Enhancing genetic discoveries with population-specific reference panels

Sanna, Serena

university of groningen

Enhancing genetic discoveries with population-specific reference panels

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 9 May 2016 at 16.15 hours

by

Serena Sanna

born on 15 November 1980
in San Gavino Monreale, Italië

Supervisors
Prof. C. Wijmenga
Prof. L.H. Franke

Assessment Committee
Prof. C.M. van Duijn
Prof. P. van der Harst
Prof. H. Snieder

Contents

ISBN printed version - 978-90-367-8822-9

ISBN online version - 978-90-367-8821-2

Chapter 1: General Introduction

1.1 Genetic dissection of complex traits and diseases

Complex or multifactorial traits and diseases are, by definition, the result of a combination of multiple genetic and environmental factors (such as lifestyle choices or risk factors exposure). The genetic components were mostly unknown for the majority of the complex diseases before the completion of the International HapMap Project in 2003 [*The International HapMap Consortium, Nature 2003*]. Technological advances benefiting from this large collaborative biological project allowed genetic studies in hundreds to thousands of individuals and assessment of 100,000–1,000,000 single nucleotide change variants (SNVs) in each of the individuals being studied, with an approach known as genome-wide association study (GWAS). The number of assessed variants was very large compared to previous approaches, such as linkage mapping, which typically used <10,000 variants to survey the entire human genome by identifying stretches of chromosome inherited from a common ancestor. In GWAS, geneticists look at one single variant at a time and evaluate whether there is any statistical correlation with the number of changed nucleotides (alleles) carried by each individual and their respective phenotypes. The power to find an association strictly depends on three factors: i) whether one tests the true causal variant or a variant in linkage disequilibrium (LD) with it, ii) the number of samples studied and iii) the impact of the true causal variant on the phenotype. The third factor is obviously unknown *a priori* and it's fixed by the underlying polygenic model of the disease. One trait can be modulated by a few highly impacting variants, or by many small impacting variants or a combination of both. After the first successful GWAS carried out using the available resources [*Klein et al, Science 2005; Menzel et al, Nat Genet 2007; Scuteri et al, Plos Gen 2007; Scott et al, Science 2007*], it became clear that most of the traits are modulated by a large number of variants with small to moderate effects, and more individuals and more variants were needed to achieve sufficient power for discovery. In this context, statistical geneticists have developed a new approach called genotype imputation to estimate with great precision the effects of many variants that are not directly genotyped with a specific genotyping array technology. This approach allowed researchers to increase both the number of variants and the number of individuals tested in a GWAS. In fact, genotype imputation permits to statistically infer and assess all 2 million variants catalogued in the International HapMap Project that were not directly genotyped. As all GWAS studies could therefore be aligned to a common reference set of SNVs, they could be compared and jointly analyzed in meta-analyses to virtually assess hundreds of thousands of samples across the world at no additional cost.

1.2. The genotype imputation method

The term genotype imputation indicates the process of predicting (or imputing) genotypes that are not directly assayed in a sample of individuals. Genotype imputation most often refers to the situation in which a reference panel of haplotypes characterized at a dense set of SNPs is used to impute into a study sample of individuals that have been genotyped at a sparse, subset of the SNPs. The fundamental idea is that short stretches of haplotypes can be shared even between unrelated individuals from distant common ancestors. Common stretches between the study samples and the reference samples can be identified using genotypes for a given set of shared SNPs, and alleles for SNPs that are measured only in the reference panel can be imputed. In a typical scenario, the study sample is genotyped with a commercial genotyping platform for hundreds of thousands to millions of SNPs located across the entire genome while the

reference panel contains haplotypes characterized for several millions of SNPs. An overview of this process is given in **Figure 1**. More than a few different statistical descriptions of genotype imputation procedures have now been published and implemented in a number of software packages, for example:

MACH/minimac: http://genome.sph.umich.edu/wiki/Minimac

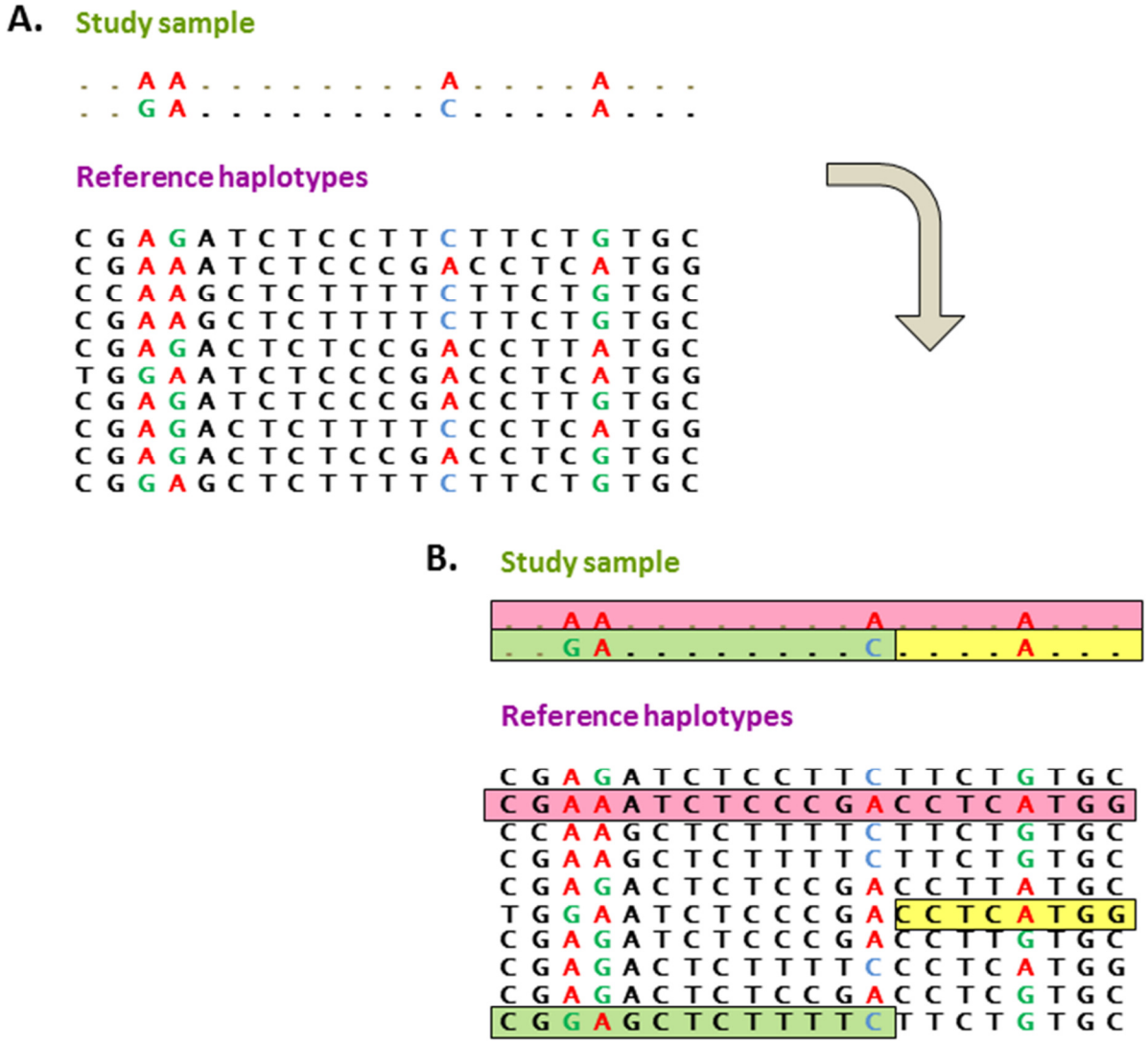
Beagle: https://faculty.washington.edu/browning/beagle/b3.html

IMPUTE/IMPUTE2: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

These tools typically provide convenient summaries of the uncertainty surrounding each genotype estimate. The imputation quality is commonly measured with a parameter called Rsq (also indicated with r2 or RSQR), i.e., the estimate of the squared correlation between imputed and true genotypes or, in other words, the ratio of the variances of imputed and true allele counts. In this context, it should be noted that the accuracy of predicted Rsq values is, in general, high for common variants, but rapid performance degradation is seen for lower minor allele frequencies, thereby limiting the applicability of such methods, especially for rare variants. The performance depends on multiple factors, including: choice of baseline array, quality of input genotypes/haplotypes and limited representation of reference haplotypes carrying rare alleles. Also and very importantly, differences in LD patterns and allele frequency spectrum significantly decrease the quality of imputation overall [*Li et al, Annu Rev Genomics Hum Genet 2009; Pistis et al, Eur J Hum Genet 2014; Porcu et al, Curr Protoc Hum Genet 2013*] .

Figure 1. Schematic representation of the genotype imputation method.

Panel A illustrates the genotypes at a restricted number of genetic markers in a sample being studied and at a larger number of markers in a reference panel of haplotypes. Panel B illustrates the process of identifying regions of a chromosome shared between a study sample and individuals in the reference panel. For chromosome 1 of the study sample, nucleotide configuration perfectly matches one reference haplotype (yellow), whereas for chromosome 2 there is not a perfect match. The chromosome could be reconstructed as a mosaic of short pieces which have recombined at a certain position. Multiple pieces and recombination points could be identified in the reference haplotypes set to reconstruct chromosome 2; the most likely arrangement (green and yellow) is selected based on several parameters including haplotype's frequency and the recombination rate map. In Panel C, observed genotypes and haplotype sharing information have been combined to fill in unobserved genotypes in the study sample for the most likely configuration. Figure adapted from Li et al, Annu Rev Genomics Hum Genet. 2009.

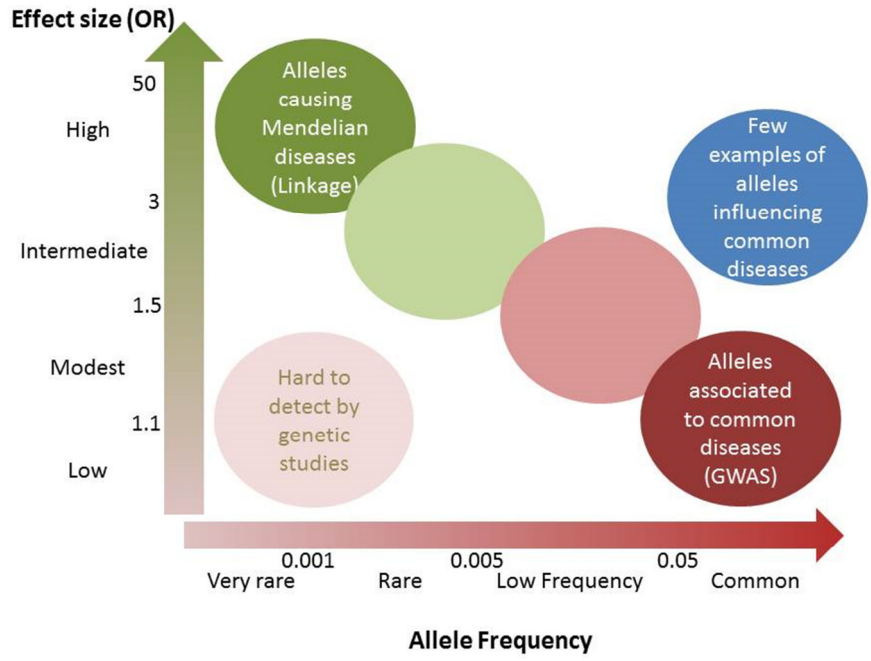

A.
Study sample
Reference haplotypes
B.
Study sample
Reference haplotypes
C.
Study sample
Reference Haplotypes

1.3 Insights from first-era GWAS studies and current strategies.

From 2005 to December 2013, 1,096 GWAS studies were carried out for several types of complex traits and diseases (according to GWAS catalog, www.genome.gov/gwastudies). Those studies revealed that, with only few exceptions, the genetic component of complex traits and diseases is fractioned into multiple variations of moderate or small impact (effect size) rather than a few with large effect size. The number of DNA variants contributing to the polygenic liability of a disease or trait's variation can be extremely large even for highly heritably traits. For example, in the past decade GWAS studies have identified up to 697 independent SNVs that influence human height, but their global contribution explains only 20% of the estimated heritability [*Wood et al, Nature Genetics 2014*]. Therefore many more associated variants have yet to be found for human height. The scenario is similar for other complex traits: despite the many SNVs identified, a substantial fraction of the heritability remains unexplained [*Manolio et al, Nature 2009*]. What is the cause of this "missing heritability"? Previous genome-wide association studies have assessed hundreds of thousands of individuals, but they analyzed 3 million of SNPs - only a small fraction of all possible variations present in the population. By using the HapMap project data scaffold, they focused mostly on common SNPs (frequency in the population >5%) catalogued within the Project, and low frequent (<5%) and rare variations (<1%) have therefore been largely unexplored (**Figure 2**). They also completely ignored any other type of genetic variation that is not a single nucleotide change, as for example insertions, deletions, copy number variations or de-novo mutations.

Figure 2. Genetic variants linked to diseases by allele frequency and effect size

The figures shows that GWAS and linkage approaches have mostly identified variants at the two extremes of frequency and effect size distributions. Sequencing based genetic studies are needed to target low-frequency and rare variation with moderate/high effect size. By contrast, very rare variants with small effect sizes are unlikely to be found even with current genetic approaches. Figure adapted from Manolio et al, Nature 2009.



Much more complete extraction of genetic variation is now accessible using next-generation sequencing (NGS) technologies. Still, efficient detection and analysis of rare and low frequency variants requires

sequencing hundreds to thousands of individuals and could be very expensive and so unfeasible for the majority of the research groups. There are however special designs which provide ideal settings to study variants in this frequency range and in a cost-effective manner. The simplest approach is to carry out a second step of genotype imputation replacing the HapMap Project with more complete, publicly available, imputation panels, as the 1000 Genomes Project [*1000 Genomes Project Consortium, Nature 2010; 1000 Genomes Project Consortium, Nature 2015*], an international project that provided a global reference for human genetic variation by sequencing the whole-genome of 2,504 individuals from four continents (America, Europe, Asia, Africa). The completion of the project, announced in September 2015, yielded the identification of a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms, 3.6 million short insertions/deletions (indels), and 60,000 structural variants (SVs)), all phased onto high-quality haplotypes that can be used for imputation [*1000 Genomes Project Consortium, Nature 2015*]. It therefore represents the most complete catalogue of genetic variation present to date. In particular, imputation with the complete version of 1000 Genomes haplotypes (phase 3) in existing GWAS studies will allow to assess both SNPs and structural variants, the latter category being largely ignored previously. The Structural Variants Analysis group catalogued the number of SVs in strong LD with known GWAS hits and found that GWAS haplotypes are enriched up to threefold for common SVs, which emphasizes the relevance of ascertaining SVs in disease studies.

A more complex approach is to sequence a subset of individuals from the population under study to build a population-specific panel for imputation. The study design including coverage, number and selection of samples, and the population being studied, have to be carefully evaluated to assure a balanced cost-benefit ratio. For rare and low frequency variants, genetically homogeneous populations represent an ideal scenario as rare and less frequent alleles may have raised in frequency due to genetic drift or selection.

1.4 Ideal settings for low frequency and rare variants mapping

Rare (MAF 0.5-1%) and low frequency variants (MAF 1-5%) require several thousands of individuals to be properly analyzed. In fact, many chromosomes have to be sequenced to detect a sufficient number of allele copies. Homogeneous populations, and especially isolates, represent a natural setting to overcome those limitations. In fact, some of the variants rare in the general population may have drifted to higher frequency, and some may even exist only in one isolate. Although associations with population-specific variants may initially appear useless to other populations, they can be useful to explain part of the missing heritability of complex traits, and moreover, to better understand the biological mechanisms underlying a complex trait variation or the etiology of a disease, and consequently suggest novel potential drug targets. In addition, extended and well ascertained pedigrees are generally available in studies on isolates which are particularly informative even for variants that are rare in the isolate itself. In fact, there is an increased chance to observe the same rare variant in more chromosomes segregating through families than in a study on unrelated individuals or small families, typical of outbred cohorts. Likewise, family-based designs can better control both genetic and environmental background and are robust to heterogeneity and population stratification. In fact, the geographically delimited area which population isolates usually live in assures restricted variability due to sharing lifestyle, sanitary conditions and exposure to pathogens, and the within-families transmissions minimize bias from population structure.

Finally, homogeneity of the population can improve genotype imputation accuracy when using a population-specific panel. The inference method in fact relies on shared stretches of chromosomes

between the study samples and a reference set of haplotypes. Shared stretches of haplotypes are expected to be longer in isolates than in open populations due to recombination within a restricted pool of variation [*Zavattari et al, Human Molecular Genetics 2000*]; therefore complexity in haplotype reconstruction is reduced.

1.5 Population isolates in genetic studies.

Population isolates are not all equal, and their different characteristics can influence the outcome of genetic studies. Population isolates vary in terms of size, time since foundation and demographic history. For example, Sardinians and Finnish are macro-isolates, counting 1.6 and 6 million inhabitants, respectively, and they contrast micro-isolates like the Amish community (~250,000 individuals) and Icelanders (~320,000 individuals)[*Zeggini, Springer 2015*]. Sardinians are also a very old isolate population (estimates of first settlement are ~10,000 years ago [*Sondaar et al, Comp Rend Acad Sci Paris 2015; Tykot et al, 1994*] compared to Finnish (~2,000 year ago)[*Kittles et al, Am J Hum Gen 1998*] and Icelanders (~1,000 years ago)[*Helgason et al, Am J Hum Gen 2000*]. Also the number of initial founders, the population growth rate and the variation of this rate over time are important to determine the amount of variability present at the settlement and the role of evolutionary mechanisms in modifying it. In fact, isolates originated by a main founder event show a substantially homogeneous genetic background with a highly reduced pool of variation, and isolates that experienced different waves of internal migrations with multiple bottlenecks and multiple founder events can reveal significant fine-scale substructure that needs to be accounted for in genetic studies.

A large pre-Neolithic settlement has been suggested in Sardinia. The island was inhabited by ~300,000 individuals during the Bronze Age and the population size did not significantly increase until around 300 years ago, experiencing very low immigration rates [*Francalacci et al, Science 2013*; *Sidore et al, Nature Genetics 2015*]. Consequently, the Sardinian population preserved a higher inter individual variability while maintaining a substantial genetic homogeneity. Being a macro-isolate, it offers the possibility to easily collect large cohorts characterized by a significant inter individual variability, while maintaining a reduced genetic substructure and it therefore provides an ideal setting to study rare and low frequency variation.

