Population- and individual-specific regulatory variation in Sardinia

Mauro Pala1–3, 12, Zachary Zappala4, 12, Mara Marongiu1, Xin Li2, Joe R Davis4, Roberto Cusano1, Francesca Crobu1, Kimberly R Kukurba3, Michael J Gloudemans5, Frederic Reinier6, Riccardo Berutti3, 6, Maria G Piras1, Antonella Mulas1, Magdalena Zoledziewska1, Michele Marongiu1, Elena P Sorokin4, GaeLEN T Hess4, Kevin S Smith2, Fabio Busonero1, Andrea Maschio1, Maristella Steri1, Carlo Sidore1, Serena Sanna1, Edoardo Fiorillo1, Michael C Bassik4, Stephen J Sawcer7, Alexis Battle8, John Novembre9, Chris Jones6, Andrea Angius1, Gonçalo R Abecasis10, David Schlessinger11, Francesco Cucca1, 3, 13 & Stephen B Montgomery2, 4, 13

Genetic studies of complex traits have mainly identified associations with noncoding variants. To further determine the contribution of regulatory variation, we combined whole-genome and transcriptome data for 624 individuals from Sardinia to identify common and rare variants that influence gene expression and splicing. We identified 21,183 expression quantitative trait loci (eQTLs) and 6,768 splicing quantitative trait loci (sQTLs), including 619 new QTLs. We identified high-frequency QTLs and found evidence of selection near genes involved in malarial resistance and increased multiple sclerosis risk, reflecting the epidemiological history of Sardinia. Using family relationships, we identified 809 segregating expression outliers (median z score of 2.97), averaging 13.3 genes per individual. Outlier genes were enriched for proximal rare variants, providing a new approach to study large-effect regulatory variants and their relevance to traits. Our results provide insight into the effects of regulatory variants and their relationship to population history and individual genetic risk.

Human migration and rapid population expansion have led to an abundance of population- and individual-specific genetic variation1–5. Within protein-coding regions of the genome, multiple studies have identified numerous rare loss-of-function alleles6–11 that affect monogenic disorders and, to a lesser extent and especially in founder populations, common diseases and complex traits12–14. Most of the variants associated with complex traits are found outside protein-coding regions, however, and their functional consequences remain elusive. Large studies of gene expression have greatly advanced the ability to identify functional variation in noncoding regions of the genome15–17, and many of these variants have been connected to common genetic diseases18, 19. However, few studies thus far have had access to whole-genome sequencing data, family relationships, and auxiliary complex trait data from research participants. Such data have the potential to empower assessment of the population- and individual-specific consequences of regulatory variants.

To overcome this, we sequenced RNA isolated from the white blood cells of 624 individuals from the founder population of Sardinia. Study of the Sardinian population has several advantages: DNA from these individuals includes the bulk of mainland European DNA variation, but, owing to a period of relative isolation for >10,000 years, many alleles have been added and many old and novel variants have reached dramatically higher frequencies, which should improve power to detect associations between those variants and traits such as gene expression20–22. In addition, the SardiNIA study cohort has been extensively genotyped and phenotyped and consists of both unrelated and related individuals23. By combining RNA-seq data with whole-genome sequencing data, we discovered eQTLs and sQTLs that are specific to the isolated Sardinian population. As this is the first eQTL and sQTL study that, to our knowledge, integrates both whole genomes and transcriptomes from multiple families, we developed a framework that leverages these family relationships to identify large-effect, rare regulatory variants. We identified extreme gene expression outliers that segregate within families and investigated the distribution and associated functional annotations of putatively causal rare variants as well as their influence on individual disease risk. This approach enhances ongoing studies of loss-of-function variants by demonstrating a new approach to identifying and studying large-effect alleles.

1Istituto di Ricerca Genetica e Biomedica (IRGB), CNR, Monserrato, Italy. 2Department of Pathology, Stanford University School of Medicine, Stanford, California, USA. 3Dipartimento di Scienze Biomediche, Universita di Sassari, Sassari, Italy. 4Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. 5Program in Biomedical Informatics, Stanford University School of Medicine, Stanford, California, USA. 6CRS4, Advanced Genomic Computing Technology, Pula, Italy. 7Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. 8Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, USA. 9Department of Human Genetics, University of Chicago, Chicago, Illinois, USA. 10Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA. 11Laboratory of Genetics, National Institute on Aging, Baltimore, Maryland, USA. 12These authors contributed equally to this work. 13These authors jointly directed this work. Correspondence should be addressed to S.B.M. (smontgom@stanford.edu) or F. Cucca (francesco.cucca@irgb.cnr.it).

