.565
ESR										
RHCE	*1:25724005* [b]	*rs630337*	*T/C*	*0.297*	*−0.109 (0.020)*	4.03×10^{-08}	*0.5*	*0.957*	*Intronic*	*0.797*
CR1	1:207684359	rs11117956	T/G	0.400	−0.153 (0.018)	9.43×10^{-18}	1.2	Genotyped	Intronic	0.989
HBB	11:5248004	rs11549407	A/G	0.048	−0.437 (0.042)	1.02×10^{-25}	1.8	0.918	Stop gained, Q40X	0.330
AACS, MIR5188	**12:125406340**	**N/A**	**G/A**	**0.007**	**1.034 (0.104)**	$\mathbf{4.40\times10^{-23}}$	**1.6**	**0.952**	**Intergenic**	--
MCP-1										
DARC,CADM3	1:159175354	rs12075	G/A	0.446	−0.405 (0.019)	1.08×10^{-96}	7.2	Genotyped	Missense, G44D	Same SNP
DARC,CADM3	*1:159164454* [c]	*rs2852718*	*C/T*	*0.022*	*−0.515 (0.063)*	3.34×10^{-16}	*1.1*	*0.999*	*Intronic*	*0.005*
DARC,CADM3	***1:159175494*** [d]	***rs34599082***	***T/C***	***0.037***	***−0.338 (0.049)***	$\mathbf{8.23\times10^{-12}}$	***0.8***	Genotyped	***Missense, R89C***	--

Europe PMC Funders Author Manuscripts

Candidate Gene	Chr:position	rs name	Effect Allele / Other	Freq	Effect (StdErr)	pvalue	R2(%)	RSQR	Variant Consequence	r^2 with previous hit
CCR2, CCR3	3:46383906	rs113403743	T/G	0.099	0.273 (0.034)	1.47×10^{-15}	1.1	0.997	Intergenic	0.988
CCR2	***3:46399764*** [e]	***rs200491743***	***A/T***	***0.005***	***0.799 (0.130)***	***9.94×10^{-10}***	***0.6***	Genotyped	***Missense, M249K***	--
N4BP1, CBLN1 *	16:49072490	rs76135610	T/C	0.005	0.969 (0.172)	1.76×10^{-08}	0.9	0.915	Intergenic	--
IL-6										
IL6R	1:154428283	rs12133641	G/A	0.255	0.118 (0.020)	6.87×10^{-09}	0.6	1	Intronic	0.998
ABO	9:136142355	rs643434	A/G	0.263	−0.221 (0.020)	5.80×10^{-27}	2.0	Genotyped	Intronic	0.980

Notes:

[a]Results refer to the conditional analyses after conditioning on rs183233091

[b]Results refer to the conditional analyses after conditioning on rs11117956

[c]Results refer to the conditional analyses after conditioning on rs12075

[d]Results refer to the conditional analyses after conditioning on rs12075 and rs2852718

[e]Results refer to the conditional analyses after conditioning on rs113403743

*Results refer to the female-specific analysis (see Supplementary Table 8 for more details); these genes do not fulfil our specific criteria for being candidates, but they are the nearest to lead SNP in the region (*N4BP1*, 428.3 Kb; *CBLN1*, 239.3 Kb)

Europe PMC Funders Author Manuscripts

Table 4

Rare variant tests

The table shows results for the rare variant association tests at genes passing the significant threshold for at least on the two statistical tests (CMC and VT). Of note, no significant results were observed for LDL-c, hsCRP and IL-6. For each gene, we indicated the genomic location assessed for analyses (in hg19 genomic build), the number of available SNPs considered, the number of SNPs passing the tests-specific criteria for inclusion, and the number and the fraction of individuals carrying a rare allele. For the CMC test, the effect size and its standard error, along with the pvalue and the phenotypic variance explained are reported. For the VT the impact on the phenotype (+ increase, − decrease) of rare variants, the pvalue and the phenotypic variance explained are reported. We also reported the pvalue observed after adjusting for the lead variant at the same or the nearby gene. Specifically, *STAB1* was adjusted for rs7639267; *CCR2* was adjusted for rs113403743 and rs200491743; *IFI16* was adjusted for rs12075, rs2852718 and rs34599082; *HBB* and *OR52H1* were adjusted for rs76728603, and *PTPRH* was adjusted for the best lead in the region (rs7253814). Genes that remain significant after adjustment are marked in bold.

Gene	Chr:Start-end	#SNPs	#Pass	Burden Fraction with Count rare		CMC test				VT test			
						Effect(StdErr)	pvalue	R2	Adjusted pvalue	Direction	Pvalue	R2	Adjusted pvalue
ADPN													
STAB1	3:52535766-52558237	25	23	752	0.12886	0.245 (0.039)	4.71×10^{-10}	0.007	**1.92×10^{-09}**	+	1.00×10^{-07}	0.007	**1.00×10^{-07}**
MCP1													
CCR2	3:46399158-46401290	4	3	105	0.01797	0.541 (0.104)	1.84×10^{-07}	0.005	0.7092	+	1.00×10^{-06}	0.005	0.92
IFI16	1:158979950-159024668	10	8	567	0.09702	0.218 (0.046)	2.50×10^{-06}	0.004	0.1564	+	$1.40\times^{-05}$	0.003788	0.115
ESR													
HBB	11:5247914-5248004	2	2	613	0.10318	−0.345 (0.039)	9.77×10^{-19}	0.013	0.015	−	1.00×10^{-07}	0.013	0.025
OR52H1	11:5565906-5566751	5	3	529	0.08904	−0.205 (0.042)	1.23×10^{-06}	0.004	0.345	−	3.40×10^{-06}	0.004	0.69
PTPRH	19:55693244-55716713	22	15	1152	0.19391	−0.146 (0.029)	8.31×10^{-07}	0.004	**4.22×10^{-06}**	−.	1.18×10^{-05}	0.0041	1.90×10^{-05}