The SardiNIA study is the largest population study existing in Sardinia. The project started in 2001 and recruited 6,921 Sardinians (age 14-102 older), from a cluster of four towns in the Lanusei Valley of the Ogliastra region: Arzana, Elini, Ilbono and Lanusei, corresponding to approximately 62% of the population eligible in the area for recruitment [*Pilia et al, Plos Genetics 2006*]. The samples can be grouped in >1000 families, up to 5 generations deep; the largest family has more than 625 genotyped individuals. All volunteers have been characterized for more than 800 quantitative traits, including anthropomorphic measures, plasma and serum markers (such as cholesterol and other biomarkers for cardiovascular disease), personality traits (using the five-factor model), as well as deep characterization of the immune system through assessment of different cell types by means of fluorescence-activated cell sorting (FACS)[*Naitza et al, Plos Genetics 2012; Orrù et al, Cell 2013*]. My thesis reports successful studies carried out in this cohort.

1.6 Large-scale genetic studies in open populations

Isolated populations offer an intrinsic gain in power to test association at variants that have risen in frequency due to founder effects, drift and selective forces. For the same reasons, however, all those

variants that were not present in the initial pool of haplotypes existing at the time of the settlement or that were initially too rare and have been lost over generation, are absent in the present-day people. Therefore, isolated populations are limited in the number of variants assessable for association. From this perspective, large-scale studies of several thousands of individuals in open populations are necessary to study the full spectrum of rare variation. Very large sample sizes can be reached with collaborative efforts among different centers and the potential applications of the data can be maximized by the creation of biobanks. Furthermore, as the rare variants have, on average, relatively recent origin, they are more prone to fine-scale structure than common variants, showing higher frequency in specific geographical areas. For this category of variants, it is possible to observe different genotype patterns than that observed in the wider continental pool, and population-specific reference sets thus remain valuable for better quality imputation. Therefore large-scale studies in open populations will still benefit from population-specific imputation panels. Small scale studies instead will lack statistical power to assess rare variants even if directly genotyped. There are two clear examples of successful efforts that combined large biobanks with ad-hoc reference panels. In the Netherlands, the BBMRI-NL biobank (Biobanking BioMolecular Resources and Research Infrastructure of the Netherlands), within the European initiative BBMRI-ERC, provides a systematic database of collections of biomaterial and associated data from about 200 major clinical and population cohorts like LifeLines, Netherlands Twin Registry, Leiden Longevity, and the Rotterdam Study. To enhance the value of the BBMRI-NL, the GoNL (Genome of the Netherlands Project) has sequenced the whole-genome of 250 Dutch parent-offspring families and set up a population-specific reference panel for imputing samples from the Dutch Biobanks [*Boomsma DI, et al Eur J Hum Gen 2014; Genome of the Netherlands Consortium, Nat Genet 2014*]. The combination of these big resources has been successful in identifying rare variants associated to variation in lipid levels [*van Leeuwen et al, Nature Comm 2015*]. Similar efforts are ongoing in the UK with the UK BioBank and UK10K Consortium. The UK Biobank is a very large and detailed prospective study with over 500,000 participants aged 40–69 years when recruited in 2006–2010. The study has collected and continues to collect extensive phenotypic and genotypic detail about its participants, including data from questionnaires, physical measures, sample assays, genome-wide genotyping and longitudinal follow-up for a wide range of health-related outcomes [*Sudlow C et al, Plos Med 2015*]. The UK10K Consortium has set up a reference panel, that can be used for imputing the UKBiobank, by sequencing the whole-genome of ~4,000 British volunteers [*UK10K Consortium, Nature 2015*]. The first scientific reports of the combined resources have highlighted novel rare and low frequent variants associated to variation in several quantitative traits [*UK10K Consortium, Nature 2015*].

1.7 Limitations of the genotype-imputation approach

The genotype imputation approach is an extraordinary cost-effective strategy, especially in the era of whole-genome sequencing. By increasing the haplotypes in the reference set and using a set of chromosomes that perfectly matches LD patterns of the population being studied, genotypes can be accurately inferred for the majority of low-frequency and rare variants. However, there are still many variants in this category that, despite being detected in the sequenced individuals, are poorly imputed in the genotyped samples and some have to be discarded for analyses. Therefore, the power to detect a variant – that strictly depends on the available sample size in GWAS for a fixed effect size - has to be further scaled for imputation quality. Moreover, imputation approaches were designed for non-overlapping, bi-allelic changes and the same algorithm has been used to impute structural variants or multi-allelic sites, by pretending that they are in different positions, when other overlapping sites exist, or pretending that each alternative allele is a different mutation when more than one alternative allele exists. While the algorithm can be modified to improve imputation accuracy at rare sites and structural variations, there is another

category of genetic modifications that is completely ignored by the approach: *de novo* mutations. A *de novo* mutation is an alteration in the genome that is present for the first time in one family member as a result of a mutation in a germ cell (egg or sperm) of one of the parents or in the fertilized egg itself. Such alterations can be identified in sequencing efforts that involve trios and have sufficient coverage. To date, only the Genome of the Netherlands Consortium was able to catalogue *de novo* events at a population scale, taking advantage of their family-based design for sequencing and of the medium coverage (~10x on average)[*Genome of the Netherlands Consortium, Nat Genet 2014*]. The role of *de novo* events in common complex traits and disease is largely unexplored; each mutation is likely to contribute only slightly to the overall heritability of a trait, but we cannot exclude that recurrent mutations in the same gene or cluster of genes may play important roles and further explain part of the missing heritability for certain complex traits and diseases.

1.8 Outline of thesis

This thesis consists of three parts. In the first part (Chapter 2 and 3), we showed the advantages of combining sequencing and genotype imputation in the isolate of Sardinia. We firstly investigated benefits in detecting traits-associated rare and low frequent variants in the worst case scenario: sequencing a reduced number of individuals and focusing on exons of already established loci. We selected 256 Sardinian individuals with extreme low-density lipoprotein cholesterol (LDL-C) levels – who were expected to be enriched for LDL-C associated genetic variants – and sequenced the exons of seven well known genes with the standard Sanger method. Discovered variants were either genotyped or imputed in a large sample of 5,524 Sardinians (from the SardiNIA study). The study revealed that at such loci better lead variants and/or additional independent variants exist, and accounting for their contribution to phenotypic variation doubles the estimates of the heritability explained at these loci compared to variants previously detected by HapMap-based GWAS. Our results also include a Sardinian specific rare variant associated with LDL-C, highlighting the benefit of sequencing in this and other isolated populations. Overall, this study provided insights about what extensive whole-genome sequencing efforts were likely to reveal for the understanding of the genetic architecture of complex traits and encouraged us to carry out future steps.

We undertook large scale whole-genome sequencing with the aim to set up a reference panel for imputation that would maximize genetic information. Considering the wide diversity of phenotypes measured in the SardiNIA cohort, we did not select individuals based on their phenotype but rather on their estimated genome sharing with other genotyped members in the family (http://genome.sph.umich.edu/wiki/ExomePicks) We sequenced up to 2,120 Sardinians using a low-coverage approach (4.16x on average) and genotyped the whole SardiNIA cohort with both genome-wide and custom arrays. Similar efforts are carried out in other populations including general Europeans, but we showed that the gain in accuracy was remarkably higher in Sardinians, especially at low frequency and rare variants (Chapter 3). We also demonstrated that the extended homogeneity of the population allows precise estimates of genotypes even when using arrays with a reduced genomic content, as for example MetaboChip or Affymetrix HumanCore.

In the second part of the thesis we took advantage of the whole-genome sequences reference panel to carry out GWAS analyses for several traits using sequencing based imputed genotypes (Chapter 4-7), including one GWAS that was carried out using only ImmunoChip and MetaboChip as genotype scaffold (Chapter 4).

In the first of the papers from this collection, the phenotypes represented an extra layer of novelty. It was in fact the first time that the immune system was so well characterized in a large cohort of healthy individuals. Taking advantage of large pedigrees, we were able to estimate the heritability of variation in hundreds of immune cell populations, showing that for some of them the inherited variability can be as high as height. In a genome-wide scan that assessed >8 million markers (genotyped with ImmunoChip and MetaboChip or imputed based on ~1000 Sardinian whole-genomes), we identified 23 independent variants that are responsible for at least 2% of the phenotypic variation of the associated immune traits (which was consistent with the estimated lowest detectable effect size based on statistical power calculations).

Up to this date, we sequenced and analyzed about 2000 Sardinian whole-genomes, data that allowed us to deeply inference the demographic history of Sardinia, and to quantify its isolation comparing it with other European populations (Chapter 5). This large amount of sequenced genomes was extremely useful for genotype imputation of the full SardiNIA cohort, which was intensively characterized with genome-wide arrays. We evaluated the power of this genotype- combined-with-sequencing and inference design, by performing genome-wide association scans in two worst case scenario: lipid levels, which were previously analyzed in very large samples (>188,000 individuals)[*Willer et al, Nature Genetics 2013; Teslovich et al, Nature Genetics 2010*], and blood inflammatory markers, which were analyzed in the same cohort using custom arrays to target low-frequency and rare variants at some specific loci. We identified fourteen signals, including two major new loci, for lipid levels, and nineteen, including two novel loci, for inflammatory markers. Of note, novel signals would have not being identified without the Sardinian sequencing panel. In fact, when repeating the analyses using 1000 Genomes for imputation, such signals were either below genome-wide significance or where not imputed because too rare or absent in the 1000 Genomes sequenced populations.

Using this highly informative map, we also carried out whole-genome association analysis for three different hemoglobin levels (Chapter 6). Those parameters are rarely measured in healthy cohorts, and there are no available studies today that have measured them in the same individuals. The Sardinians represent an ideal population to study hemoglobin variations. In fact malaria, which is one of the strongest selective pressures that have shaped the genetics of red blood cells indices, and, consequently hemoglobin levels, was endemic on the Island until a few decades ago. In the genome-wide scan we identified 5 novel signals, including 3 that are highly differentiated in frequency among other Europeans (frequencies 1%,10% and 0.7% in Sardinians versus 0%, 1% and 0.4%, respectively, in other Europeans) and one that is currently Sardinian specific.

Finally, we analyzed another well studied trait: height (Chapter 7). Recent meta-analyses have investigated up to 3 million relatively common polymorphisms in a very large number of samples: 253,288 individuals [*Wood et al, Nature Genetics 2014*]. However, this gigantic effort ignored low frequent and rare variants. In our GWAS we analyzed 6,307 Sardinian individuals but included all variants detected with sequencing; we were able to detect a novel signal pointing at a rare variant that is barely present outside Sardinia. Our GWAS also revealed other potentialities of complex traits mapping in isolates. Taking advantage of the families, we searched for parent-of-origin effects of variants in a previously known height-associated locus located in an imprinted region. This analysis, coupled with replication in other Sardinian and European cohorts, allowed us to identify the most likely candidate variant of the locus and to refine the mechanism of action.

We therefore demonstrated several advantages of population-specific reference panels in isolated populations, using Sardinians as example. We showed that such populations offer the possibility for cost-

effective designs to enhance accuracy in genotype imputation of rare variants, that their particular demographic history have shifted to high frequency rare variants with likely functional impact, that large pedigrees easy collectable allow the assessment of deviation from the classical additive model of inheritance.

In the last section (Chapter 8) we discuss other ongoing whole-genome sequencing efforts projected to build population-specific imputation panels in other isolates but also in more open populations. We review the benefits in terms of overall accuracy of estimated genotypes and present key examples of novel genetic discoveries yielded.

References

The 1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing*. **Nature** 467, 1061–1073 (2010)

1000 Genomes Project Consortium. *A global reference for human genetic variation.* **Nature** 526, 68–74 (01 October 2015) doi:10.1038/nature15393

Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, Francioli LC, Hottenga JJ, Laros JF, Li Q, Li Y, Cao H, Chen R, Du Y, Li N, Cao S, van Setten J, Menelaou A, Pulit SL, Hehir-Kwa JY, Beekman M, Elbers CC, Byelas H, de Craen AJ, Deelen P, Dijkstra M, den Dunnen JT, de Knijff P, Houwing-Duistermaat J, Koval V, Estrada K, Hofman A, Kanterakis A, Enckevort Dv, Mai H, Kattenberg M, van Leeuwen EM, Neerincx PB, Oostra B, Rivadeneira F, Suchiman EH, Uitterlinden AG, Willemsen G, Wolffenbuttel BH, Wang J, de Bakker PI, van Ommen GJ, van Duijn CM. *The Genome of the Netherlands: design, and project goals.* **Eur J Hum Genet.** 2014 Feb;22(2):221-7. doi: 10.1038/ejhg.2013.118. Epub 2013 May 29.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, Sanna D, Useli A, Urru MF, Marcelli M, Cusano R, Oppo M, Zoledziewska M, Pitzalis M, Deidda F, Porcu E, Poddie F, Kang HM, Lyons R, Tarrier B, Gresham JB, Li B, Tofanelli S, Alonso S, Dei M, Lai S, Mulas A, Whalen MB, Uzzau S, Jones C, Schlessinger D, Abecasis GR, Sanna S, Sidore C, Cucca F*. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny.* **Science**. 2013 Aug 2;341(6145):565-9. doi: 10.1126/science.1237947

Genome of the Netherlands Consortium. *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* **Nat Genet.** 2014 Aug;46(8):818-25. doi: 10.1038/ng.3021. Epub 2014 Jun 29.

Helgason A, Sigureth ardóttir S, Nicholson J, Sykes B, Hill EW, Bradley DG, Bosnes V, Gulcher JR, Ward R, Stefánsson K.. *Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland*. **Am. J. Hum. Genet.** 67, 697–717 (2000).

The International HapMap Consortium. *The International HapMap Project.* **Nature** 426, 789-796. 2003

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC. Dual origins of Finns revealed by Y chromosome haplotype variation. **Am J Hum Genet.** 1998 May;62(5):1171-9.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. *Complement factor H polymorphism in age-related macular degeneration*. **Science**. 2005 Apr 15;308(5720):385-9. Epub 2005 Mar 10.

Li Y, Willer C, Sanna S, Abecasis G. *Genotype imputation.* **Annu Rev Genomics Hum Genet.** 2009;10:387-406. doi: 10.1146/annurev.genom.9.081307.164242. Review.PMID:19715440 | PMCID:PMC2925172

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, et al. *Finding the missing heritability of complex diseases.* **Nature**. 2009 Oct 8;461(7265):747-53. doi: 10.1038/nature08494.

Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, Foglio M, Zelenika D, Boland A, Rooks H, Best S, Spector TD, Farrall M, Lathrop M, Thein SL. *A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15.* **Nat Genet.** 2007 Oct;39(10):1197-9. Epub 2007 Sep 2.

Naitza S, Porcu E, Steri M, Taub DD, Mulas A, Xiao X, Strait J, Dei M, Lai S, Busonero F, Maschio A, Usala G, Zoledziewska M, Sidore C, Zara I, Pitzalis M, Loi A, Virdis F, Piras R, Deidda F, Whalen MB, Crisponi L, Concas A, Podda C, Uzzau S, Scheet P, Longo DL, Lakatta E, Abecasis GR, Cao A, Schlessinger D, Uda M, Sanna S, Cucca F. *A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation*. **PLoS Genet.** 2012 Jan;8(1):e1002480. doi: 10.1371/journal.pgen.1002480. Epub 2012 Jan 26.

Orrù V, Steri M, Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A, Floris M, Mentzen WI, Urru SA, Olla S, Marongiu M, Piras MG, Lobina M, Maschio A, Pitzalis M, Urru MF, Marcelli M, Cusano R, Deidda F, Serra V, Oppo M, Pilu R, Reinier F, Berutti R, Pireddu L, Zara I, Porcu E, Kwong A, Brennan C, Tarrier B, Lyons R, Kang HM, Uzzau S, Atzeni R, Valentini M, Firinu D, Leoni L, Rotta G, Naitza S, Angius A, Congia M, Whalen MB, Jones CM, Schlessinger D, Abecasis GR, Fiorillo E, Sanna S, Cucca F. *Genetic variants regulating immune cell levels in health and disease.* **Cell**. 2013 Sep 26;155(1):242-56. doi:10.1016/j.cell.2013.08.041. PubMed PMID: 24074872.

Pilia G, Chen WM, Scuteri A, Orrú M, Albai G, Dei M, Lai S, Usala G, Lai M, Loi P, Mameli C, Vacca L, Deiana M, Olla N, Masala M, Cao A, Najjar SS, Terracciano A, Nedorezov T, Sharov A, Zonderman AB, Abecasis GR, Costa P, Lakatta E, Schlessinger D. *Heritability of cardiovascular and personality traits in 6,148 Sardinians.* **PLoS Genet**. 2006 Aug 25;2(8):e132. Epub 2006 Jul 1

Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A, Brennan C, Lai S, Miller MB, Marcelli M, Urru MF, Pitzalis M, Lyons RH, Kang HM, Jones CM, Angius A, Iacono WG, Schlessinger D, McGue M, Cucca F, Abecasis GR, Sanna S. *Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs.* Eur J Hum Genet. 2015 Jul;23(7):975-83. doi: 10.1038/ejhg.2014.216. Epub 2014 Oct 8

Porcu E, Sanna S, Fuchsberger C, Fritsche LG. *Genotype imputation in genome-wide association studies.* **Curr Protoc Hum Genet**. 2013 Jul;Chapter 1:Unit 1.25. doi: 10.1002/0471142905.hg0125s78.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN,

Tuomilehto J, Collins FS, Boehnke M. *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* **Science**. 2007 Jun 1;316(5829):1341-5. Epub 2007 Apr 26.

Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orrú M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, Cao A, Lakatta E, Abecasis GR. *Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.* **PLoS Genet.** 2007 Jul;3(7):e115.

Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, Kwong A, Ortega del Vecchyo VD, Chiang CWK, Bragg-Gresham J, Pitzalis M, Nagaraja R, Tarrier B, Brennan C, Uzzau S, Fuchsberger C, Atzeni R, Reinier F, Berutti R, Huang J, Timpson NJ, Toniolo D, Gasparini P, Malerba G, Dedoussis G, Zeggini E, Soranzo N, Jones C, Lyons R, Angius A, Kang HM, Novembre J, Sanna S, Schlessinger D, Cucca F, Abecasis GR. *Genome sequencing elucidates Sardinian genetic architecture and augments GWAS findings: the examples of lipids and blood inflammatory markers*. **Nature Genetics** 47, 1272–1281 (2015) doi:10.1038/ng.3368 Epub 15 Sept 2015

Sondaar, P.Y., Elburg, R., Klein Hofmeijer, G., Martini, F., Sanges, M., Spaan, A. and De Visser, J.A. (1995) The human colonization of Sardinia: a late-Pleistocene human fossil from Corbeddu Cave. **C. R. Acad. Sci.** Paris, 320, 145–150

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. *UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age.* **PLoS Med.** 2015 Mar 31;12(3):e1001779. doi: 10.1371/journal.pmed.1001779.

Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, et al. *Biological, clinical and population relevance of 95 loci for blood lipids*. **Nature**. 2010 Aug 5;466(7307):707-13. doi: 10.1038/nature09270.

Tykot RH, in Radiocarbon dating and Italian Prehistory, R. Skeates, R. Withehouse, Eds.(Accordia Specialist Studies on Italy, London, 1994), pp. 115-145.

UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Richards JB, Greenwood CM, Timpson NJ, Durbin R, Soranzo N. *The UK10K project identifies rare variants in health and disease*. **Nature**. 2015 Oct 1;526(7571):82-90. doi: 10.1038/nature14962. Epub 2015 Sep 14.

van Leeuwen EM, et al. *Population-specific genotype imputations using minimac or IMPUTE2.* **Nat Protoc**. 2015 Sep;10(9):1285-96. doi: 10.1038/nprot.2015.077. Epub 2015 Jul 30.

Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang HY, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, et al. *Discovery and refinement of loci associated with lipid levels.* **Nat Genet**. 2013 Nov;45(11):1274-83. doi: 10.1038/ng.2797. Epub 2013 Oct 6.

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen

J, Lo KS, Locke AE, et al. *Defining the role of common variation in the genomic and biological architecture of adult human height.* **Nat Genet**. 2014 Nov;46(11):1173-86. doi: 10.1038/ng.3097. Epub 2014 Oct 5.PMID:25282103 | PMCID:PMC4250049

Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, Loddo M, Eaves I, Mastio G, Todd JA, Cucca F. *Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection.* **Hum. Mol. Genet**. 9, 2947–2957 (2000)

Zeggini E, Morris A. *Assessing Rare Variation in Complex Traits, Design and Analysis of Genetic Studies*. **Springer** 2015. Doi:10.1007/978-1-4939-2824-8

Part I: Exploring the advantages in isolates of genotyping-combined-with-sequencing imputation approaches.

Chapter 2: Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability.

Based on: ***Sanna S****, Li B*, Mulas A*,Sidore C, Kang HM, Jackson AU, Piras MG, Usala G, Maninchedda G, Sassu A, Serra F, Palmas MA, Wood WH 3rd, Njølstad I, Laakso M, Hveem K, Tuomilehto J, Lakka TA, Rauramaa R, Boehnke M, Cucca F, Uda M, Schlessinger D, Nagaraja R, Abecasis GR. PLoS Genet. 2011 Jul;7(7):e1002198. doi:10.1371/journal.pgen.1002198.*

[,#] indicate equal contributions*

ABSTRACT

Complex trait genome-wide association studies (GWAS) provide an efficient strategy for evaluating large numbers of common variants in large numbers of individuals and identifying trait associated variants. Nevertheless, GWAS often leave much of the trait heritability unexplained. We hypothesized that some of this unexplained heritability might be due to common and rare variants that reside in GWAS identified loci but lack appropriate proxies in modern genotyping arrays. To assess this hypothesis, we re-examined 7 genes (APOE, APOC1, APOC2, SORT1, LDLR, APOB and PCSK9) in 5 loci associated with low-density lipoprotein cholesterol (LDL-C) in multiple GWAS. For each gene, we first catalogued genetic variation by re-sequencing 256 Sardinian individuals with extreme LDL-C values. Next, we genotyped variants identified by us and by the 1000 Genomes Project (totaling 3,277 SNPs) in 5,524 volunteers. We found that in one locus (PCSK9), the GWAS signal could be explained by a previously described low frequency variant, and that in three loci (PCSK9, APOE, and LDLR) there were additional variants independently associated with LDL-C, including a novel and rare LDLR variant that seems specific to Sardinians. Overall, this more detailed assessment of SNP variation in these loci increased estimates of the heritability of LDL-C accounted for by these genes from 3.1% to 6.5%. All association signals and the heritability estimates were successfully confirmed in a sample of ~10,000 Finnish and Norwegian individuals. Our results thus suggest that focusing on variants accessible via GWAS can lead to clear underestimates of the trait heritability explained by a set of loci. Further, our results suggest that, as prelude to large-scale sequencing efforts, targeted re-sequencing efforts paired with large scale genotyping will increase estimates of complex trait heritability explained by known loci.

INTRODUCTION

In the past few years, Genome-Wide Association Studies (GWAS) have identified hundreds of genetic variants associated with quantitative traits and diseases, providing valuable information about their underlying mechanisms (for a recent example, see [1]). More than 2,000 common variants appear associated with over 200 conditions (as reported by the NHGRI GWA catalog on 12/2010) and for a few, like age-related macular degeneration [2] and type 1 diabetes [3], these common variants already account for a large fraction of trait heritability. In contrast, for most complex traits and diseases, common variants identified by GWAS confer relatively small increments in risk and explain only a small proportion of trait heritability [4]. For example, for low-density lipoprotein cholesterol (LDL-C), GWAS based on up to ~100,000 individuals examined at ~2.5 million common variants [1,5,6], have identified 35 loci associated with trait variation, with some also involved in modulating the risk of cardiovascular diseases. Common variants at these loci are estimated to account for 12.2% of the variability in LDL-C levels, about one-fourth of its genetic variance [1]. Several hypotheses have been formulated about the nature of the remaining heritability of lipid levels and other complex traits [4,7], ranging from the potential role of copy number variation to contributions from a large number of common variants each with very small effects. In our view, common and rare variants that are poorly represented in common genotyping arrays might account for an important fraction of trait heritability. Ignoring these variants might not only preclude identification of important trait associated loci but also compromise estimates of heritability. Thus, fine mapping appears the logical next step after GWAS. Here, we have focused on seven genes located in 5 of the loci associated with LDL-C in our original GWAS for blood lipid levels (*APOE*, *APOC1*, *APOC2*, *SORT1*, *LDLR*, *APOB* and *PCSK9*) [5]. A sixth locus (corresponding to SNP rs16996148) that included a large number of genes and no obvious functional candidates was not further examined here. Together, the 5 SNPs identified in the original GWAS analyses of these 5 loci in >8,000 individuals (with follow-up genotyping of >10,000 individuals) explained only 3.1% of LDL-C variability. We set out to re-assess the contribution of these loci

to trait heritability by evaluating a broader spectrum of variants. To catalog genetic variation in these regions, we first sequenced the exons and flanking regions of the seven genes in 256 unrelated Sardinians [8], each with extremely low or high LDL-C, and in an additional 120 HapMap samples (parents from the 30 CEU and 30 YRI trios). To assess the effect of identified polymorphisms, we genotyped detected variants and additional variants selected based on an early release of the 1000 Genomes Project in a cohort of 5,524 volunteers from the SardiNIA project [8]. Our results show that at these five loci, a combination of rare and common variants, some novel and some previously identified, are associated with LDL-C, and that, taken together they double the variance explained by the common variants detected in GWAS.

RESULTS

To refine the contribution of five loci implicated by GWAS in the variability of LDL-C, we sequenced the exons and flanking regions of seven genes in 256 unrelated Sardinians [8] with LDL-C levels that were either extremely low (116 individuals, mean LDL-C=70.4±16.0 mg/dl) or high (140 individuals, mean LDL-C=205.9±19.6 mg/dl) (**Materials and Methods**), as well as an additional 120 HapMap samples (parents from the 30 CEU and 30 YRI trios). Observed heterozygosity per base pair per individual was 1.28×10^{-3} in the selected Sardinian individuals, 1.31×10^{-3} in the CEU and 1.99×10^{-3} in the YRI.

Sequencing identified 782 variants, all submitted to dbSNP and now included in dbSNP releases 130 and later (for a complete list see **Supplementary Table 1**). As expected, more variants were found in the HapMap YRI samples than in the HapMap CEU or in Sardinian individuals with extreme lipid levels (**Supplementary Table 2**). Overall, we observed a 2:1 trend for enrichment of rare variants (MAF <1%) in the high LDL-C group compared to the low LDL group, similar to the observation by Johansen and colleagues [9] (**Supplementary Table 3**), but this enrichment was only statistically significant for *APOB* (P = 0.03 using an exact test). To test for LDL-C association, we used logistic regression to compare individuals in the two categories, yielding 10 variants (in *APOE*, *APOC1*, *SORT1*, *APOB*, and *PCKS9)* with P <0.1 (**Supplementary Table 4**). Because of the modest number of sequenced individuals and because no signal reached significance after Bonferroni adjustment, we judged these initial association analyses – which focused only on sequenced samples and only at coding regions – inconclusive.

In addition to the loci discussed so far, our re-sequencing and genotyping effort also included *B3GALT4* and *B4GALT4,* two loci that approached genome-wide significance in our initial GWAS analysis (each with $5 \times 10^{-8} < p < 5 \times 10^{-6}$) [5]. SNPs in these loci did not reach genome-wide significance in two subsequent meta-analyses [1,6] and were not significantly associated with LDL-C in the data generated here (**Table 1, Supplementary Figure 1**). Because we have no evidence that these two genes are associated with LDL-C, they are not discussed further. Variants identified in the two genes have been also deposited in dbSNP.

To increase the power to detect association, we genotyped 5,524 individuals in the SardiNIA cohort [8] using the Metabochip (see **Materials and Methods)**. The chip included 285 variants newly discovered by sequencing, together with an additional 2,992 derived from an early analysis of 1000 Genome Project Pilot haplotypes (considering variants ± 250Kb from each gene). To further supplement the number of variants at each locus, we carried out two rounds of genotype imputation. First, we used haplotypes for 256 sequenced SardiNIA samples to impute genotypes for 554 SNPs that failed assay design or genotyping on the Metabochip. Second, using the haplotypes of 60 CEU samples from the 1000 Genomes Pilot, we successfully imputed an additional 5,066 variants [10] (**Materials and Methods** and **Supplementary Table**

5). After imputation, 8,897 SNPs were available for analysis, with an average minor allele frequency of 18% and an average imputation r^2 of 0.84 for 5,620 imputed SNPs (**Supplementary Table 5)**.

At three loci, *SORT1*, *APOB* and *LDLR,* GWAS-identified variants were very strong proxies for the best available association signal, with similar allele frequencies and r^2>0.88 **(Table 1, Figure 1A** and **Supplementary Figure 2**). In those three genes, the variant showing strongest association was non-coding and not in strong linkage disequilibrium (r^2 > 0.4) with any tested coding variant. The most strongly associated marker at the *SORT1* locus, rs583104 (p-value=$1.2x10^{-9}$) was in high LD (r^2 =0.77) with rs12740374 (p-value=$2.2x10^{-8}$), an intronic SNP in the *CELSR2* gene that alters the hepatic expression of the *SORT1* gene by creating a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site [11]. Both markers were genotyped, so that under the hypothesis that rs12740374 is the causal variant underlying this association signal, the modest difference in p-values may be attributable to statistical fluctuation.

At the remaining two loci, *APOE* and *PCSK9*, evidence for association peaked at low frequency (1-5%) variants not in strong linkage disequilibrium with the original GWAS signals. In both cases our analyses pointed to variants that were well studied in other contexts, but which are not included in typical GWAS panels or in the HapMap panel of European haplotypes commonly used to impute missing genotypes. Thus these variants were missed in previous GWAS analyses. In *PCSK9,* variant rs11591147, which leads to a non-synonymous R46L change in exon 1, was more strongly associated (*P*=$2.9x10^{-15}$, frequency (T)=0.037, effect=-12.9 mg/dl; **Table 1**) than GWAS variant rs11206510, a SNP ~10Kb upstream of the transcription start site of the gene (*P* =$5.7x10^{-7}$, frequency (C)=0.24, effect=-3.7 mg/dl) (**Figure 1C**). Furthermore, rs11591147 totally explained the GWAS association signal (association at GWAS variant rs11206510 became non-significant (p=0.999) when non-synonymous variant R46L / rs11591147 was included as a covariate, **Figure 1D**). This coding variant has previously been implicated in the regulation of blood lipid levels, including LDL-C, and in the susceptibility to coronary and ischemic heart disease [12,13]. At the *APOE* gene cluster, the strongest evidence of association was observed at the missense variant (R176C, also known as R158C [14]) rs7412 (*P* =$1.8x10^{-31}$, frequency (T)=0.037, effect=-18.8 mg/dl) (**Figure 1E**). This variant did not account for the previously reported GWAS signal; marker rs4420638 indeed remained significantly associated (*P* =$6.4x10^{-10}$) after adjusting for rs7412. The missense variants at *APOE* and *PCSK9* were not typed in the HapMap II data set, and were only recently added to genotyping arrays (Illumina 1MDuo). Thus they have not been assessed by any GWAS reported to date.

We next conditioned on the top association signal at each of the 5 loci and sought to identify additional independently associated variants. To declare statistical significance at secondary signals, we used a p-value threshold of $1x10^{-4}$; corresponding to an adjustment for 500 independent tests across the five regions examined. At *LDLR*, we found an independently associated rare missense variant (rs72658864 / V578A, *P*=$2.5x10^{-6}$ in the basic model, *P*=$3.9x10^{-6}$ in the conditional model, frequency (C)=0.005; effect=23.7 mg/dl) (**Table 1 and Figure 1B**). This variant appears to be specific to Sardinia (where we identified it in our SardiNIA cohort [8] by Sanger sequencing in 3/256 individuals with extreme LDL-C; by Illumina genotyping in 51/5,800 randomly ascertained individuals; and by Solexa sequencing of 505 individuals, unpublished data). It is absent in the HapMap data set, not detected in 280 Northern European individuals sequenced within the 1000 Genomes Project, and monomorphic in >10,000 Finnish [15,16] and Norwegian [17,18,19] individuals genotyped with the MetaboChip (**Materials and Methods**, **Supplementary Table 6** and **Supplementary Table 7**). Reassuringly, the variant was also observed, albeit with a lower frequency (0.00035), in TaqMan genotyping an independent sample of 5,661 Sardinians from different villages in Sardinia [20] (**Materials and Methods**). The change in lipid levels associated with this rare variant (23.7

mg/dl) is 4 times greater than that observed for the strongest associated common variant at the locus (5.7 mg/dl for rs73015013). At the *APOE* locus, we found a strong independent signal at non-synonymous variant rs429358 (C130R, also known as C112R [14]) (**Table 1 and Figure 1F**)(P=1.2 x10^{-12} in the basic model, P=5.8 x10^{-11} in the conditional analysis, frequency (C) = 0.071, effect=9.3 mg/dl), which, together with rs7412, defines the three major isoforms of *APOE (*ε2, ε3 and ε4) [14,21]. This variant strongly correlates (r^2=0.96) with the originally reported GWAS signal, rs4420638 (P=4.6x10^{-12}, frequency (G)=0.097, effect=7.8 mg/dl). So, at this locus, the initial GWAS analysis picked up one independent signal (a proxy of rs429358/C130R) but missed the strongest associated variant in the region (rs7412/R176C). There was no clear evidence for residual association after accounting for the two missense variants (**Supplementary Figure 3**). Interestingly, the frequency of the derived allele C at rs429358 was remarkably lower in Sardinia (freq=7%, see **Table 1**) than that observed in the Finnish and Norwegian individuals (see **Supplementary Table 7**) and several other European ancestry samples (freq ~20%)[22,23,24], resulting in a strikingly lower frequency of the ε4 haplotype (2.5% vs. 15%)[22]. Finally, at *PCSK9*, we observed a possible independent association at SNP rs2479415, in the non-coding region flanking the transcript (P=1.1x10^{-7} in the basic model, P=8x10^{-5} in the conditional model, frequency (T)=0.59, effect = -3.6 mg/dl) (**Table 1** and **Figure 1D**). This variant showed an independent trend also in ~10,000 Finnish and Norwegian individuals (one-sided P=0.055 after conditioning for rs11591147).

When the 5 GWAS SNPs were replaced by the 8 variants described here (1 each for *SORT1* and *APOB,* 2 for *APOE*, *PCSK9* and *LDLR*) the variance accounted for by those loci increased from 3.1% to 6.5%. Similar estimates were also obtained with ~10,000 Finnish and Norwegian individuals, where, on average, analysis of these 8 variants increased variance explained from 3.5% to 7.1% (**Table 2** and **Material and Methods**).

DISCUSSION

We conducted fine mapping of five loci associated with LDL-C at an unprecedented level of resolution. In particular, we sequenced individuals with extreme phenotype levels, and subsequently genotyped variants identified by us and by the 1000 Genomes Project in a larger sample. In a final step we also imputed additional variants in the region to account for limitations of genotyping assay design. At all but one of the loci, *APOB*, the most strongly associated variant was directly genotyped or sequenced, suggesting that our initial selection included the crucial variants. In three loci, we found strongly associated rare or low frequency variants – which (except for a variant in LDLR, which appears to be specific to Sardinia) had been extensively characterized in previous non-GWAS studies. In these cases, although the associated variants had been previously described, they had not been thoroughly examined in together with GWAS associated variants at the same loci – so that the relative contributions of GWAS identified SNPs and previously described variants remained unclear.

In summary, we observed that:

(a) At ,*SORT1* and *APOB* loci, association peaked at variants with similar effect size and frequency to the variants identified in GWAS;
(b) At the *LDLR* locus, in addition to confirming the GWAS signal, a rare variant with a large effect was found. This variant is currently unique to the island of Sardinia;
(c) At the *APOE* locus, an independently associated low frequency variant was identified. The signal was previously missed in GWAS because the variant was not included in the available

genotyping chips or in the HapMap reference panels. An independently associated common variant similar in frequency and effect size to the original GWAS signal was also identified.

(d) At the last locus, *PCSK9*, the GWAS signal could be explained by a low frequency coding variant not included in the available GWAS genotyping chips or in the HapMap reference panels. Furthermore, there was evidence for one other independently associated variant.

The strongest signals identified at *APOE* (both variants) and *PCSK9* (the top hit) are likely to be the causal variants underlying the association signals. For *SORT1*, the variant exhibiting strongest association appears to be in strong linkage disequilibrium with a recently proposed functional polymorphism. In contrast, biological interpretation for other associated variants remains unclear for the other identified polymorphisms and requires further studies.Our results lead to several important major conclusions. First, it is striking that prior LDL-C GWAS have often missed signals due to low frequency variants (in two of the loci examined here, we identified strongly associated variants with frequency 1-5% that were missed in the original GWAS, because they were untyped or missing on imputation panels and poorly tagged by nearby SNPs). Sequencing in individuals with extreme trait values, along with large-scale imputation and genotyping, provided a better evaluation of the contribution of these loci to variation in LDL-C levels. A similar design was recently used to fine-map loci associated with fetal hemoglobin levels, a trait for which three loci can now account for about half of total variance [25].