Received 2 August 2016; accepted 17 March 2017; published online 10 April 2017; doi:10.1038/ng.3840
RESULTS
Expression and splicing QTL discovery in Sardinia
The 624 participants, all from four towns in the Lanusei Valley in the Ogliastra region of Sardinia, were enrolled from a cohort of 6,921 individuals in the SardiNIA longitudinal study of aging. The entire SardiNIA cohort was genotyped using the Cardio-MetaboChip, Immunochip, ExomeChip, and OmniExpress arrays. A subset of 2,120 Sardinians were additionally whole-genome sequenced at low coverage (average fourfold coverage), producing an integrated map of approximately ~15 million SNPs after imputation. This cohort and imputation pipeline has previously been described. For RNA, we sequenced a total of 36 billion reads in total. After quantification and quality control, we obtained a median of ~59 million 51-bp paired-end reads per participant (over 15,243 and 12,603 genes were sufficiently expressed for eQTL and sQTL analyses, respectively (Table 1). To account for confounding effects that can reduce power to discover cis QTLs, we performed hidden factor correction with PEER. We were able to identify and remove factors correlated with sex, age, various blood cell counts, and sequencing (Supplementary Fig. 1 and Supplementary Table 1).

To discover eQTLs, we tested the association of genotype with expression level for all variants within ±1 Mb of a target gene's transcription start site (TSS) for all individuals with genetic data in the integrated map (n = 606). At a false discovery rate (FDR) of 5%, we identified eQTLs for the majority of tested genes (Table 1). We then used a forward stepwise regression approach to characterize the number of independent eQTLs per gene (Online Methods). We found that approximately half of all protein-coding and long noncoding RNA (lncRNA) transcripts were influenced by at least two independent eQTLs; microRNAs (miRNAs), however, were mostly associated with a single eQTL (Table 2). At the extreme, we found a single protein-coding gene, ITGB1BP1, affected by 14 independent eQTLs. ITGB1BP1 encodes an integrin-binding protein that is implicated in upstream regulation of immune critical TNF-NF-kB transcriptional regulation. We also identified an lncRNA of unknown function, NBPF1, that was affected by 11 independent eQTLs (Supplementary Table 2). In total, we mapped at least one eQTL for 73% of the tested genes, corresponding to 11,167 primary eQTLs. Our forward stepwise regression analysis identified an additional 1,648 secondary eQTLs, for a total of 6,768 sQTLs (Table 2).

We compared our forward stepwise regression approach to an alternative method implemented in the MLMM software package, which uses a stepwise mixed-model regression with forward inclusion and backward elimination to identify independent associations. The two approaches resulted in a similar number of independent eQTL associations (Supplementary Table 6) and sQTL associations (Supplementary Table 7) that were largely consistent with our original findings (Supplementary Note). We also performed simulations to assess the impact of statistical noise and missing SNPs on our independent association analyses. We ran our pipeline on simulated expression traits where a single, randomly selected SNP in the region located ±1 Mb with respect to each gene explained 25% of the trait's variance (we did this for each gene where at least one eQTL was found in the original analysis). Only a small fraction of these simulations resulted in multiple independent associations in comparison to the actual analysis (Supplementary Table 8). We repeated this simulation a second time but excluded the randomly chosen causal SNP from the association mapping phase. While we observed more independent associations relative to the first analysis (Supplementary Table 9), the similarity between the results of these simulations and the consistency observed between our pipeline and MLMM suggest that our approach is identifying independent associations and is robust to statistical noise and residual linkage disequilibrium (LD) blocks near these genes.

Comparison of Sardinian and European eQTLs
We next measured the replication of Sardinian QTLs with European eQTLs found in lymphoblastoid cell lines (LCLs; GEUVADIS) and whole blood (Depression Genes and Networks, DGN). For the Sardinian eQTLs that were tested in each study, the replication rate was 92% in DGN and 72% in GEUVADIS, reflecting the high degree of sharing of common European alleles within Sardinia (Supplementary Table 10). For sQTLs, the replication rate was 72% in DGN and 76% in GEUVADIS. Additionally, we tested eQTLs and sQTLs found in either DGN or GEUVADIS for replication in the Sardinian cohort and found that 89–92% of eQTLs and 70–97% of sQTLs replicated (Supplementary Table 11). Replication could not be tested for 2,568 eQTLs and 1,152 sQTLs found in Sardinia because the SNPs were either absent in Europe or only present at a minor allele frequency (MAF) below 1%.
QTLs, 437 eQTLs and 182 sQTLs were novel in Sardinia when compared to the 1000 Genomes Project, dbSNP, UK10K, and Exome Aggregation Consortium (ExAC) databases, representing new and/or previously uncaptured functional variation.

We first observed that novel eQTLs were depleted from known disease-associated genes (Supplementary Fig. 3). To determine whether these new eQTLs were associated with traits measured in Sardinia, we tested all 437 new eQTL variants for associations with 15 blood cell measurements in the whole SardinIA cohort (n = 6,000; Supplementary Table 12). We identified five associations (five traits and four variants, in total) that were significant after correcting for multiple testing (P < 8.8 × 10^{-6}). For each association, we then retested the trait association for all variants within ±4 Mb of the target gene to identify the subset of loci where the Sardinia-specific eQTL variant and the top trait-associated variant within this region were in high LD (r^2 > 0.8). We identified a Sardinia-specific eQTL for ARHGDIB that was linked to the top trait-associated variant for neutrophil percentage, which is also Sardinia specific (Supplementary Fig. 4; top neutrophil percentage variant chr12:g.14190223T>C, P = 3.8 × 10^{-6}; top eQTL chr12:g.15553026G>T, P = 7.69 × 10^{-6}; r^2 = 0.86). We further performed an eQTL–trait colocalization analysis with eCAVIAR29 and observed strong local colocalization of the ARHGDIB eQTL with variants for both neutrophil and lymphocyte percentages (Supplementary Fig. 5). Within this locus, only 3 of 14 variants that passed our LD filter have previously been reported outside of Sardinia (allele frequencies in Europe below 0.002). Of note, one of these variants (chr12:g.15095546G>C, P = 3.85 × 10^{-6}; r^2 with top neutrophil signal = 0.84) is a nonsense mutation that had been observed only once in the ExAC database but has a frequency >1% in Sardinia, with the direction of effect on expression consistent with nonsense-mediated decay. ARHGDIB presents a biologically plausible role in inhibition of cell migration and lymphocytes from Arhdgb/-/- mice show changes in expansion and survival in culture30.

**Table 2** Independent QTLs segmented by gene type

| Max number of independent QTLs | Protein coding | IncRNA | miRNA precursor | Protein coding | IncRNA |
|---------------------------------|----------------|--------|-----------------|----------------|--------|
| (number of genes)               | (number of genes) |        |                  | (number of genes) |        |
| 1                               | 4,215           | 598    | 44              | 3,489          | 281    |
| 2                               | 2,833           | 386    | 8               | 799            | 60     |
| 3                               | 1,235           | 170    | 2               | 165            | 22     |
| 4                               | 428             | 66     | 0               | 36             | 8      |
| 5                               | 175             | 18     | 1               | 18             | 1      |
| 6                               | 82              | 6      | 0               | 3              | 1      |
| 7                               | 27              | 5      | 0               | 5              | 0      |
| 8                               | 14              | 3      | 0               | 0              | 0      |
| ≥9                              | 10              | 6      | 0               | 0              | 0      |
| Total                           | 9,019           | 1,258  | 55              | 4,515          | 373    |

We report the number of independent QTLs for gene-level and isoform-level analyses. Isoform results are grouped by the respective gene.

Functional enrichment in highly differentiated eQTLs

We next tested whether two epidemiological factors present in Sardinia were reflected among highly differentiated eQTLs. Until the mid-twentieth century, the Sardinian population suffered high mortality rates due to malaria34,35 and continues to have a higher prevalence of multiple sclerosis relative to other European-descent populations in the Mediterranean basin36,37. We identified a significant enrichment for known malarial resistance genes (P = 0.0015) and genes associated with multiple sclerosis (EBI/NHGRI GWAS Catalog nominal P = 1.84 × 10^{-5} and ImmunoBase nominal P = 1.17 × 10^{-8}) among the top 1% of differentiated eQTLs (mean allele frequency difference of ~17%) (Fig. 2c,d and Supplementary Table 14). Multiple sclerosis had the highest enrichment among 354 traits tested from the EBI/NHGRI GWAS catalog and among 19 traits tested from the ImmunoBase catalog (Fig. 2d and Supplementary Fig. 7). We also
found that genome-wide association study (GWAS) hits for multiple sclerosis showed evidence for colocalization with eQTLs identified in Sardinia, suggesting that regulation of these genes mediates the association signals at these loci (Supplementary Table 15).

One of the most differentiated eQTLs was associated with expression levels of the BAFF gene ($P = 8.05 \times 10^{-12}$), allele frequency difference between Sardinians and Europeans ($\Delta \text{AF}_{\text{SARD–EUR}} = 0.25$), which is known to be involved in the response to and survival of malaria infection\(^{38–40}\) and has unique evolutionary history in Sardinia (M.S., V. Orrù, M.I. Ilda, M. Pitzalis and M.P. unpublished data). We also identified several regulatory variants for the CR1 gene, whose product is involved in complement activation and immune complex formation during malaria infection\(^{41,42}\). CR1 has two eQTLs (chr1:g.207275799G>A and chr1:g.207667190G>C) and nine sQTLs. The eQTL at chr1:g.207667190G>C is highly differentiated between Sardinia and Europe ($\Delta \text{AF}_{\text{SARD–EUR}} = 0.25$) (Supplementary Fig. 8). Among the nine sQTLs associated with CR1, two are highly differentiated: chr1:g.207685105C>G ($\Delta \text{AF}_{\text{SARD–EUR}} = 0.42$) influences the abundance of ENST00000367051 and chr1:g.207716099A>C ($\Delta \text{AF}_{\text{SARD–EUR}} = 0.43$) influences the abundance of ENST00000529814. Both sQTLs are tightly linked and in high LD ($r^2 = 0.99$ and 0.95, respectively) with variant chr1:g.207757515A>G that has previously been associated with erythrocyte sedimentation rate in the SardiNIA cohort\(^43\).