Second, we show that in one of the five loci we fine-mapped, a previously missed low frequency variant can account for the GWAS signal – consistent with the hypothesis that at least some GWAS signals will be due to disequilibrium with nearby low frequency or rare variants [26]. There is considerable debate on how frequently this scenario will occur [27]. Our observations are compatible with some of the arguments made on both sides of this debate [26, 27]. For example, in the case of *PCSK9*, a single low frequency variant explains the observed common variant association signal but did not appear to reduce the ability of the genome-wide association study to localize the functional element of interest. Furthermore, the effect of this variant was too small to be detectable in most linkage studies (including our own linkage analysis of >35,000 relative pairs in Sardinia). Further, a single low frequency variant (and not a cluster of variants) was sufficient to explain this association signal.

Finally, our results show that if estimates are based only on the common variation assessed through GWAS, heritability at identified loci is likely to be underestimated. A more complete dissection, including common, low frequency and rare variants (some of which will be population specific), dramatically increased the proportion of heritability associated with the 5 loci examined here, from 3.1% to 6.5%. Notably, the variance explained by each locus increased when a rare variant was found as a primary or secondary hit (*LDLR*, *APOE* and *PCSK9*), even when the top GWAS SNP highly correlates with the strongest observed signal (*LDLR* and *APOE*). By contrast, only slight improvements were observed at loci where the most associated marker highly correlates with the GWAS SNPs and there was no evidence for additional independent signals, even when the GWAS variant is unlikely to be functional (*SORT1* and *APOB*).

Genome-wide association studies have proven to be an extremely productive strategy for identifying regions of the genome associated with complex traits, often leading to unexpected insights into complex trait biology. A major efficiency of these studies is that, by focusing on a subset of variants that can be genotyped using array based platforms, they can conveniently and economically survey many common variants in large numbers of individuals. Our results emphasize the utility of these genome-wide studies in identifying trait association regions, but also emphasize that caution is needed when genome-wide study

results are used to quantify the overall contribution of a locus to trait heritability. In our opinion, and consistent with our results, accurate estimates of heritability will require more extensive examination of each identified locus.

Broadly, this observation is consistent with recent simulation studies [28] which explore, in the context of a dichotomous trait, the relationship between effect sizes observed at GWAS SNPs and at true causal variants for the same locus. These simulation studies suggest that, most of the time, effect sizes estimated from GWAS would be similar to true effect sizes but that, some of the time, effect sizes estimated from GWAS might substantially underestimate the true effect size – especially in a scenario where rare variants are more likely to be causal. In cases where the effect size was underestimated by GWAS variants, a noticeable increase in heritability ensues.

It is also interesting to note that the effect sizes estimated here for rare and low frequency variants (all >10mg/dl) are larger than the effect sizes of any of the common variants identified in GWAS studies. Effect sizes of even rarer alleles associated with familial hypercholesterolemia are even larger (see [29] for examples of *PCSK9* variants with effects >100mg/dl). This is consistent with the intuition that alleles with a large impact on LDL-cholesterol levels will be under strong natural selection and will, thus, be prevented from reaching high frequency in the population. Although rare and low frequency alleles with more modest impacts on LDL-cholesterol values are also likely to exist, we cannot detect them using available sample sizes and their detection must await studies of much larger sample sizes.

In conclusion, these results underline that the subsequent sequencing of the coding regions around GWAS associations in individuals with extreme values followed by large scale imputation and genotyping is an important step in assessing the contribution of associated genomic regions to trait heritability. If similar trends to those described here are observed at the remaining LDL-C associated loci, extending our approach described to all known LDL-C susceptibility loci could lead to an increase in the proportion of variance they explain from ~12% to ~24%, exceeding half of the genetic variance for this trait. Due to economic considerations, our sequencing efforts focused on the coding regions of each gene and only on genes that appeared very likely to be involved in lipid metabolism. In each locus, we augmented the set of discovered variants with variants discovered by the 1000 Genomes Project, but that will likely miss very rare as well as population specific variants. We expect that more extensive fine-mapping efforts that more comprehensively examine non-coding regions could identify additional trait associated variants. Ultimately, unbiased whole genome sequencing based association analyses might be required to fully explain the heritability of a trait like LDL-C, facilitating the comprehensive assessment of rare, population specific, and non-SNP variation. In the meantime, directed sequencing and large scale genotyping appears to be a promising approach.

MATERIALS AND METHODS

Ethics statement

All individuals studied and all analyses on their samples were done according to the Declaration of Helsinki and were approved by the local medical ethics and institutional review committees.

Samples description

The SardiNIA project is a population based study of aging-related traits that includes 6,148 related individuals from the Ogliastra region of Sardinia, Italy [8,30]. During physical examination, a blood sample was collected from each individual and divided into two aliquots, one for DNA extraction and the other to characterize several blood phenotypes, including lipids levels. Specifically, LDL-C values were derived using the Friedwald formula that combines HDL and total cholesterol levels. The Finnish and Norwegian individuals are Type 2 Diabetes patients and unaffected individuals collected from several studies. Specifically, Finnish studies are: Dehko 2D 2007 (D2D 2007), Dose Responses to Exercise Training (DrsEXTRA), Diabetes Prevention Study (DPS), FUSION stage 2 [15] samples (from ACTION LADA, D2D 2004, FINRISK 1987, FINRISK 2002, Health 2000, Savitaipale) and Metabolic Syndrome in Men (METSIM)[16]; Norwegian studies are: The Nord- Trøndelag Health Study (HUNT 2)[17,18] and The Tromsø Study (TROMSØ)[19]. Baseline clinic characteristics of the SardiNIA, Finnish and Norwegian studies are reported in **Supplementary Table 7**.

The independent Sardinian sample used for assessing the frequency of the rare variant at *LDLR* consists of 5,661 individuals belonging to 884 families enrolled in the SharDNA study [20], which recruited volunteers from a cluster of villages located in the Ogliastra region: Talana, Urzulei, Baunei,Triei, Seui, Seulo, Ussassai, Perdasdefogu, Escalaplano and Loceri. Observed heterozygotes were unrelated to those observed in the SardiNIA study by using demographic records to track origin of individuals up to 10 generations.

Sequencing

Sequencing of the 256 Sardinians and the 120 HapMap samples (parents from the 30 CEU and 30 YRI trios) was carried out at the University of Washington Genome Sequencing Center through the NHLBI Resequencing & Genotyping Service (Debbie Nickerson, PI). To select the 256 individuals to be sequenced, we adjusted LDL levels by age and sex and then identified individuals in the top and bottom 5% of the distribution (individuals under lipid-lowering therapy were not considered). Among those, we selected all unrelated individuals who had at least one sibling in the study and were genotyped with 500K or 10K arrays [28], to facilitate downstream follow-up and imputation analyses.

Among the 782 variants detected by sequencing, two loss-of-function variants were observed. However, these were identified only on HapMap samples (see **Supplementary Table 8**). A common in-frame insertion in *APOB* was observed in Sardinia and in HapMap CEU samples but was not associated with LDL-C after multiple testing adjustment (rs17240441, $P = 3.0 \times 10^{-4}$; see **Supplementary Figure 1C and 1D, Supplementary Table 8**). The observed heterozygosity per bp/per individual was 0.00128, 0.00131 and 0.00199 in Sardinia, CEU and YRI samples, respectively. Concordance rate of HapMap II and III phases genotypes with those obtained from Sanger sequencing was 99.63%, while a lower rate (98.1%) was observed with genotypes obtained from the low-pass sequencing 1000 Genomes Project (43 CEU and 42 YRI samples were common between the two datasets), indicating the slightly lower accuracy of next-generation sequencing technologies and in particular of low-pass sequencing approaches [31].

Genotyping

Genotyping was carried out with Metabochip arrays (Illumina), which were designed in collaboration with several international consortia [5,32,33] with the aim to fine map association loci detected through GWAS for a variety of traits. Part of the design included a set of wild-card SNPs chosen by individual research groups, and the SardiNIA study promoted the inclusion of all variants detected by sequencing individuals with extreme LDL-C values. In particular, assays were successfully designed for 285 of the 782 variants discovered by sequencing and 178 passed quality controls filters (some of those were polymorphic only in HapMap individuals, but we included all detected variant on the chip to assess heterozygosity on a large sample). Briefly, 3,277 variants were included on MetaboChip, and 1,868 passed quality checks. For a detailed description of markers discarded by each filter see **Supplementary Table 9**. Concordance rate of Sanger and Metabochip genotypes was 99.47% at QCed markers, evaluated comparing genotypes of the 256 sequenced samples.

Metabochip genotyping was performed using Illumina Infinium HD Assay protocol with Multisample Beadchip format, and GenomeStudio was used for genotype calling. All samples had a call rate >98%, and there was no evidence for mis-specified family relationships (evaluated using Relpair software [34]). We discarded markers if any of the following was true: a) call rate <95%, b) MAF=0, c) Hardy-Weinberg Equilibrium $P<10^{-6}$ or d) excess of Mendelian Errors (**Supplementary Table 9**).

A total of 5,524 Sardinian individuals were genotyped, of which 5,382 had lipid measurements available and were not under lipid lowering therapy. In the Finnish and Norwegian studies, a total of 10,823 samples were genotyped, of which 10,027 had LDL-c measurement available and were not under lipid lowering therapy.

Genotyping of the rare *LDLR* variant rs72658864 on the SharDNA samples was carried out using TaqMan single SNP genotyping assays (Applied Biosystems). Given the rarity of the variant, DNA of a known heterozygote from the SardiNIA project was included in each well plate to allow detection of intensities of both alleles. The genotype of this sample was called as heterozygote in all plates.

Imputation and Statistical Analysis

To better represent genomic variation, we merged genotypes from the 256 sequenced Sardinian samples with genotypes available from Affymetrix 500K [30] and/or Metabochip for all variants +/-2Mb spanning the gene's transcript. We then phased the haplotypes using MACH [10] and used this reference set of haplotypes to impute sequence variants in the rest of the cohort [35]. We then focused on variants within +/250Kb of the gene transcript. To further fine map the region, we used 120 haplotypes from the 60 CEU samples sequenced within the 1000 Genomes Project (June 2010 release of haplotypes based on March 2010 genotypes release) to impute variants outside the coding regions and flanking sequences targeted in our sequencing study. MACH software was used for imputation, with the same sized window used for the Sardinian-based imputation (+/-2Mb). The results obtained with these two rounds of imputation are identified in the text, as well in table and figure legends, as "Affy+Sanger" and "1000G", respectively.

For association, LDL levels were adjusted for age, age squared and sex, and the distribution of residuals was normalized using a quantile transformation. The association test was performed using Merlin (--fastassoc option), which uses a variance component framework to account for genetic correlation across family members [35,36].

Comparison of imputed genotypes with experimental genotypes, carried out on a set of 1,097 individuals that were genotyped with the 6.0 Affymetrix Arrays (unpublished data), showed that the average per genotype error rate between imputed and experimental genotypes was 3.7% and 4.1% for imputations based on 1000 Genomes and Sanger haplotypes, respectively.

In the Finnish and Norwegian studies we applied a similar strategy to analyze variants (rs547235 and rs562338 on *APOB*, rs2479415 on *PCSK9* and rs429358 on *APOE*) that were not included on Metabochip. We defined a set of reference haplotypes of the 60 HapMap CEU founders by merging genotypes from the 1000 Genomes project and those from our Sanger sequencing, using SNPs located +/- 2Mb of *APOB*, *PCSK9* and *APOE*. We then used this reference panel to carry out imputation and successively used imputed dosages for testing association with LDL-C. Association analysis was performed using the same trait transformation and covariates as in the SardiNIA study. Imputation and association tests were performed separately for Finnish diabetics (N=1,742), Finnish non-diabetics (N=5,678), Norwegian diabetics (N=1,171) and Norwegian non-diabetics (N=1,436). Results were then meta-analyzed using an inverse-variance method, which combines p-values from each study using weights proportional to the variance of the beta coefficient (effect) (**Supplementary Table 7**). A combined estimate of allele frequencies was obtained using the same weights.

Variance explained

We evaluated the variance explained by a set of markers by including all of them into the linear model in addition to the clinical covariates (age, age squared, gender), and by subtracting the variance explained by this model versus the basic model (only clinical covariates). Analyses were performed using the lmekin function in R kinship package which uses a variance component framework to account for genetic correlation across family members. Notably, variance is not purely additive across loci, thus heritability in **Table 2** has been calculated using all 8 SNPs (or 5 SNPs) in the model rather than adding values observed at specific loci (**Table 1**). For the Finnish and Norwegian samples, the LDL-C variance explained was calculated in each study group separately, and a combined estimate was calculated by weighting each study according to its sample size (**Table 2**).

Conditional analyses

We conducted conditional analyses to test for residual associations after accounting for a key SNP. The procedure consists of adding a SNP into the regression model as covariate and testing the effect of another SNP. Specifically, we performed this analysis by adding the strongest associated variant (key SNP) as covariate in order to test 1) whether that variant could explain the GWA association signal; and 2) if additional independent signals were present. For the latter analysis, a threshold of $P < 1x10^{-4}$ was used to declare significance, corresponding to a Bonferroni threshold for 500 independent tests. A graphical representation of association results from the conditional analysis is shown in **Figure 1B, 1D, 1F** and in **Supplementary Figure 2B**, **2D.**

URLs:

MACH software: http://www.sph.umich.edu/csg/abecasis/mach/;

HapMap project: http://www.hapmap.org/;

1000 Genomes Project: http://www.1000genomes.org/;

1000 Genomes Haplotypes for imputation:

http://www.sph.umich.edu/csg/abecasis/MACH/download/1000G-2010-06.html;

Locus Zoom: http://csg.sph.umich.edu/locuszoom/

R kinship package http://cran.r-project.org/web/packages/kinship/index.html

REFERENCES

1. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466,707-713
2. Chen W, Stambolian D, Edwards AO, Branham KE, Othman M, et al (2010). Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. Proc Natl Acad Sci U S A 07(16):7401-6.
3. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6):703-7
4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461, 747-53.
5. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40, 161-169.
6. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 41(1):56-65.
7. Cirulli, E.T. & Goldstein, D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6), 415-25.
8. Pilia G, Chen WM, Scuteri A, Orrú M, Albai G, et al. (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. Plos Genet 2, e132.
9. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42(8), 684-7
10. Li, Y., Willer, C., Sanna, S. & Abecasis, G.R. (2009) Genotype Imputation. Annu Rev Genomics Hum Genet 10, 387-406.
11. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 5;466(7307):714-9.
12. Cohen, J.C., Boerwinkle, E, Mosley T.H. Jr & Hobbs HH. (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 354(12), 1264-72 .
13. Benn, M.J., Nordestgaard, B.G., Grande, P., Schnohr, P. and Tybjaerg-Hansen A. (2010) PCSK9 R46L, low-density lipoprotein cholesterol levels, and risk of ischemic heart disease: 3 independent studies and meta-analyses. J Am Coll Cardiol 55(25), 2833-42

14. Hansena, P.S., Gerdesa, L.U., Klausena, I.C., Gregersenb, N. & Faergeman, O. (1994) Genotyping compared with protein phenotyping of the common apolipoprotein E polymorphism. Clin Chim Acta. 31, 224(2):131-7.
15. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316,1341-5.
16. Stancáková A, Kuulasmaa T, Paananen J, Jackson AU, Bonnycastle LL et al. (2009) Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men. Diabetes 58, 1212-21
17. Midthjell K, Krüger O, Holmen J, Tverdal A, Claudi T, et al. (1999) Rapid changes in the prevalence of obesity and known diabetes in an adult Norwegian population. The Nord-Trøndelag Health Surveys: 1984-1986 and 1995-1997. Diabetes Care 22, 1813-20.
18. Holmen J, Midthjell K, Kruger O, Langhammer A, Lingaas Holmen T, et al (2003) The Nord-Trondelag Health Study 1995–97 (HUNT 2): Objectives, contents, methods and participation. Norsk Epidemiol 13, 19-32 .
19. Joseph J, Svartberg J, Njolstad I, Schirmer H. (2010) Incidence of and risk factors for type-2 diabetes in a general population. Scand j Public Health 38:768-75.
20. Biino G, Balduini C L, Casula L, Cavallo P, et al. (2011). Analysis of 12,517 inhabitants of a Sardinian geographic isolate reveals that predispositions to thrombocytopenia and thrombocytosis are inherited traits. Haematologica 96(1), 96-101
21. Weisgraber KH, Rall SC Jr, Mahley RW. (1981) Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms. J Biol Chem. 256(17):9077-83.
22. Sing CF, Davignon (1985) J. Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. Am J Hum Genet 37, 268-85.
23. Nickerson DA, Taylor SL, Fullerton SM, Weiss KM, Clark AG, et al (2000). Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res 10(10), 1532-45
24. Stengård JH, Clark AG, Weiss KM, Kardia S, Nickerson DA, et al (2002). Contributions of 18 additional DNA sequence variations in the gene encoding apolipoprotein E to explaining variation in quantitative measures of lipid metabolism. Am J Hum Genet 71 (3), 501-517
25. Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, et al. (2011) Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nat Genet 42, 1049-51
26. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. (2010) Rare variants create synthetic genome-wide associations. PLoS Biol. 8(1):e1000294.
27. Anderson CA, Soranzo N, Zeggini E, Barrett JC. (2011) Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. 9(1):e1000580.
28. Spencer C, Hechter E, Vukcevic D, Donnelly P. (2011) Quantifying the underestimation of relative risks from genome-wide association studies PLoS Genet. Mar;7(3):e1001337.
29. Abifadel M, Varret M, Rabès JP, Allard D, Ouguerram K, et al.(2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia Nat Genet. 34(2):154-6
30. Scuteri A, Sanna S, Chen WM, Uda M, Albai G. et al. (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. Plos Genet 3(7), e115.

31. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. (2011) Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. May 4 [Epub ahead of print].
32. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N. et al (2009). Variants in MTNR1B influence fasting glucose levels. Nat Genet 41(1), 77-81
33. Preuss M, König IR, Thompson JR, Erdmann J, Absher D et al. (2010). Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. Circ Cardiovasc Genet 3(5), 475-83.
34. Epstein MP, Duren WL and Boehnke M. (2000) Improved inference of relationship for pairs of individuals. Am J Hum Genet 67, 1219-1231.
35. Chen W. and Abecasis GR (2007) Family-based association tests for genome-wide association scans. Am J Hum Genet 81, 913-926
36. Abecasis, G.R, Cherny, S. S., Cookson, W.O. & Cardon, L.R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30, 97 - 101
37. Pruim RJ, Welch RP, Sanna S, Teslovich TM et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 26(18):2336-7

Figure 1. Regional Association plots

Association results around *LDLR*, *PCSK9* cluster and *APOE*. In each panel, the box at left (A, C and E) shows the association results in the main analysis; and at right (B, D and F) the results after conditioning for the strongest associated variant, highlighted with a purple dot in both plots, and its name written at the top. Arrows highlight independent signals and the most associated SNP detected in the previous GWAS [5]. Each SNP is also colored according to its LD (r^2) in Sardinians with the top variant, with symbols that reflect genomic annotation as indicated in the legend. The rugs above indicate the position of the SNPs that were analyzed by direct typing (MetaboChip), or imputed by using haplotypes from sequenced samples (Affy+Sanger) or 1000 Genomes haplotypes (1000G). Plots were drawn using the LocusZoom standalone version [37].Genomic coordinates are given according to build 36 (hg18).