Finally, as difference in allele frequency itself does not account for background selection near genes, we used an alternative method to define differentiated Sardinian eQTLs based on $F_{\text{ST}}$ values (Supplementary Note). Differentiated eQTLs identified with this method were similarly enriched near genes associated with malaria ($P = 4.91 \times 10^{-3}$; Supplementary Table 16) and near loci for multiple sclerosis (Fig. 2d), with multiple sclerosis being the most significantly enriched trait in both the EBI/NHGRI GWAS Catalog (nominal $P = 2.11 \times 10^{-3}$; Supplementary Table 17) and ImmunoBase (nominal $P = 7.41 \times 10^{-5}$; Supplementary Table 18).

Heritable patterns of extreme gene expression in families

Beyond the unique history of the Sardinian population, the availability of family relationship data in the SardiNIA cohort provided an opportunity to identify the impact of rare, large-effect regulatory variation. Specifically, we developed a likelihood-ratio test to identify patterns of extreme gene expression that segregated in families (Fig. 3a and Online Methods). We tested 61 Sardinian trios for the 15,243 genes included in our eQTL analyses and identified 809 genes where a parent and child were both expression outliers (median $z$ score = 2.97) at an FDR of 10% (Fig. 3b). On average, we found 13.3 shared expression outlier genes per child.

Several lines of evidence suggest that shared expression outliers are not due simply to parent–offspring shared environment. There was little correlation of gene expression between the outlier parent and the non-outlier partner (Pearson $r = 0.20$) (Fig. 3d). Mothers and fathers were equally likely to be the outlier parent ($P = 0.20$), regardless of the sex of the child ($P = 0.83$) (Supplementary Fig. 9). We used a separate method to identify outliers on the basis of $z$ scores alone and found that approximately 10% of the average child’s extreme expression outliers were shared with one parent alone and the remaining 90% are likely not caused by genetics\(^44\) (Supplementary Fig. 10). These results are concordant with those of Tabassum et al.\(^44\), who found ~100 expression outliers per individual that could be largely explained by extrinsic factors, for example, cell type proportions.

We found almost twice as many shared underexpression outliers (529 outliers; 65%) as shared overexpression outliers (280 outliers; 35%), consistent with observations of the effects of random substitutions in promoters and enhancers in massively parallel reporter assays\(^45–47\). Furthermore, because rare variants tend to be heterozygous and thus only influence one allele, we hypothesized that outlier parents and children would be enriched for allele-specific expression in comparison to non-outlier controls. We found that allele-specific expression was significantly enriched in outlier individuals for both under- and overexpression outliers (adjusted Wilcoxon rank-sum $P = 6.0 \times 10^{-6}$) (Fig. 3c). This is likely a conservative estimate of the true enrichment, given the inherently low levels of read depth in underexpression outliers that limit the ability to measure allelic effects in outlier genes. These allelic effects were consistent between outlier parents and children (Pearson $r = 0.84$) but not between children and the other, non-outlier parent (Fig. 3d). The strength of the outlier effect was also significantly associated with the enrichment of allele-specific expression (Spearman correlation).
\( \rho = 0.338, P < 1 \times 10^{-6} \), reflecting the capacity of allele-specific effects to have an impact on total expression (Fig. 3e).

**Rare variants can underlie extreme gene expression in families**

Using the combination of whole-genome data and family relationships, we were able to characterize potential causal variants underlying expression outliers. We first identified 3,464 rare variants (Sardinia MAF < 1%) that were located in 250-kb windows adjacent to the TSS and transcription end site (TES) of outlier genes and were unambiguously transmitted from the outlier parent to the outlier child (the variant was heterozygous in both the outlier parent and child, and the other parent was homozygous for the reference allele). We also identified an equivalent set of 245,165 rare variants in the same genomic loci that were unambiguously transmitted from non-outlier parents to their children. We found at least one shared rare variant for 509 of the outlier genes (63%), with an average of 6.8 variants shared by outliers versus 4.0 shared by non-outliers (enrichment = 1.71, 95% confidence interval = 1.65–1.77). Of interest, rare variants shared by outlier individuals were concentrated within 5 kb of the TSS (enrichment = 3.61, 95% confidence interval = 2.96–4.24) and TES (enrichment = 3.00, 95% confidence interval = 2.44–3.54) (Fig. 4a) of outlier genes, similar to what has been observed for common regulatory variation.

Furthermore, rare variants shared by outliers were enriched in multiple functional annotations (Fig. 4b and Supplementary Fig. 11). For variants in the 50-kb window adjacent to the TSS, this enrichment was most notable in splice donor/acceptor sites (log odds = 4.05, \( P = 2.52 \times 10^{-7} \)) and regions associated with active transcription, including promoters (log odds = 0.91, \( P = 8.8 \times 10^{-9} \)) and enhancers (log odds = 0.42, \( P = 0.0094 \)) (enrichment data for different genomic window sizes are provided in Supplementary Tables 19 and 20). We further investigated whether other carriers of these variants had the same outlier expression profile as the outliers.
parent–child pairs. We analyzed 2,912 variants (84% of the 3,464 outlier variants) that were heterozygous in at least four individuals in the cohort, regressing outlier gene expression on genotype at the rare variant position. The largest and most significant of these genotype–expression associations for both over- and underexpression outliers were concentrated at the TSS of outlier genes (Fig. 4c). Additionally, we found that metrics of conservation (GERP, PhyloP) and predicted functional relevance (FitCons, CADD) all discriminated the most significant associations (Fig. 4d).