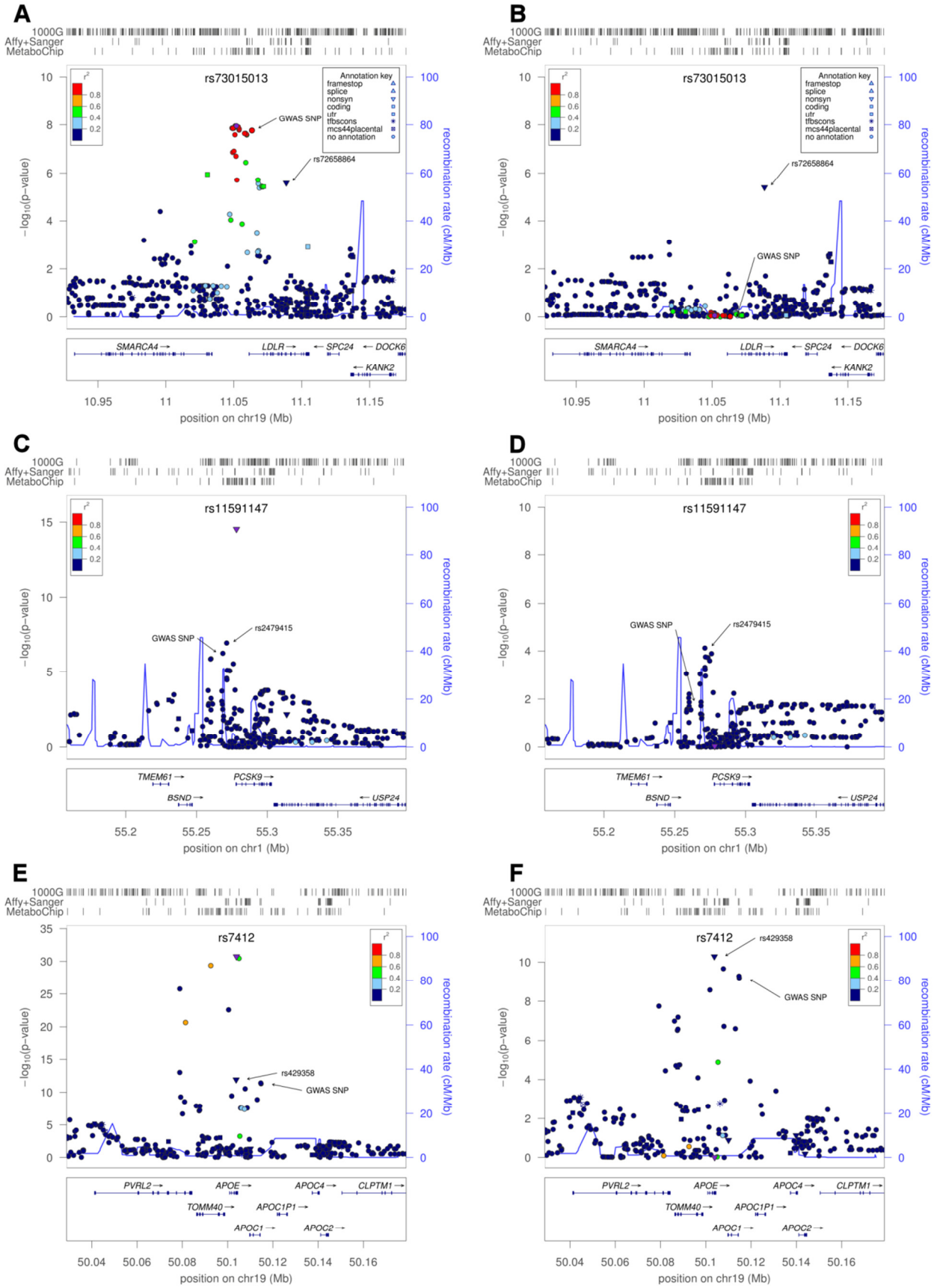

A
1000G
Affy+Sanger
MetaboChip
rs73015013
GWAS SNP
rs72658864
Annotation key
framestop
splice
nonsyn
coding
utr
tfbscons
mcs44placental
no annotation
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
SMARCA4
LDLR
SPC24
DOCK6
KANK2
10.95
11
11.05
11.1
11.15
position on chr19 (Mb)
B
1000G
Affy+Sanger
MetaboChip
rs73015013
rs72658864
GWAS SNP
Annotation key
framestop
splice
nonsyn
coding
utr
tfbscons
mcs44placental
no annotation
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
SMARCA4
LDLR
SPC24
DOCK6
KANK2
10.95
11
11.05
11.1
11.15
position on chr19 (Mb)
C
1000G
Affy+Sanger
MetaboChip
rs11591147
GWAS SNP
rs2479415
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
TMEM61
PCSK9
BSND
USP24
55.2
55.25
55.3
55.35
position on chr1 (Mb)
D
1000G
Affy+Sanger
MetaboChip
rs11591147
GWAS SNP
rs2479415
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
TMEM61
PCSK9
BSND
USP24
55.2
55.25
55.3
55.35
position on chr1 (Mb)
E
1000G
Affy+Sanger
MetaboChip
rs7412
rs429358
GWAS SNP
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
PVRL2
APOE
APOC4
CLPTM1
TOMM40
APOC1P1
APOC1
APOC2
50.04
50.06
50.08
50.1
50.12
50.14
50.16
position on chr19 (Mb)
F
1000G
Affy+Sanger
MetaboChip
rs7412
rs429358
GWAS SNP
r²
0.8
0.6
0.4
0.2
−log10(p−value)
recombination rate (cM/Mb)
PVRL2
APOE
APOC4
CLPTM1
TOMM40
APOC1P1
APOC1
APOC2
50.04
50.06
50.08
50.1
50.12
50.14
50.16
position on chr19 (Mb)

Table 1. Association Analysis results

The left panel shows the association results at 7 loci. For each gene, the strongest variant is listed first, and any second detected independent signal is listed with results from the conditional analysis (**Materials and Methods**). The column Type indicates whether the SNP was directly genotyped (Metabochip) or imputed using 1000G reference haplotype (1000G) or the Sardinian reference panel (Affy+Sanger). The right panel shows the association results for the GWAS SNPs previously described[5], the correlation with the top SNP listed in the left panel, and its p-value in the conditional analysis (Adjusted P-value).

Locus	SNPname	Type	Effect Allele / Other	Freq Effect Allele	Effect (SE)[a]	P-value	Genomic Annotation	Variance explained by the locus	Top GWAS SNP	Effect Allele / Other	Freq Effect Allele	Effect (SE)[a]	P-value	r2	Adjusted P-value	Variance explained by the locus
PCSK9	rs11591147	Metabochip	T/G	0.037	-0.380 (0.048)	$2.90x10^{-15}$	**missense (R46L)**	1.19 %	rs11206510	C/T	0.243	-0.106 (0.023)	$5.71x10^{-07}$	0.101	0.013	0.23%
	rs2479415	1000G	C/T	0.413	0.076 (0.019)	$7.50x10^{-05}$	8Kb from *PCSK9*									
SORT1	rs583104	Metabochip	T/ G	0.177	0.149 (0.024)	$1.28x10^{-09}$	31Kb from *SORT1*[b]	0.63%	rs599839	G/A	0.276	-0.148 (0.025)	$1.43x10^{-09}$	0.991	0.90	0.61%
B3GALT4	rs28361085	1000G	C/T	0.073	0.114 (0.036)	0.00169	146Kb from *B3GALT3*	0.22%	rs2254287	G/C	0.492	0.005 (0.018)	0.771	0.413	0.84	0.02%
B4GALT4	rs34507110	1000G	G/A	0.154	0.122 (0.030)	$4.99x10^{-05}$	83Kb from *B4GALT4*	0.48%	rs12695382	A/G	0.075	-0.074 (0.035)	0.035	0.795	0.48	0.03%
APOB	rs547235	1000G	A/G	0.187	-0.144 (0.024)	$1.69x10^{-09}$	140Kb from *APOB*	0.51%	rs562338	A/G	0.173	-0.139 (0.025)	$1.43x10^{-8}$	0. 878	0.98	0.43%
LDLR	rs73015013	Metabochip	T/C	0.138	-0.155 (0.027)	$1.12x10^{-08}$	9kb from *LDLR*	1.17%	rs6511720	T/G	0.132	-0.160 (0.027)	$1.71x10^{-08}$	0.934	0.97	0.59%
	rs72658864	Metabochip	C/T	0.005	0.626 (0.136)	$3.90x10^{-06}$	**missense (V578A)**									
APOC1/C2/E	rs7412	Metabochip	T/C	0.037	-0.563 (0.048)	$1.80x10^{-31}$	**missense (R176C) *APOE***	3.33%	rs4420638[c]	G/A	0.097	0.218 (0.031)	$4.67x10^{-12}$	0.0003	$6.41x10^{-10}$	1.07%
	rs429358	Affy+Sanger	C /T	0.071	0.260 (0.036)	$5.82x10^{-11}$	**missense (C130R) *APOE***									

a. Effect sizes are standardized (see **Materials and Methods**), and represent the change in trait LDL-C values associated with each copy of the reference allele, measured in standard deviation units.

b. SNP rs583104 is also 1Kb from *PSRC1* transcript

c. r^2=0.967 with Metabochip second-independent SNP, rs429358. After adjusting for the two independent SNPs, rs7412 and rs429358, the p-value for rs4420638 was 0.5

Table 2. Heritability estimates in all study samples

The table shows the LDL-C variance accounted for by the 5 GWAS SNPs and the 8 SNPs here described in all studies. A sample size weighted average estimate is given for the Finnish and Norwegian samples.

Study	**N samples**	**Variance explained by 5 GWAS SNPs**	**Variance explained by 8 SNPs**
SardiNIA	5,382	3.1%	6.5%
Norwegian T2D	1,171	5.8%	9.3%
Norwegian controls	1,436	3.1%	8.5%
Finnish T2D	1,742	2.1%	5.0%
Finnish controls	5,678	3.4%	7.0%
Average Finnish and Norwegian	*10,027*	*3.5%*	*7.1%*

Chapter 3: Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs

Based on:

Pistis G, Porcu E*, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A, Brennan, C, Lai S, Miller MB, Marcelli M, Urru MF, Pitzalis M, Lyons RH, Kang HM, Jones CM, Angius A, Iacono WG, Schlessinger D, McGue M, Cucca F[#], Abecasis GR[#], **Sanna S**[#].*

Eur J Hum Genet. 2014 Oct 8. doi: 10.1038/ejhg.2014.216

[,#] indicate equal contributions*

ABSTRACT

The utility of genotype imputation in genome-wide association studies is increasing as progressively larger reference panels are improved and expanded through whole-genome sequencing. Developing general guidelines for optimally cost-effective imputation, however, requires evaluation of performance issues that include the relative utility of study-specific compared with general/multipopulation reference panels; genotyping with various array scaffolds; effects of different ethnic backgrounds; and assessment of ranges of allele frequencies. Here we compared the effectiveness of study-specific reference panels to the commonly used 1000 Genomes Project (1000G) reference panels in the isolated Sardinian population and in cohorts of European ancestry including samples from Minnesota (USA). We also examined different combinations of genome-wide and custom arrays for baseline genotypes. In Sardinians, the study-specific reference panel provided better coverage and genotype imputation accuracy than the 1000G panels and other large European panels. In fact, even gene-centered custom arrays (interrogating ~ 200 000 variants) provided highly informative content across the entire genome. Gain in accuracy was also observed for Minnesotans using the study-specific reference panel, although the increase was smaller than in Sardinians, especially for rare variants. Notably, a combined panel including both study-specific and 1000G reference panels improved imputation accuracy only in the Minnesota sample, and only at rare sites. Finally, we found that when imputation is performed with a study-specific reference panel, cutoffs different from the standard thresholds of MACH-Rsq and IMPUTE-INFO metrics should be used to efficiently filter badly imputed rare variants. This study thus provides general guidelines for researchers planning large-scale genetic studies.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified thousands of common, single nucleotide polymorphisms (SNPs) associated with complex traits. However, existing genotyping arrays used in GWAS survey only a limited repertoire of sequence variation, and under-represent rare and population-specific variants. Much more complete extraction of genetic variation is now accessible using next generation sequencing (NGS) technologies, but efficient detection of rare and low frequency variants requires sequencing hundreds to thousands of individuals[1].

An alternative cost-effective approach to enlarge the frequency spectrum of variants assessed in GWAS capitalizes on publicly available sequencing reference panels, especially the 1000 Genomes Project (1000G) reference panels. Indeed, “probabilistic” sequenced genomes can be reconstructed by means of imputation methods, inferring untyped variants by combining partial haplotypes found in a study sample with the full haplotypes available in a more densely characterized reference set. It has, however, been unclear how well general reference panels represent variation in populations that were poorly or not at all represented in projects like 1000 Genomes. Furthermore, even for well represented populations, a complete evaluation is needed to assess the benefits of sequencing more study samples for successfully impute rare or low frequency variants.

How can imputation be further improved? Imputation works very well for common variants, but rapid performance degradation is seen for lower minor allele frequencies. The performance depends on multiple factors, including: choice of baseline array, quality of input genotypes/haplotypes, and limited representation of reference haplotypes carrying rare alleles. Also and very importantly, differences in linkage disequilibrium (LD) patterns and allele frequency spectrum significantly decrease the quality of imputation overall, especially when using public reference panels for ancestral or geographically isolated populations[2,3].

To investigate these factors, we compared imputation quality using three complementary sets of reference panels: 1,488 Sardinians from Sardinia, Italy; 1,325 individuals of Northern European ancestry from Minnesota, USA; and 1,092 individuals from the 1000 Genomes project. These reference panels permit comparison of the relative efficiency of study-specific imputation in founder (i.e., Sardinia) and continental (i.e., Northern European) populations that have also been genotyped, and contrast those results with the current standard approach (i.e., 1000 Genomes). Finally, we evaluated the efficiency of the conventional quality thresholds to discard poorly imputed rare and low frequency variants, focusing on metrics defined by the two most commonly used imputation software, MACH[4] and IMPUTE[5]. **Figure 1** shows a schematic representation of the study.

MATERIALS AND METHODS

Sample description and genotyping

The study sample consists of the SardiNIA and the MCTFR cohorts. Both studies were approved by the corresponding *institutional review boards and a signed informed consent was obtained from every volunteer.* The SardiNIA cohort comprises 6,921 individuals, representing >60% of the adult population of four villages in the Lanusei Valley in Sardinia. Details on the study have been previously described[6]. *The Minnesota Center for Twin and Family Research (MCTFR[7,8]) at the University of Minnesota specializes in the use of genetically informative family cohorts to investigate etiology of behavioral and psychiatric phenotypes. The MCTFR consists of two complementary cohorts. One is a population-based cohort of twins and their parents, the other is a family adoption study.*

The entire SardiNIA cohort was genotyped using the HumanOmniExpress GWAS array, containing ~750K markers, and three different Illumina custom arrays: the Cardio-MetaboChip, the ImmunoChip and the HumanExome, each containing about 200,000 markers[9,10]. Genotype calling was performed using the Illumina GenCall algorithm, and an additional 2,968 rare variants were called for HumanExome using Zcall[11]. A subset of 1,072 samples was also previously genotyped with Affymetrix 6.0[12].

After performing quality control checks (see **Supplementary Information** and **Table S1** for details), we used the quality checked (QCed) autosomal markers from the HumanOmniExpress, ImmunoChip and Cardio-MetaboChip arrays as baseline genotypes to impute variants detected through sequencing, as described below. In order to have fully comparable data sets for all analyses described here, we considered only the 6,602 samples for which all four Illumina arrays were

successfully genotyped. Data from the Affymetrix 6.0 array was instead not combined with the Illumina arrays, given the smaller number of samples available (1,072 vs 6,602); for this set quality control filters have been already described[13].

From the QCed set of markers we extracted a subset of 227,745 SNPs representing most of the content of the Illumina HumanCore array (78.9% prior QC), a low density genome-wide array. Given the extensive overlap, and considering that after quality control filtering the effective content of an array is always reduced, we treated this subset of markers as an approximation of the genomic content accessible with the HumanCore array, which we refer to here as "pseudo-HumanCore".

Genotyping protocols and quality control for the MCTFR study have been described previously[8,14]. In short, the full MCTFR study sample was genotyped with the Illumina 660W-quad array, with 7,278 (97.8%) samples and 527,829 (94.3%) markers passing quality control filters. The full sample was also genotyped with the Illumina HumanExome array, with 7,244 (97.4%) samples and 144,075 (58.1%) markers passing quality control filters. We initially used 6,610 individuals of European ancestry, and noticed that the inclusion of the 1,181 individuals who were also in the reference panel biased accuracy estimates at rare variants, due to perfect match of haplotypes (**Table S2**). We therefore restricted the analyses to the 5,429 samples not overlapping with the reference panel.

Sequencing and variant calling

Samples to be sequenced were selected in trios, taking advantage of their highly informative content for haplotypes reconstruction. Trios (or parent-offspring pairs for incomplete trios) were selected starting from the founders of all available families to assure the representation of all haplotypes that have been propagated within families (using ExomePicks, see **URLs**). For the Sardinians, 2,120 samples from 695 nuclear families were sequenced to an average coverage of 4.16-fold. Of those, 1,122 samples were part of the SardiNIA project[6], whereas the other 998 were individuals enrolled in case-control studies of Multiple Sclerosis and Type 1 Diabetes[15,16]. The sequencing effort has been described in part previously[17], and updated details are provided in **Supplementary Information**.

In the MCTFR study, 1,328 individuals from 602 families were sequenced to an average coverage of 10.4-fold. Three samples gave unacceptable sequence quality, leaving 1,325 total sequenced samples for analysis.

Variant calling was performed in both studies using GotCloud[18]. Sequencing yielded 17.6 and 27.1 million autosomal bi-allelic SNPs in Sardinians and Minnesota samples, respectively, of which 30.6% and 48.4% were not described in dbSNP135.

Genotype imputation

Genotype imputation for all scenario were performed on haploid data using Minimac (see **URLs**), a modified version of the MACH[4] software. For SardiNIA, phased haplotypes were generated using MACH (*--phase* option) with 400 states and 30 rounds by subdividing the variants in 344 groups of

2,500 with an overlap of 500, and imputation was subsequently performed independently on each phased chunk (for a description of the code, see the "1000G imputation cookbook" URL). Imputation performance was evaluated on seven different input genotype datasets: i) HumanOmniExpress (OmExp), ii) Cardio-MetaboChip (Metab), iii) ImmunoChip (Imm), iv) Cardio-MetaboChip and ImmunoChip (MetabImm), v) HumanOmniExpress, Cardio-MetaboChip and ImmunoChip (OMI), vi) pseudo-HumanCore (pHumCore), and vii) Affymetrix 6.0 (Affy 6.0).

For simplicity, we phased the Cardio-MetaboChip, ImmunoChip and HumanOmniExpress arrays jointly, and then extracted haplotypes at relevant SNPs to perform imputation for each particular genotyping set. In actual practice, Cardio-MetaboChip and ImmunoChip will be phased without the additional support of a genome-wide array, so we assessed the impact of our procedure by phasing separately each SNP set, for chromosome 20. We noticed that only imputations performed with the SardSeq panel or its combination with 1000G were slightly overestimated (see **Supplementary Information** and **Table S3**).

In the MCTFR study, haplotypes were phased using SHAPEIT2[19] (v2.644) with the following model options: --thread 8 --burn 10 --prune 8 --main 20 --states 200. Imputation was performed using Minimac and the Illumina *660W-quad array as baseline genotypes.*

We used as reference panels the 1000G-ALL (1,092 samples) and 1000G-EUR (379 samples) data sets from the 1000 Genomes March 2012 release; the full MCTFR sequencing data (1,325 samples, named MinnSeq in the text); a subset of the Sardinian sequencing data (1,488 samples, named SardSeq in the text); and combinations of those (see "***Combination of reference panels***" below). Considering the overall high inbreeding in Sardinia, the SardSeq reference panel was created by selecting only haplotypes of parents at each sequenced trios to avoid over-representation of rare variants.

We also performed imputation with IMPUTE2 (newest release of IMPUTE), to test a different approach for reference panels combination (see "**Combination of reference panels" paragraph) and t**o assess the efficiency of its imputation accuracy metric INFO (see "**Evaluation of imputation accuracy" paragraph).**

Simulation of European haplotypes

Because the Minnesota samples were genotyped with different arrays from those used for Sardinians, they could not be used to assess relative efficiency of arrays in genotype imputation. We therefore generated, by simulation with the HAPGEN[20] software and 1000G-EUR as reference, 6,602 unrelated individuals of European ancestry for SNPs present in each different genotyping array considered in the SardiNIA study. For simplicity, we focused only on chromosome 20. Haplotypes were phased using MACH (--*phase* option) with 400 states and 30 rounds, and imputation performed using Minimac, as in the SardiNIA and MCTFR data sets. This simulated data set was only used for assessing the efficiency of different genotyping arrays and reference panels in genotype imputation.

Combination of reference panels

We used VCFtools[21] to combine the SardSeq and the MinnSeq panels with 1000G-EUR and1000G-ALL reference panels for chromosome 20. The variants in each set were 331,799, 602,317, 851,702 and 377,494 for SardSeq, MinnSeq, 1000G-ALL and 1000G-EUR, respectively. During the merging procedure, we removed the variants present only in one panel, leading to SardSeq + 1000G-ALL, SardSeq + 1000G-EUR, MinnSeq + 1000G-ALL and MinnSeq + 1000G-EUR reference panels containing 249,624; 227,405; 304,899; and 267,550 variants, respectively. Imputation was then performed using Minimac, as for single reference panels. For combinations with 1000G and SardSeq panels, we also performed imputation with IMPUTE2 using the *--merge_ref_panels* option, which imputes variants unique to one panel into the other, prior imputation. We observed no difference in imputation accuracy at all frequency ranges when using this approach, which should be preferable for research studies, allowing imputation of all available variants, including those that are study-specific, in the same run (**Table S4**).

In addition, to assess the impact of adding a smaller number of population-specific haplotypes, we created two additional reference panels using 500 and 1000 randomly chosen samples from the SardSeq reference panel and merging them with 1000G reference panels (500SardSeq + 1000G and 1000SardSeq + 1000G, respectively). This analysis was restricted to the SardSeq panel and the SardiNIA cohort, because the advantage in accuracy was substantial for this population.

Evaluation of imputation accuracy

Imputation accuracy was assessed using both the MACH Rsq metric and the squared Pearson correlation (R^2)[4] between dosages and the real genotypes (considered as allele count) available for the same individuals, extracted from the HumanExome array. The Rsq metric is also known as variance ratio, being calculated as the proportion of the empirically observed variance (based on the imputation) to the expected binomial variance *p(1-p)*, where *p* is the minor allele frequency. In SardiNIA we tested 21,398 SNPs across autosomes for genome-wide evaluation of imputation accuracy and tested a subset of 558 SNPs for comparisons restricted to chromosome 20. For the MCTFR study, as the baseline array was different, we used a subset of 541 SNPs. The number of SNPs tested for comparing imputation with SardSeq versus 500SardSeq + 1000G and 1000SardSeq + 1000G was reduced to 517 because 41 SNPs (MAF range 0.0008% - 0.0072%) were not detected in the selected subset of sequenced samples.

We also assessed efficiency in discriminating between well and poorly imputed markers of the imputation accuracy metrics estimated by MACH (Rsq) and IMPUTE (INFO)[4]. The INFO metric, also known as imputed information score (INFO), is a measure of the relative statistical information about the SNP allele frequency from the imputed data. We defined good and bad quality imputed SNPs as in the original MACH paper, i.e. those with $R^2 > 0.5$ and with $R^2 < 0.2$, respectively, and stratified imputed SNPs based on their Rsq and INFO scores. This analysis was restricted to chromosome 20, and performed using as baseline genotypes the OmExp for the SardiNIA study and the Illumina 660W-quad for the MCTFR cohort.

RESULTS

Effect of baseline genotyping array

This subsection is restricted to the SardiNIA study and the simulated European haplotypes, because the MCTFR study used only one array. We found clear differences in imputation performance depending on the baseline genotyping set. Comparable differences were seen when assessments were done with either the Rsq metric - the imputation quality metric from MACH[4] - or the R^2 metric, the squared Pearson correlation, between dosages and real genotypes[4] (**Table 1**).

When using the 1000G reference panels for Sardinians, the two custom arrays (Cardio-MetaboChip and ImmunoChip) provided very limited information for imputation and far less accuracy than the genome-wide arrays, reflecting their low marker density. However, the Cardio-MetaboChip array performed very well when imputing with the SardSeq panel, allowing accurate inference of the rest of the genome (mean Rsq = 0.62, and mean R^2 = 0.70 at HumanExome SNPs). The relative efficiency was similar when considering all autosomes (**Table 1** and **Table S5**) or focusing only on chromosome 20 **(Figure2** and **Table S6**). The extended LD in the population and the increased genetic similarity of the reference panel aid in haplotype reconstruction when using a relatively small set of markers.

The addition of the two custom arrays to the OmExp genome-wide array (OmExp+Metab+Imm, called OMI here) did not improve quality for common or low frequency variants compared to that reached using OmExp alone. Thus, such arrays provide direct genotyping of low frequency and rare variants in genes of interest but do not contribute to an overall improvement in imputation accuracy. We also observed negligible differences in imputation accuracy between the two tested Illumina genome-wide arrays, OmExp and pHumCore (**Table1, Table S5** and **Table S6**), when imputing the SardSeq panel. In particular, we noticed that the low density genome-wide array pHumCore provided only slightly less accuracy than the denser OmExp array when the SardSeq sequencing panel was used for imputation (mean R^2 = 0.85 and 0.87, for pHumCore and OmExp, respectively, at HumanExome SNPs, **Table S5**) and a very similar genomic coverage (92.6% and 91.8% of markers imputed with Rsq > 0.3, **Table 1**). Of note, performance was patently lower for both arrays and more significantly for pHumCore when imputation was performed with the 1000G panels (mean R^2 = 0.54 and 0.64, for pHumCore and OmExp, respectively, imputing with the 1000G-ALL) (**Table 1**, **Table S5**, **Table S6** and **Figure 2**). By contrast, in the simulated European data, the Cardio-MetaboChip performed poorly, with insufficient genomic coverage. Contrarily to previous observations[22], the pHumCore was fairly comparable in efficiency to the OmExp array (**Figure 3, Table S7**), but we expect performance to be overestimated (because the genotypes were simulated based on 1000Genomes). In fact, when we extracted subset of SNPs that are present in HumanOmniExpress and HumanCore from the MCTFR genotypes, the difference between the two arrays was clearly evident (**Table S8**). This difference has also been observed for another European population[23].

Thus, in founder populations it appears that highly accurate imputation can be achieved with cost-effective sparse genotyping arrays when a population-specific reference panel is available.

Effect of study-specific reference panels

Study-specific reference panels increased the accuracy and completeness of coverage in both Sardinian and Minnesota samples, but the gain in accuracy was greater for the Sardinia founder population.

In Sardinians, the 1000G-ALL reference panel provided the highest number of imputed variants - ~37 million including both indels and SNPs *vs* ~15 million SNPs for the SardSeq panel - but the majority were of poor quality and were subsequently discarded. For example, for the Metab/SardSeq combination 11.5 million imputed SNPs passed the standard Rsq > 0.3 filter, but only 2.7 million and 3.0 million reached that threshold for Metab/1000G-ALL and Metab/1000G-EUR, respectively. The gap was less striking but still marked when denser genotype datasets were considered, and was still noticeable even considering only SNPs present in all reference panels [which are enriched for high frequency variants (**Table 1**)]. Consistent results were seen for the OmExp, OMI, pHumCore and Affy6.0 datasets, with accuracy consistently better when using SardSeq (**Figure 2**).

The benefit in overall accuracy was clear at all frequency ranges and even greater for low frequency and rare variants. For example, using the OMI dataset, the average R^2 for SNPs with MAF ranging from 0.5% to 1% is 0.91, 0.57 and 0.52 when using SardSeq, 1000G-ALL and 1000G-EUR reference panels, respectively (**Table S5**). This reinforces the finding that on average, low frequency variants are hard to impute in founder populations when using external reference panels, because those variants appear in fewer haplotypes[2]. Of note, the results remained the same after removing 646 Sardinian samples that appear in both the genotyping set and the SardSeq reference panel (**Table S2**).

To assess whether the advantage with the SardSeq panel was attributable to the lower number of European haplotypes present in the 1000Genomes reference, we performed imputation using the MinnSeq panel. There was no appreciable gain in accuracy within Sardinians compared to 1000G-based imputations (**Figure 4A**, **Table S9A** and **Table S10**).

Similar to results with Sardinians, the MinnSeq panel outperformed the 1000G panels in the MCTFR study, at all frequency ranges (**Figure 4B** and **Table S9B**). However, the gain in accuracy was far less than that observed in Sardinians with the SardSeq panels. For example, for variants with MAF ranging from 1% to 5%, we observed 11% and 42% additional gain in mean R^2 for Minnesota and Sardinians, respectively. Of note, in both cohorts the study-specific panel also yielded a higher number of SNPs useful for analyses (considering an Rsq > 0.3) even when the other reference sets contain more SNPs (**Table S10**).

Effect of combined reference panels

We also evaluated the impact on imputation accuracy of extended panels created by combining the two study-specific panels and 1000G haplotypes.

The combined SardSeq + 1000G panels provided only marginally higher accuracy at rarer shared SNPs in Sardinians (**Figure 2, Table S5** and **Table S6**). Slight increase in accuracy was also observable for more frequent variants [except for the two custom arrays (Metab and Imm), for which the improvement was substantial across all frequency ranges (**Figure 2**, **Table S5** and **Table S6)**]**.** Thus for Sardinians, the inclusion of 1000G haplotypes would only be beneficial for very rare variants if a genome-wide array were used for baseline imputation.

In the simulated European set, the addition of SardSeq haplotypes to the 1000G panels remarkably increased imputation accuracy for custom genotyping arrays (Metab and Imm) for both common and rare variants (**Figure 3** and **Table S7**). For example, for variants with MAF > 40% and MAF ≤ 50 % the mean R^2 is 0.57 and 0.98, when imputing with 1000G-ALL and SardSeq + 1000G-ALL and using the Metab dataset (**Figure 3** and **Table S7**). The impact of a combined panel was instead negligible for the more comprehensive genotype data (OmExp, OMI, pHumCore and Affy6.0). However, imputation on simulated data could give slight overestimations, and this could mask the advantage of adding SardSeq to 1000G panels. Indeed, when considering the MCTFR study, the combined SardSeq + 1000G-ALL panel provided benefit at all frequency ranges compared with 1000G-ALL imputation, and for MAF ≤ 0.5% variants accuracy becomes fairly similar to that observed when using the MCTFR specific panel (**Figure 4** and **Table S9**). Thus the Sardinian panel could be generally useful to increase the overall accuracy in population cohorts other than Sardinians, especially where only custom array genotyping is available or when a study-specific reference is not available.

Compared to imputation with MinnSeq alone, the addition of the 1000G haplotypes to the MinnSeq reference panel was useful only for rare variants in Minnesotans. The difference in accuracy was >4 fold higher that what seen in Sardinians comparing imputations with SardSeq and SardSeq + 1000G panels. Thus for Europeans, the inclusion of 1000G haplotypes in a study-specific panel is sensitively beneficial for very rare variants. Of note, for the Minnesotans, genotype imputation at the full spectrum of frequency ranges never reaches the same accuracy as in SardiNIA with the SardSeq panel, even when using the combined MinnSeq + 1000G with almost twice as many individuals as there are in the SardSeq panel.

Given the great utility of the Sardinian haplotypes, we further examined whether the advantage achieved by imputing with the SardSeq panel could have been reached sequencing a smaller number of samples and merging their haplotypes with the 1000G panels. For simplicity, we again focused on chromosome 20 and the OmExp array. Only for variants with MAF > 5% does adding 500 Sardinian samples to the 1000G panels provide the same accuracy as the SardSeq panel alone. Instead, adding 1000 Sardinians to the 1000G panels provides the same accuracy given by the SardSeq panel for all frequency bins, with only a modest difference in accuracy for the very rare variants (MAF < 0.5%) (**Figure S1** and **Table S11**).

Thus, sequencing a smaller number of individuals and combining their haplotypes with the 1000G panels could give imputation accuracy that is highly comparable to a panel comprising a large number of samples. However, the caveat remains that the genotype accuracy and variant discovery in low-pass sequencing is highly dependent on the number of sequenced samples. Consequently, sequencing only 500 samples would not provide genotypes as precise as those obtained by randomly

selecting 500 samples from a set of 2,000 sequenced genomes. For example, when we performed variant calling on a subset of 508 samples, the heterozygous error rate increased from 2.6% to 11.3% at rare sites (**Table S12**).

Performance of imputation quality metrics

To determine whether the commonly used MACH-Rsq threshold > 0.3 and IMPUTE-INFO > 0.4 can be applied to all frequency ranges (and if not, to infer appropriate cutoffs), we investigated how well imputation quality metrics can predict true imputation accuracy, especially for rare and less common variants. We found that for MAF ≥ 1% imputation accuracy and therefore concordance between real genotypes and dosages using study-specific panels was almost perfect in both Sardinians and Minnesotans (**Table 2**, **Table 3** and **Figure S2**). At these frequency ranges, high but clearly less concordance was also seen when imputing with the 1000G panels. Whatever the reference panel used and the population under study, the standard Rsq cutoff of > 0.3 efficiently discarded most badly imputed markers while keeping most of those imputed well (see **Materials and Methods**). In particular, imputation was so accurate overall that even an Rsq cutoff of > 0 would leave no badly imputed markers on chromosome 20 (**Table 2A, Table 3A**) (and only 8 over the entire genome in Sardinians, **Table S13**). Similarly for the INFO metrics, the standard > 0.4 threshold was efficient to discriminate between well and poorly inferred genotypes at this range of frequency (**Table 2B** and **Table 3B**).

By contrast, for MAF < 1%, we noticed that both metrics were slightly overestimated when using the study-specific panels, possibly because of the inclusion of relatives with similar haplotypes in the target dataset; but overall concordance was better than 1000G imputation for this range of frequency as well. Specifically, in this range and when imputation was performed with the 1000G panels, the threshold of Rsq > 0.3 was less efficient, aggressively discarding some well imputed variants (eliminating 7-18% and 7-25% of the well imputed markers for ALL and EUR panels) and retaining an excess of the badly imputed ones (**Table 2A, Table 3A, Table S14** and **Table S15**). The INFO > 0.4 threshold instead worked efficiently on selecting well imputed variants, but was too lenient on discarding those of poor quality (**Table 2B** and **Table 3B, Table S14** and **Table S15**). Nevertheless, Rsq > 0.3 and INFO > 0.4 still remain the optimal thresholds.

When imputation was performed with the study-specific panels, both the Rsq and INFO thresholds were more efficient in capturing all well imputed markers, but less efficient in discarding the poorly imputed.

In such cases, e.g., for MAF < 1% and when imputation is performed with a reference panel that is genetically close to the study population, an Rsq threshold of > 0.6 and INFO > 0.7 should be preferred in lieu of the standard thresholds of 0.3 and 0.4, respectively.

DISCUSSION

We used different reference panels and genotype input sets to investigate effects on imputation, in founder and non-founder populations of European ancestry. We found that a study-specific reference panel considerably improved imputation accuracy and genomic coverage compared to external equally large reference panels, regardless of the genotype array, especially for rare variants. However, the benefit was strikingly higher in the founder population of Sardinians, with a precision that was not obtainable in Europeans even with a reference panel twice the size. In fact, in such homogenous populations each sequenced genome provides information that can be extended to distant relatives as well, whereas in continental Europeans, haplotypes carrying rare variants can only inform closely related samples.

We also observed that in Sardinians a study-specific panel boosts imputation even for low coverage genotyping array(s), like the Cardio-MetaboChip, which are barely informative when imputing with the 1000G panels alone, or for the HumanCore, which becomes highly comparable for all frequency ranges to the wider HumanOmniExpress. Given the low cost of the sparser arrays, accurate population-scale imputation is more feasible in the Sardinian founder population than in non-founder populations when combined with large-scale sequencing. For example, at current cost schedules, with an investment of 500,000 dollars one could genotype ~8,300 Sardinian samples with the HumanCore array instead of ~4,500 with the HumanOmniExpress. The power to detect association for variants accounting for 0.5% of the trait variance thereby rises from 24% to 84%.

Finally, we observed that standard thresholds on metrics for evaluating accuracy, estimated by two commonly used imputation software, are somewhat imprecise for rare variants. We propose that all cohorts using study-specific reference panels for imputation consider adopting different thresholds for common and rare variants to filter inaccurate genotypes.