On the basis of these observations, we developed a strict set of rules to distinguish putatively causal rare variants by prioritizing variants that were close to the TSS or likely involved in splicing, highly conserved, and replicated their effects in the larger population (Online Methods). We identified candidate causal variants for 30 outlier genes (Supplementary Table 21), including five rare splicing variants. One of these splicing variants, chr12:g.121570899G>T, is found at the first exon–intron boundary of the P2RX7 gene, which encodes a ligand-gated ion channel responsible for ATP-dependent lysis of macrophages. While chr12:g.121570899G>T is rare in all European populations, including Sardinia, where it is most frequent with MAF = 0.009% (Supplementary Table 22), it has previously been shown to disrupt proper splicing of P2RX7, leading to an elongated transcript that is subsequently degraded by nonsense-mediated decay and results in monoallelic expression. As expected, all carriers of chr12:g.121570899G>T (n = 12) in the Sardinian cohort underexpressed P2RX7, and all reads showed the same allele. While the other splicing variants have not been characterized, we saw similar trends for all five splicing variants, suggesting that all of these putative splicing variants are effectively null alleles (Fig. 5).

Because the SardiNIA cohort has been extensively phenotyped, we were able to test for the association of rare variants with measured traits. Of the 30 putatively causal variants, 11 were associated with the expression of genes near significant GWAS loci. Of these, five genes (SPECC1, GLB1, CADM1, BRI3BP, and ANXA5) were associated with traits measured in the Sardinian cohort. However, we found no significant association between the five candidate variants for these genes and their matched GWAS trait (Supplementary Table 23). We found no significant relationship between expression levels of these genes and their matched GWAS trait (Supplementary Table 23), suggesting that either the gene is not involved in the trait or dosage is not a critical factor. We next searched for outlier genes that have established roles in the manifestation of rare clinical traits. We were able to identify three outlier genes associated with clinical traits in our database: VPS13D is known to repress IL-6 production; TSSC1 suppresses osteolysis; and mutations in POMGNT1 disrupt dystroglycan and can interfere with skeletal muscle function. For each gene, we tested for association of the genotype of the candidate rare variant with levels of the appropriate trait and then for overall association between gene expression and the trait. We were, however, unable
to find any significant evidence for association (Supplementary Table 24), consistent with recent observations in British-Pakistani cohorts for association testing of rare protein-coding variants in trait-associated genes. While we were unable to identify any direct association between rare variants and clinical traits, we did observe a modest enrichment of outliers in potential disease-associated genes and a marked enrichment of outlier genes in loss-of-function intolerant genes relative to common eQTLs (Supplementary Fig. 12).

**DISCUSSION**

Our study focused on identifying the effect of population- and individual-specific regulatory variants in Sardinia. We identified hundreds of new or highly differentiated regulatory alleles and observed that these alleles reveal new trait associations and reflect the island’s epidemiological history of multiple sclerosis and malaria. By combining whole-genome sequencing data with transcriptomes from many families, we were able to identify patterns of outlier gene expression and implicate the functional role of rare regulatory variants. While such observations have previously been limited to unrelated individuals, we were able to identify hundreds of genes with large heritable effects and candidate rare regulatory variants. Relating the effects of candidate rare regulatory variants to phenotypes remained a major challenge, comparable to systematic efforts to identify the phenotypic consequences of rare, protein-coding loss-of-function alleles. However, we observed that outlier expression effects were more prevalent in genes intolerant of loss-of-function variation, consistent with their increased potential for important individual consequences.

As gene expression assays complement whole-genome sequencing, discovery of population-specific and rare, large-effect regulatory variants will enable the generation of new hypotheses to understand the molecular etiology of diverse disorders and increase understanding of the utility of different genes as potential therapeutic targets. In particular, identifying extreme patterns of gene expression can be used to provide a more nuanced view of genic dosage tolerance than is revealed by naturally occurring knockouts. We anticipate that large catalogs of rare, large-effect regulatory variants, found in either isolated populations or families, will yield new opportunities for clinical interpretation of the noncoding genome, precision health, and understanding of genome biology.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**ACKNOWLEDGMENTS**

All participants gave informed consent, with protocols approved by institutional review boards for ASL4 in Sardinia and by the University of Michigan. IRB exemption (OHSPR 11916) applied to analyses on coded data at collaborating institutions. M.P. is supported by the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement 633964 (ImmunoAgeing). Z.Z. is supported by the National Science Foundation (NSF) GRFP (DGE-114747) and by the Stanford Center for Computational, Evolutionary, and Human Genomics (CEHG). Z.Z., J.R.D., and G.T.H. also acknowledge support from the Stanford Genome Training Program (SGTP; NIH/NHGRI T32HG000044). J.R.D. is supported by the Stanford Graduate Fellowship. K.R.K. is supported by the Department of Defense, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship 32 CFR 168a. S.J.S. is supported by the NIH-Cambridge Biomedical Research Centre. The SardiNIA project is supported in part by the intramural program of the National Institute on Aging through contract HHSN271201100005C to the Consiglio Nazionale delle Ricerche di Italy. The RNA sequencing was supported by the PB05 InterOmics Flagship grant; by the FareBio2011 “Farmaci e Reti Biotecnologiche di Qualità” grant; and by Sardinian Autonomous Region (L.R. no. 7/2009) grant CRP3-154 to F. Cucca, who is also supported by the Italian Foundation for Multiple Sclerosis (FISM 2015/R/09) and by the Fondazione di Sardegna (ex Fondazione Banco di Sardegna, Prot. U1301.2015/AL1157.BE Prat. 2015-1651). S.B.M. is supported by the US National Institutes of Health through R01HG008150, R01HH081184, U01HG007436, and U01HG009980. All of the authors would like to thank the CRS4 and the SCGPM for the computational infrastructure supporting this project.

**AUTHOR CONTRIBUTIONS**

M.P., Z.Z., Mara Marongiu, G.R.A., D.S., F.Cucca, and S.B.M. conceived and designed the experiments. Mara Marongiu, R.C., F.Crobu, M.G.P., A.Mulas, M.Z., F.B., A. Maschio, E.F., and A.A. performed the experiments. M.P., Z.Z., X.L., J.R.D., M.J.G., G.R.A., F.Cucca, and S.B.M. performed statistical analysis. M.P., Z.Z., X.L., J.R.D., K.R.K., M.J.G., F.R., R.B., Michele Marongiu, M.S., C.S., S.S., A.B., J.N., G.R.A., D.S., F. Cucca, and S.B.M. analyzed the data. M.P., Z.Z., M.C.B., A.B., J.N., C.J., S.J.S., G.R.A., D.S., F. Cucca, G.T.H., E.P.S., K.S.S., and S.B.M. contributed reagents, materials, and/or analysis tools. M.P., Z.Z., J.N., G.R.A., D.S., F. Cucca, and S.B.M. wrote the manuscript. M.P. and Z.Z. contributed equally. F. Cucca and S.B.M. jointly directed research. All authors read and approved the final version of the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.
ONLINE METHODS

Study population and sample acquisition. Our study was performed on a subset of 624 participants from the larger Sardinia cohort. All 624 participants live in the Lanusei Valley in the Ogliastra region of Sardinia. Participants represented a mixture of related individuals, including 61 complete trios, and unrelated individuals (n = 188; Supplementary Fig. 13). Whole genomes for 606 of these samples were available from a previously published study20. For each participant, leukocytes were isolated from whole blood using the LeuKLOCK fractionation kit and RNA was extracted using TRI Reagent (Ambion, AM9738) and isolated using the PureLink RNA Mini Kit (Ambion, 12183018A). The quantity and integrity of isolated RNA samples was evaluated using the Agilent Technologies 2100 Bioanalyzer platform with the RNA 6000 LabChip kit (Agilent, 5067-1511); samples with an RNA integrity number (RIN) less than 7.5 were discarded. Poly(A)⁺RNA was isolated from 4 μg of high-quality total RNA samples through two rounds of positive selection and purification using magnetic beads following the TruSeq RNA Sample Preparation manual (Illumina, 15015050).

Sequencing library preparation, alignment, and quality control. Before library preparation, we added one of two ERCC RNA Spike-in Control Mixes (Ambion, 4456740) to 288 samples at a 1:625 final dilution to assess the uniformity of library preparation across samples. Purified RNA samples were then processed into indexed, paired-end cDNA libraries using the TruSeq RNA–Seq Library Preparation kit. Following purification, amplification, and cleanup, cDNA libraries were quantified using the Agilent Technologies 2100 Bioanalyzer with the Agilent DNA 1000 assay (Agilent, 5067-1504). Sample-specific cDNA libraries were then pooled to obtain equimolar concentrations and loaded onto a paired-end flow cell using the Illumina cBot System and TruSeq PE Cluster Generation kit v3 (Illumina, PE-401-3001). 51-bp paired-end reads were generated on an Illumina HiSeq 2000 using TruSeq SBS v3 reagents (Illumina, FC-401-3001). Demultiplexed FASTQ files were generated and aligned to the human reference genome (hg19) using STAR (version 2.2.0c)69. Three of the 627 samples were discarded because of extreme GC-content biases, and we observed several other well-known technical biases that we ultimately corrected for (Supplementary Fig. 14). A full description of library preparation and quality control procedures is available in the Supplementary Note.