Taken together, these imputation-based analyses can guide genetic studies, and complement recent reports [22,24] with several novel aspects that can improve performance:

- They exploit imputation accuracy with the two larger study-specific reference panels so far publi shed, including one that is population-specific.
- They also provide the first evaluation of imputation performance of the 1000 Genomes Project h aplotypes in an isolated population.
- They include analyses of large cohorts coupled with the use of HumanExome array, allowing app ropriate assessment of results for less frequent and rare variants.
- Using real data sets, they based analyses on a subset of quality controlled SNPs instead of the ful l list of markers present on an array (excluding many that are likely to be imperfectly genotyped i n a case study).
- They evaluate two widely used custom genotyping arrays, Cardio-MetaboChip and ImmunoChip, providing information for cohorts that are limited to that source of genotypes.
- They also evaluate for rare variants the efficiency of accuracy metric thresholds that were previo usly suggested for common variants.

Ultimately, full genome sequencing could make imputation methods superfluous, but the time scale remains indeterminate. It should be considered that increasing sample size can augment genome-wide power to assess rare variants more than increasing array density -- even up to full genotyping

of the complete 1000 Genomes Project variant set[22,24]. Thus aids to imputation are increasingly valuable, because most studies are likely to be collecting increasing numbers of samples and using this inferential process rather than sequencing full genomes.

Overall, population specific panels might have been thought to be "private", with potential discoveries limited to that population. Instead, the effectiveness of population-specific reference panels can be appreciable for other populations, but will vary depending on the size of the panels and the demographic history of the isolate. Intuitively in Europe, their value may be greater for populations like Basques and Greeks, who are relatively genetically distant from the European samples selected for the 1000 Genomes Project. Here, we show that sequencing efforts from the Sardinian founder population can, when coupled with available panels, improve rare variants imputation accuracy in other population backgrounds as well. This reinforces the value of isolated populations for discovery of variants that are locally enriched but rarer and thus harder to detect in international surveys[25].

REFERENCES

1. 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.

2. Porcu E, Sanna S, Fuchsberger C, Fritsche LG: Genotype imputation in genome-wide association studies. *Curr Protoc Hum Genet* 2013; **Chapter 1:** Unit1 25.

3. Marchini J, Howie B: Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11:** 499-511.

4. 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.

5. Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39:** 906-913.

6. Pilia G, Chen WM, Scuteri A *et al*: Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* 2006; **2:** e132.

7. Iacono WG, McGue M, Krueger RF: Minnesota Center for Twin and Family Research. *Twin Res Hum Genet* 2006; **9:** 978-984.

8. Miller MB, Basu S, Cunningham J *et al*: The Minnesota Center for Twin and Family Research genome-wide association study. *Twin Research and Human Genetics* 2012; **15:** 767-774.

9. Voight BF, Kang HM, Ding J *et al*: The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012; **8:** e1002793.

10. Cortes A, Brown MA: Promise and pitfalls of the Immunochip. *Arthritis Res Ther* 2011; **13:** 101.

11. Goldstein JI, Crenshaw A, Carey J *et al*: zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* 2012; **28:** 2543-2545.

12. Scuteri A, Sanna S, Chen WM *et al*: Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 2007; **3:** e115.

13. Naitza S, Porcu E, Steri M *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.

14. Vrieze SI, Feng S, Miller MB *et al*: Rare nonsynonymous exonic variants in addiction and behavioral disinhibition. *Biol Psychiatry* 2013; **75:** 783-789.

15. Sanna S, Pitzalis M, Zoledziewska M *et al*: Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nat Genet* 2010; **42:** 495-497.

16. Pitzalis M, Zavattari P, Murru R *et al*: Genetic loci linked to type 1 diabetes and multiple sclerosis families in Sardinia. *BMC Med Genet* 2008; **9:** 3.

17. Orru V, Steri M, Sole G *et al*: Genetic variants regulating immune cell levels in health and disease. *Cell* 2013; **155:** 242-256.

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. *Manuscript in Preparation* 2014.

19. Delaneau O, Zagury JF, Marchini J: Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10:** 5-6.

20. Su Z, Marchini J, Donnelly P: HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 2011; **27:** 2304-2305.

21. Danecek P, Auton A, Abecasis G *et al*: The variant call format and VCFtools. *Bioinformatics* 2011; **27:** 2156-2158.

22. Nelson SC, Doheny KF, Pugh EW *et al*: Imputation-based genomic coverage assessments of current human genotyping arrays. *G3 (Bethesda)* 2013; **3:** 1795-1807.

23. Francioli LC, Menelaou A, Pulit SL *et al*: Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; **46:** 818-825.

24. Lindquist KJ, Jorgenson E, Hoffmann TJ, Witte JS: The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. *Genet Epidemiol* 2013; **37:** 383-392.

25. Holm H, Gudbjartsson DF, Sulem P *et al*: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 2011; **43:** 316-320.

Table 1. Basic imputation statistics on the SardiNIA samples, for different panels/genotyping arrays.

The table shows, for each genotyping array/reference panel combination, the number of imputed SNPs and the corresponding mean Rsq and standard deviation, the percentage of SNPs with Rsq > 0.3, with the corresponding mean Rsq and standard deviation, mean Rsq and standard deviation evaluated for 8,842,944 SNPs that were imputed in all genotyping array/reference panel combination (called "Shared imputed SNPs").

		Whole imputed SNPs set		Rsq > 0.3		Shared imputed SNPs
Array	**Reference Panel**	**N SNPs**	**Mean(SD) Rsq**	**% SNPs**	**Mean (SD) Rsq**	**Mean (SD) Rsq**
Imm	SardSeq	15,071,719	0.258 (0.312)	33.33	0.652 (0.213)	0.299 (0.321)
	1000G-ALL	37,798,002	0.037 (0.134)	3.90	0.638 (0.232)	0.099 (0.213)
	1000G-EUR	16,873,087	0.085 (0.203)	9.68	0.647 (0.231)	0.115 (0.232)
Metab	SardSeq	15,069,660	0.617 (0.335)	76.91	0.777 (0.181)	0.685 (0.301)
	1000G-ALL	37,782,741	0.064 (0.170)	7.20	0.614 (0.217)	0.175 (0.260)
	1000G-EUR	16,878,099	0.149 (0.253)	18.05	0.634 (0.219)	0.201 (0.282)
MetabImm	SardSeq	14,977,409	0.734 (0.300)	86.51	0.835 (0.163)	0.808 (0.239)
	1000G-ALL	37,721,853	0.100 (0.218)	11.71	0.644 (0.221)	0.272 (0.311)
	1000G-EUR	16,781,983	0.219 (0.303)	27.12	0.667 (0.222)	0.297 (0.328)
OmExp	SardSeq	14,580,754	0.861 (0.256)	92.61	0.924 (0.131)	0.935 (0.161)
	1000G-ALL	37,424,729	0.297 (0.382)	33.61	0.796 (0.224)	0.742 (0.322)
	1000G-EUR	16,453,325	0.543 (0.406)	60.89	0.84 (0.206)	0.729 (0.341)
OMI	SardSeq	14,319,695	0.862 (0.256)	92.57	0.925 (0.131)	0.937 (0.159)
	1000G-ALL	37,211,511	0.300 (0.385)	34.00	0.799 (0.131)	0.753 (0.318)
	1000G-EUR	16,255,689	0.549 (0.406)	61.50	0.842 (0.206)	0.739 (0.337)
pHumCore	SardSeq	15,020,615	0.840 (0.264)	91.81	0.908 (0.139)	0.913 (0.179)
	1000G-ALL	37,793,052	0.234 (0.341)	26.66	0.759 (0.221)	0.614 (0.354)
	1000G-EUR	16,825,817	0.455 (0.398)	52.64	0.802 (0.207)	0.615 (0.367)

	SardSeq	14,550,658	0.798 (0.342)	84.51	0.937 (0.116)	0.905 (0.232)
Affy6.0	1000G-ALL	37,328,716	0.263 (0.379)	29.55	0.814 (0.217)	0.721 (0.341)
	1000G-EUR	16,350,040	0.515 (0.416)	57.63	0.843 (0.205)	0.708 (0.357)

Table 2. Efficiency of imputation quality metrics in the SardiNIA cohort

The table shows the number and the percentage of poorly and well imputed SNPs (see **Materials and Methods**) that are captured for each Rsq (Panel A) and INFO (Panel B) threshold. Imputation was performed on chromosome 20 HuamnOmniExpress SNPs, using the SardSeq and 1000G-ALL panels. Statistics are reported separately for common and rare variants.

A	MAF < 1%				MAF ≥ 1%			
	SardSeq		1000G-ALL		SardSeq		1000G-ALL	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (14)	100 (222)	100 (98)	100 (124)	0 (0)	100 (301)	100 (20)	100 (255)
> 0.1	92.86 (13)	100 (222)	44.9 (44)	92.74 (115)	0 (0)	100 (301)	90 (18)	99.61 (254)
> 0.2	85.71 (12)	100 (222)	19.39 (19)	86.29 (107)	0 (0)	100 (301)	75 (15)	99.61 (254)
> 0.3	78.57 (11)	100 (222)	11.22 (11)	81.45 (101)	0 (0)	100 (301)	65 (13)	98.43 (251)
> 0.4	71.43 (10)	100 (222)	5.1 (5)	70.16 (87)	0 (0)	100 (301)	45 (9)	97.25 (248)
> 0.5	64.29 (9)	99.55 (221)	3.06 (3)	62.9 (78)	0 (0)	100 (301)	30 (6)	94.9 (242)
> 0.6	42.86 (6)	95.95 (213)	2.04 (2)	50.81 (63)	0 (0)	100 (301)	20 (4)	89.41 (228)
> 0.7	28.57 (4)	91.89 (204)	0 (0)	43.55 (54)	0 (0)	100 (301)	15 (3)	82.35 (210)
> 0.8	14.29 (2)	83.78 (186)	0 (0)	33.87 (42)	0 (0)	100 (301)	0 (0)	72.16 (184)
> 0.9	7.14 (1)	58.11 (129)	0 (0)	23.39 (29)	0 (0)	98.01 (295)	0 (0)	59.22 (151)
> 1	0 (0)	0.9 (2)	0 (0)	0 (0)	0 (0)	3.65 (11)	0 (0)	2.75 (7)

B	MAF < 1%				MAF ≥ 1%			
	SardSeq		1000G-ALL		SardSeq		1000G-ALL	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (7)	100 (189)	100 (81)	100 (83)	0 (0)	100 (307)	100 (32)	100 (251)
> 0.1	100 (7)	100 (189)	100 (81)	98.8 (82)	0 (0)	100 (307)	100 (32)	100 (251)
> 0.2	100 (7)	99.47 (188)	90.12 (73)	97.59 (81)	0 (0)	100 (307)	100 (32)	100 (251)

> 0.3	100 (7)	99.47 (188)	62.96 (51)	96.39 (80)	0 (0)	100 (307)	96.88 (31)	100 (251)
> 0.4	100 (7)	99.47 (188)	48.15 (39)	93.98 (78)	0 (0)	100 (307)	96.88 (31)	100 (251)
> 0.5	100 (7)	99.47 (188)	27.16 (22)	89.16 (74)	0 (0)	100 (307)	84.38 (27)	99.2 (249)
> 0.6	100 (7)	98.94 (187)	17.28 (14)	85.54 (71)	0 (0)	100 (307)	59.38 (19)	98.01 (246)
> 0.7	71.43 (5)	97.35 (184)	11.11 (9)	73.49 (61)	0 (0)	100 (307)	37.5 (12)	95.62 (240)
> 0.8	42.86 (3)	92.59 (175)	3.7 (3)	60.24 (50)	0 (0)	100 (307)	15.62 (5)	88.84 (223)
> 0.9	14.29 (1)	76.19 (144)	0 (0)	44.58 (37)	0 (0)	99.35 (305)	6.25 (2)	72.91 (183)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Table 3. Efficiency of imputation quality metrics in the MCTFR cohort.

The table shows the number and the percentage of poorly and well imputed SNPs (see **Materials and Methods**) that are captured for each Rsq (Panel A) and INFO (Panel B) threshold. Imputation was performed on chromosome 20 Illumina 660W-quad array SNPs, using MinnSeq and 1000G-ALL as reference panels. Statistics are reported separately for common and rare variants.

A

	MAF < 1%				MAF ≥ 1%			
	MinnSeq		1000G-ALL		MinnSeq		1000G-ALL	
Rsq	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (38)	100 (129)	100 (80)	100 (92)	0 (0)	100 (284)	100 (4)	100 (258)
> 0.1	81.58 (31)	100 (129)	72.5 (58)	96.74 (89)	0 (0)	100 (284)	100 (4)	100 (258)
> 0.2	73.68 (28)	100 (129)	41.25 (33)	95.65 (88)	0 (0)	100 (284)	25 (1)	100 (258)
> 0.3	57.89 (22)	100 (129)	26.25 (21)	92.39 (85)	0 (0)	100 (284)	25 (1)	99.61 (257)
> 0.4	47.37 (18)	100 (129)	17.5 (14)	83.7 (77)	0 (0)	100 (284)	25 (1)	98.84 (255)
> 0.5	28.95 (11)	96.9 (125)	10 (8)	72.83 (67)	0 (0)	100 (284)	0 (0)	96.12 (248)
> 0.6	21.05 (8)	92.25 (119)	3.75 (3)	59.78 (55)	0 (0)	95.07 (270)	0 (0)	87.21 (225)
> 0.7	2.63 (1)	72.09 (93)	1.25 (1)	48.91 (45)	0 (0)	89.79 (255)	0 (0)	77.13 (199)
> 0.8	0 (0)	51.94 (67)	0 (0)	31.52 (29)	0 (0)	79.58 (226)	0 (0)	62.02 (160)
> 0.9	0 (0)	28.68 (37)	0 (0)	17.39 (16)	0 (0)	59.51 (169)	0 (0)	48.45 (125)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

B

	MAF < 1%				MAF ≥ 1%			
	MinnSeq		1000G-ALL		MinnSeq		1000G-ALL	
INFO	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)	% bad (n)	% good (n)
> 0	100 (38)	100 (96)	100 (82)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.1	100 (38)	100 (96)	100 (82)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.2	100 (38)	100 (96)	97.56 (80)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)

> 0.3	97.37 (37)	100 (96)	95.12 (78)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.4	94.74 (36)	100 (96)	69.51 (57)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.5	73.68 (28)	100 (96)	41.46 (34)	100 (67)	100 (1)	100 (277)	100 (9)	100 (241)
> 0.6	50 (19)	98.96 (95)	14.63 (12)	100 (67)	0 (0)	100 (277)	44.44 (4)	100 (241)
> 0.7	23.68 (9)	95.83 (92)	7.32 (6)	92.54 (62)	0 (0)	99.28 (275)	0 (0)	99.59 (240)
> 0.8	5.26 (2)	77.08 (74)	0 (0)	71.64 (48)	0 (0)	91.7 (254)	0 (0)	87.97 (212)
> 0.9	2.63 (1)	40.62 (39)	0 (0)	46.27 (31)	0 (0)	72.2 (200)	0 (0)	63.49 (153)
> 1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Figure 1. Graphical representation of analyses and study aims

The figure shows a scheme of analyses carried out. For each genotype input set we carried out several imputation runs (genome-wide for SardiNIA, and on chromosome 20 for other European populations) with different reference panels. We assessed imputation quality of each genotype array/reference panel combination by looking at the mean imputation quality (MACH-Rsq) and by comparing imputed markers with those directly typed with the HumanExome array (R^2). Finally, we assessed the efficiency of standard thresholds at the commonly used accuracy metrics (MACH-Rsq/IMPUTE-INFO) in filtering bad imputed markers.

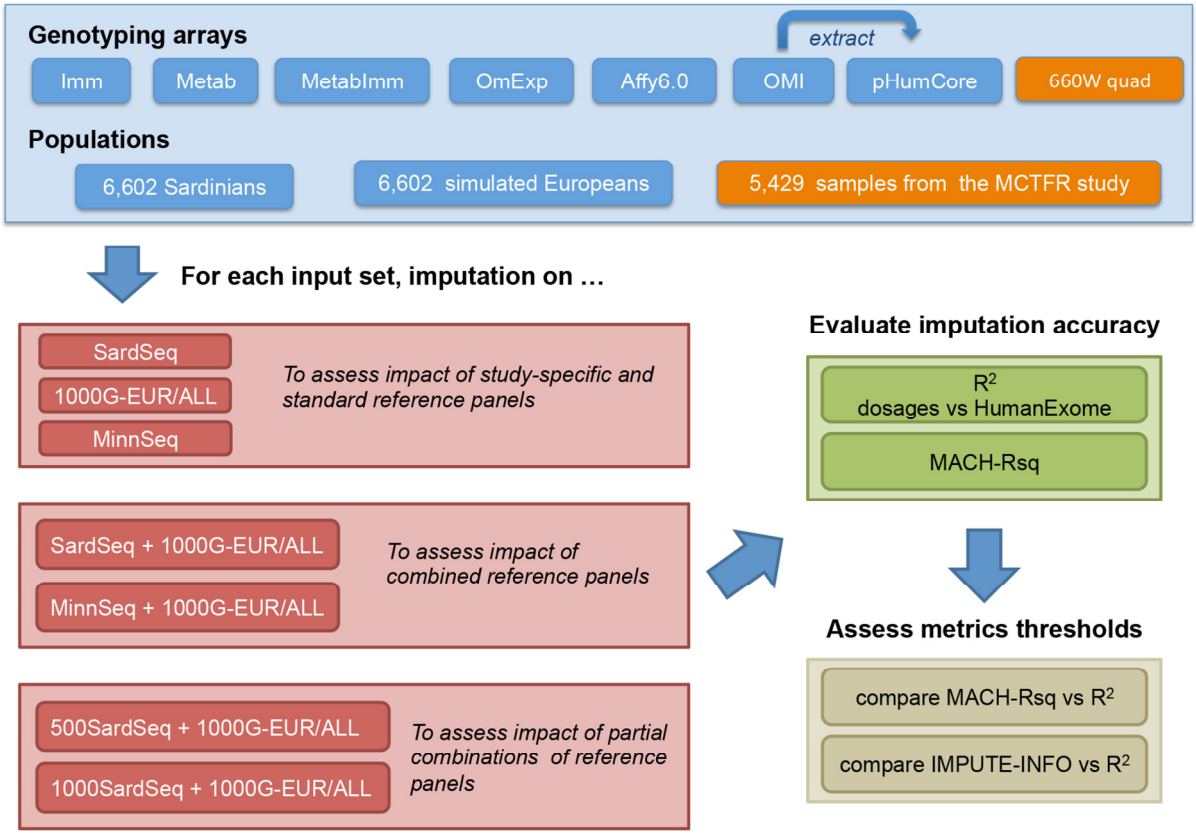


Figure 2. Mean R^2 for each particular genotype array/reference panel in the SardiNIA cohort

The figure shows the mean R^2 at different allele frequencies ranges, for each particular genotyping array/reference panel combination including the combination of SardSeq and 1000G panels. Results are restricted to chromosome 20.