Quantification and normalization of gene, isoform, and allele-specific expression. Gene levels were quantified using HTSeq10 (0.5.4p5) over the ENCODE v14 annotation; counts were converted to FPKM2728 and variance was stabilized using Deseq23 (1.10.1). We then ran PEER26 (v1.3) to identify and remove confounding factors. The number of hidden factors to remove was decided by empirically optimizing our ability to discover eQTLs on a random subset of 1,500 genes. eQTLs were mapped using Merlin64 on PEER residuals after removing k hidden factors (we tested various k values in the range of 0 to 100; Supplementary Note); we found that removing 30 hidden factors maximized our power to discover eQTLs (Supplementary Fig. 1). We attempted to identify the biological or technical sources of these hidden factors; many corresponded to known technical biases like GC content, 3′ and 5′ biases, etc. (Supplementary Table 1). We additionally filtered out non-autosomal genes, genes with a mean FPKM less than 0.3 across all 624 samples, and genes with an FPKM of 0 in 50% or more of the 624 samples. Ultimately, we mapped eQTLs for 15,243 genes that passed these filters. Isoform quantification was performed for these 15,243 genes using Cufflinks67 (v2.2.1). Isoform proportions were computed as the ratio of the isoform FPKM relative to the sum of the FPKMs for all the isoforms for each gene. We then filtered out genes with only one expressed isoform and where the isoform ratios did not follow a normal distribution (Supplementary Note). For the 606 samples where whole-genome data were available, allele-specific expression data were generated using SAMtools65 (v1.2) mpileup and quantified as the deviation of the reference allele ratio from 0.5. We only considered heterozygous sites with at least 30 reads and where reads for both the reference and alternate alleles comprised at least 2% of all supporting reads. We additionally restricted our analyses to sites with an ENCODE mappability score equal to one. Finally, we excluded allele-specific expression data for 49 genes that showed significantly biased trends in allelic effects across individuals in our study, the DGN cohort, or the GEUVADIS cohort (Supplementary Fig. 15 and Supplementary Table 25).

Quantitative trait locus mapping. We used an integrated map of ~15 million SNPs for the 606 genotyped samples to map eQTLs and sQTLs using Merlin64 (v1.1.2). We excluded variants that were not in Hardy–Weinberg equilibrium (P < 1 × 10⁻⁶), that had a MAF <1% in the 606 samples, or that had an imputation quality value less than 0.3. Expression values (either expression residuals or isof orm ratios) were standardized using Merlin’s inverse normal option. For each gene and isoform, we tested the association of the trait with all cis variants within 1 Mb of the TSS of the gene. We estimated the overall FDR by permutation (Supplementary Note). We additionally calculated adjusted P values by selecting the top association for each gene or isoform, applying a gene-level Bonferroni correction, and applying the Benjamini–Hochberg procedure66 to the collection of top associations. Independent gene or isoform QTLs were identified by forward stepwise regression, in which significant QTLs were iteratively regressed out until the next best QTL was no longer significant at an FDR of 5% (Supplementary Fig. 16). We also identified independent QTLs using MLMim29, a stepwise linear mixed-model approach, and found results similar to those from our Merlin-based pipeline (Supplementary Tables 6 and 7). Additionally, we performed simulations to show that our independent QTL results were not a result of statistical noise, residual LD, or genotyping errors. Specifically, we repeated the following simulation ten times. For each gene with at least one eQTL, we chose a common SNP (MAF > 5%) within 1 Mb of the TSS to explain 25% of the gene expression variance in the simulated trait. We then ran our Merlin-based pipeline to detect independent eQTLs on the simulated expression traits, iteratively regressing out significant SNPs. We repeated these simulations a second time, excluding the randomly selected causal SNP from the association stage. We then compared the number of independent eQTLs identified in the real data versus the simulated data sets (Supplementary Tables 8 and 9).

We mapped eQTLs by computing the Spearman correlation of allelic imbalance in the 15,243 expressed genes with the genotype of nearby cis variants (within 1 Mb of the heterozygous site). Genotypes at cis variants were encoded as 0 (samples homozygous for the reference or non-reference allele) or 1 (heterozygous samples). To compare effect sizes across studies, we identified eQTLs and asQTLs in 188 unrelated Sardinians and compared them with a randomly chosen subset of 188 unrelated individuals in DGN17 and GEUVADIS15. eQTLs in the 188 unrelated individuals for each cohort were recalculated using Matrix eQTL57. We estimated the reproducibility of the Sardinian eQTLs using the π1 statistic68 after reprocessing each data set with our pipeline (Supplementary Tables 10, 11, and 26). A full description of how we controlled for power differences across studies is available in the Supplementary Note.

Colocalization of GWAS and eQTL signals. Colocalization analyses were performed with eCAVIAR39 using default parameters. eCAVIAR calculates a posterior probability that two association signals overlap (CLPP score), accounting for LD in the study population where the two signals are measured. The supplied LD was computed with VCFtools (for the GWAS signals outside of Sardinia, we used LD calculated for European genotypes in the 1000 Genomes Project). For the ARHGDB colocalization analyses, associations with neutrophil and lymphocyte percentages were calculated within Sardinia. For the colocalization analysis between multiple sclerosis GWAS and eQTL associations, we used GWAS data provided by the International Multiple Sclerosis Genetics Consortium40 and primary eQTL association data from Sardinia (the association for each SNP without adjusting for conditionally independent eQTLs). We calculated the CLPP score for the identified multiple sclerosis–associated gene as well as nearby genes (±1 Mb with respect to the GWAS SNP) and report the rank of the identified gene in the list of all genes tested for that GWAS locus (Supplementary Table 15). For the 21 genes we tested, 14 of the target genes had the highest evidence of colocalization versus background and 3 had the second highest evidence of colocalization. Only two genes showed very little evidence of colocalization (CD6 and CTD-2006C1.2).
Allelic differentiation, selection, and disease association. Analysis of allelic differentiation and selection was carried out on a subsample (n = 691) of the SardiNIA cohort for which phased genotyped data were already available and on data from the 1000 Genomes Project Phase 3 (ref. 5). iHS values were computed using selscan software on common variants (MAF ≥ 1%) that passed quality control filters (Supplementary Note). The delta allele frequency for Sardinian QTLs, ΔAF_{SARD–EUR}, was computed as the deviation between the Sardinian MAF and the European MAF (as computed by the 1000 Genomes Project). We then tested for enrichment of different eQTLs near significant GWAS loci. Briefly, we identified significant eQTLs in high LD with significant GWAS SNPs (r² > 0.8). For each GWAS trait, we then built a 2 x 2 count table where the rows separated differentiated eQTLs from non-differentiated eQTLs and the columns separated eQTLs in LD with a GWAS SNP from eQTLs not in LD with a GWAS SNP. We then performed a Fisher’s exact test on each GWAS contingency table, where a significant P value after Bonferroni correction for the number of traits tested implied an enrichment of differentiated eQTLs for the GWAS trait relative to all significant eQTLs in Sardinia (Supplementary Table 14). We repeated these analyses using different thresholds for differentiation (ΔAF_{SARD–EUR} > 0.05, 0.10, 0.15, 0.20, and 0.25) (Supplementary Fig. 7).

Identifying heritable patterns of outlier gene expression. For the 61 trios in our study, we developed a generalized likelihood-ratio test that identifies extreme gene expression signatures that are shared by one parent and their child (a full derivation of the test is given in the Supplementary Note). In practice, we ran our outlier pipeline on the same PEER-normalized data as we did for the eQTL analyses; we tested another normalization pipeline to determine whether PEER was overcorrecting outlier signals but found fewer results overall (instead of using PEER, we regressed out covariates that were highly correlated with PEER factors, as described in Supplementary Table 27). We tested each trio for all 15,243 expressed genes used in the eQTL analyses and evaluated significance via permutation, selecting the most significant trio for each gene and applying the Benjamini–Hochberg adjustment. For all genes with an outlier trio at a 10% FDR, we compared allele-specific expression in the outlier individuals with allele-specific expression in the rest of the participants (non-outliers). We next identified rare variants shared by outlier parents and children in the 250-kb window of the outlier gene and measured the relative enrichment of these variants with similarly identified variants in non-outlier individuals; confidence intervals were calculated via bootstrap resampling (B = 1,000) of all observed shared rare variants. Shared rare variants were annotated with chromatin state annotations from peripheral blood mononuclear cells (E062) from the Roadmap Epigenomics Consortium (Supplementary Table 28). Log odds scores and confidence intervals were calculated using Fisher’s exact tests for all functional annotations (Supplementary Tables 19 and 20). We then tested whether the effect of these shared rare variants on expression replicated in the larger study cohort (where there were at least four carriers in the population).

Clinical relevance of candidate causal rare variants. We prioritized 30 of the shared rare variants as candidate causal variants on the basis of several annotations (for example, variants with proximity to the TSS, that were highly conserved or deleterious, or that were potential splicing variants) (Supplementary Table 21). Five of these were associated with genes near significant GWAS loci, and three were associated with genes previously implicated in manifestation of the clinical traits available to us for study. We tested these rare variants for association with the complex traits or disease they were predicted to affect. For categorical traits (for example, celiac disease and bipolar disorder), we performed a likelihood-ratio test comparing two nested logistic regression models with the full model (genotype at the rare variant locus, sex, age, and age²) and the reduced null model (without the above covariates). Empirical P values were computed by permuting sample genotypes 1,000 times. To test rare variants for continuous traits (for example, BMI), we ran the lmekin function from the kinship R package to perform a likelihood-ratio test comparing two nested linear mixed models with the full model (genotype at the rare variant locus, sex, age, and age²) and the reduced null model (without the above covariates).

We then calculated the Pearson correlation between outlier gene expression and the adjusted trait data and calculated the correlation of gene expression with each clinical trait for each outlier gene–trait association; significance was assessed as the percentile of the empirical distribution obtained from the P values for all tested genes (Supplementary Tables 23 and 24).