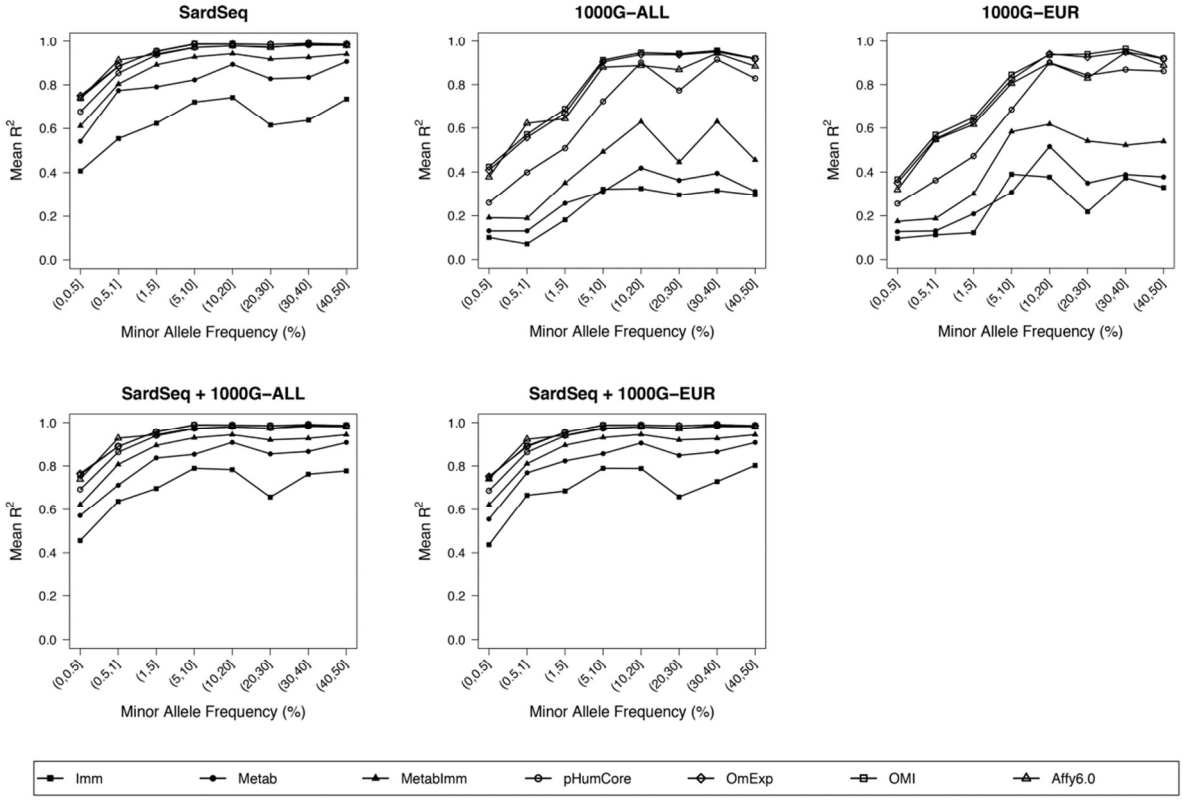


Figure 3. Mean R^2 for each combination of genotype array/reference panel in the European simulated data set

The figure shows the mean R^2 at different allele frequencies ranges, for each particular genotyping array/reference panel, including the combination of SardSeq and 1000G panels. Results are restricted to chromosome 20.

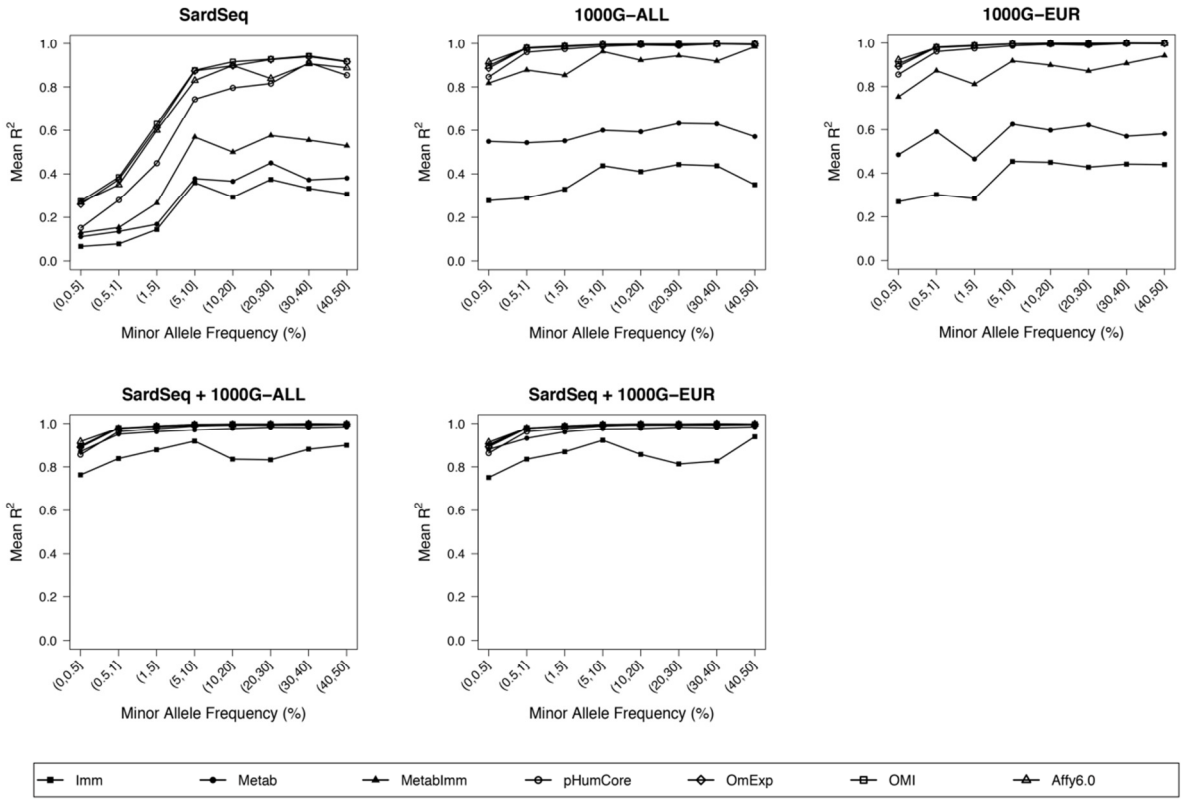
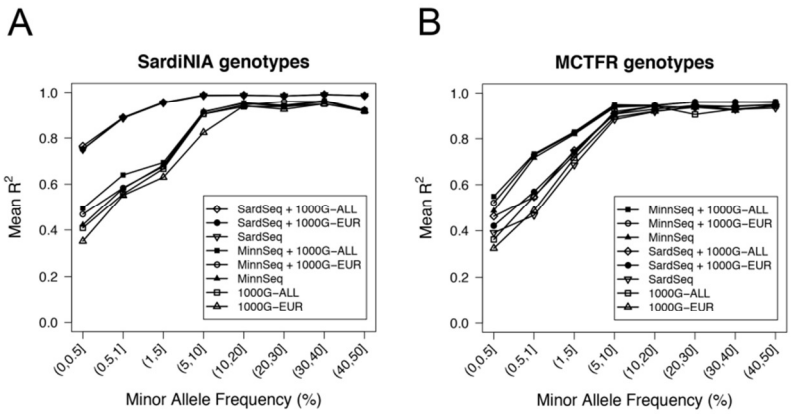


Figure 4. Impact of cross-studies reference panels

The figure shows the mean R^2 at different allele frequencies ranges for the chromosome 20 of OmExp genotyping array for SardiNIA (panel A) and the Illumina 660W-quad array for the MCTFR (panel B) study, when using different reference panels, including combination of SardSeq/MinnSeq and 1000G panels and cross-studies references.



Part II: Sequencing-based GWAS in the isolated Sardinian population

Chapter 4: Genetic variants regulating immune cell levels in health and disease.

Based on:

*Orru V[*], Steri M[*], Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A, Floris M, Mentzen WI, Urru SA, Olla S, Marongiu M, Piras MG, Lobina M, Maschio A, Pitzalis M, Urru MF, Marcelli M, Cusano R, Deidda F, Serra V, Oppo M, Pilu R, Reinier F, Berutti R, Pireddu L, Zara I, Porcu E, Kwong A, Brennan C, Tarrier B, Lyons R, Kang HM, Uzzau S, Atzeni R, Valentini M, Firinu D, Leoni L, Rotta G, Naitza S, Angius A, Congia M, Whalen MB, Jones CM, Schlessinger D, Abecasis GR, Fiorillo E[#], **Sanna S**[#], Cucca F[#]*

Cell. 2013 Sep 26;155(1):242-56. doi: 10.1016/j.cell.2013.08.041

[,#] indicate equal contributions*

ABSTRACT

The complex network of specialized cells and molecules in the immune system has evolved to defend against pathogens, but inadvertent immune system attacks on "self" result in autoimmune disease. Both genetic regulation of immune cell levels and their relationships with autoimmunity are largely undetermined. Here we report genetic contributions to quantitative levels of 95 cell types encompassing 272 immune traits, in a cohort of 1,629 individuals from four clustered Sardinian villages. We first estimated trait heritability, showing that it can be substantial, accounting for up to 87% of the variance (mean 41%). Next, by assessing ~8.2 million variants we identified, and confirmed in an extended set of 2,870 individuals, 23 independent variants at 13 loci associated with at least one trait. Notably, variants at 3 loci (HLA, IL2RA, SH2B3/ATXN2) overlap with known autoimmune disease associations. These results connect specific cellular phenotypes to specific genetic variants, helping to explicate their involvement in disease.

INTRODUCTION

The immune system must defend against a huge variety of microbes and remember them. To accomplish this and kill cancer transformed and virus infected cells, while recognizing and tolerating our own untransformed components, requires the formation and regulation of a wide range of both generalist and specialist white cell (leukocyte) types. Fluorescence activated cell sorting (FACS) approach has facilitated highly sensitive, simultaneous analysis of levels of these leukocyte subpopulations and is being used by the Human Immunology Project to characterize the immunological profile of healthy and sick individuals (Davis, 2008; Maecker et al., 2012).

Despite methodological advances, searches for connections between genetic variants and cellular immune phenotypes have typically proceeded by examining broad classes of immune cells (Ferreira et al., 2010; Nalls et al., 2011; Okada et al., 2011), and even the extent to which variation in immune cell subtypes is heritable is still unknown.

Here we use FACS to profile extensively the human immune cell repertoire for a large population sample. Applying state-of-the-art genotyping and sequencing technologies to the same individuals, we proceed to dissect the inherited phenotypic structure of the human immune cell repertoire. Importantly, our results demonstrate connections between known immune-related disease risk alleles and levels of particular immune cell types, thus representing an important extension of previous autoimmune disease GWAS. Our hypothesis-generating approach, using individuals from a general population, is also distinct from hypothesis-driven comparisons of immune cell types between cases and controls, which can be hampered by limited *a priori* knowledge and affected by second order effects due to the disease process or its therapy.

RESULTS AND DISCUSSION

Profiling the human immune cell repertoire

By FACS analyses, we characterized a wide range of circulating cell subtypes in an initial sample of 1,629 individuals enrolled in the SardiNIA study population cohort. The cells comprise the major leukocyte populations in peripheral blood (**Figure 1**), including monocytes, granulocytes, circulating dendritic cells (cDCs), natural killer (NK), B and T cells with a more detailed characterization of T cell subsets. More specifically, because of their functional relevance and potential involvement in many autoimmune and inflammatory diseases, we focused on T cells subdivided according to their maturation and activation status, including subsets of regulatory T cells (Tregs) (Shevach, 2000; Wing and Sakaguchi, 2010). Overall, we defined a total of 95 cell types that were further assessed with respect to their parental and grandparental cell lineages, resulting in 272 evaluated immunophenotypic traits (**Experimental Procedures**; **Figures 1** and **S1**; **Table S1A,B,C**).

Heritability and correlations between traits

We estimated heritability of circulating immune cell counts in the first 1,629 phenotyped individuals (**Experimental Procedures**), observing values from 3% to 87% (mean 41%). The most heritable traits corresponded to Tregs and their subsets (mean 55%) (**Figure S2**; with **Table S2A** giving descriptive statistics and impact of age and gender covariates). Remarkably, most cell populations with very high heritability (>60%) were positive for the CD39 marker (see below). Gender typically had negligible effects on phenotypic variation; age was important for a subset of cellular phenotypes, especially the previously characterized reduction in naive CD8 T cells that might explain reduced vaccination success in the elderly (Buchholz et al., 2011; Sansoni et al., 2008).

By their nature, many traits are hierarchically and functionally correlated, as the different immune cell types originate from a limited number of common progenitors and interact continuously. To examine these relationships, we performed a bivariate analysis to estimate phenotypic and genetic correlation coefficients, i.e. the proportion of variance between each pair of traits due to the combined contribution of genetic and environmental factors and the variance attributable to genetic causes only, respectively (**Extended Experimental Procedures**). A depiction of the genetic and phenotypic correlations between cell counts and CD4:CD8 and T:B cell ratios is presented in the heat map (**Figure 2)**. Similarities but also important differences in the patterns of genetic and phenotype correlation coefficients --reported in the upper and lower triangles of the figure delimited by the central diagonal-- are immediately apparent. On the one hand, two large squares in the upper part of the diagonal are indicative of conjoint genetic and phenotypic correlations, and tend to involve cells with markers, such as CD39 and CD45RA, whose expression is under strong genetic control (see next section) suggesting that the extent of similarity between traits reflects intrinsic relations dictated mainly by ontogenesis and coordinated evolution of traits --and hence shared antigen expression. On the other hand, the overall phenotypic correlations tend to be stronger than genetic correlations, consistent with additional effects of non-genetic factors on cell levels. An example of strong positive phenotypic but not corresponding strong genetic correlation is observed between some DC and Treg subsets (corresponding to the lower-right red cluster), in line with a mechanism by which an increase of DCs is controlled by an increase of Tregs (Wing and Sakaguchi, 2010).

Genetic changes affecting immune cell traits

To identify the genetic variation accounting for the inherited component of the 272 immunophenotypic traits, we next performed a sequencing-based GWAS, assessing ~8.2 million variants in the 1,629 phenotyped individuals (**Experimental Procedures**). At the significance threshold of $p<5.26x10^{-10}$, we identified 21 signals at 11 loci linking genetic variation to multiple cellular immunophenotypes and resulting in a total of 180 SNP-trait associations. We also replicated ($p<5x10^{-8}$) two previously suggested associations (Ferreira et al., 2010) resulting in a total of 23 association signals at 13 loci which were then assessed and unequivocally confirmed in the extended sample set of 2,870 individuals including 1,241 additional volunteers (**Figures 3**, **4** and **S3**; **Tables 1**, **S4A,B** and **S5A**).

The amount of phenotypic variation explained was always >2%, consistent with the expected statistical power of our sample, with nine variants explaining >5% and three variants >15% (**Table 1**). Considering the 132 traits for which we observed at least one genome-wide significant signal, the heritability explained ranged from 3.7 to 90.3%, and the proportion of explained heritability was >50% for 35 traits and >80% for four traits (**Figure 5**; **Table S2A**), showing relatively large effects for human quantitative traits (Teslovich et al., 2010; Lango Allen et al., 2010).

Among the largest genetic effects detected, a single intronic variant of *ENTPD1*, coding for CD39, accounted for 60.8% of phenotypic variation (and 72% of the heritability) of the levels of CD39+ activated CD4+ Tregs (**Table 1; Figure 4**). Thus, this association has an obvious candidate mechanism in which *cis*-acting variation regulates the expression of a key marker in individual cells and therefore determines the number of cells expressing this molecule. CD39 is an ectoenzyme, expressed on monocytes, neutrophils, B, T and NK cells (Pulte et al., 2007), which hydrolyzes extracellular ATP and ADP to AMP. Notably, among T cells, CD39 is mainly expressed by activated CD4+ Tregs, where it has an anti-inflammatory function by reducing extracellular pro-inflammatory ATP (Borsellino et al., 2007).

Other clear biological candidates among our lead associations included a variant near *IL2RA*, a gene encoding the transmembrane protein CD25, associated with variation of T cells expressing high CD25 levels (CD45RA- CD25hi CD4+ not Treg cells); a variant near the *CD8A* and *CD8B* genes, encoding the cell surface glycoprotein CD8, associated with variation in the level of T cells expressing CD8 (CD4+ CD8dim); a variant near the HLA class II transactivator (*CIITA*) gene, associated with the levels of activated T cells (i.e., HLA DR+ T lymphocytes); and a variant in the *TNFSF13B* gene, associated with the levels of B cells. Notably, *CIITA* encodes a transcription factor influencing HLA class II expression, whereas *TNFSF13B* encodes the B cell Activating Factor of the TNF Family (BAFF), inactivation of which is specifically associated with loss of mature circulating B cells (**Table 1)** (Mackay et al., 2009).

Overall, 19 of the 23 variants reported here were associated with multiple traits often with divergent effects on different traits (**Tables S4A** and **S5A**). A further layer of complexity was added by instances of multiple independent associations with the same traits within a single associated region. For example, independent variants within a region encompassing the *GALM* and *HNRPLL* genes (**Table 1**) increased the percentage of naive and terminally differentiated T cell subpopulations (those that are CD45RA positive), with corresponding decreases in the percentage of the memory T cell

subsets (which are CD45RA negative). Association with *HNRPLL* is fully concordant with its role as the master regulator of *CD45* splicing, a hallmark of T cell maturation (Wu et al., 2010). By contrast, the biology underlying the associations with *GALM* is less clear, though variants in *GALM* may act in long distance regulation of *HNRPLL*, because they fall in DNA regions known to interact with its promoter (**Table S6A)** (Li et al., 2012).

Other examples of multiple independent signals clustered in the same gene regions and associated with several traits were found near *ENTPD1* and in the *HLA* region (**Table 1)**, where *multilocus* and *multiallelic* associations with complex diseases have been extensively documented (Marrosu et al., 2001). These results illustrate a new role for *HLA* variants, modulating immune system function by affecting the level of specific immune cell types. Of note, several variants in *HLA* class I alleles were associated with variation in the levels of numerous distinct CD8+ T cell subtypes, consistent with the notion that self-class I MHC molecules support CD8+ T cell survival (Takada and Jameson, 2009).

In general, most of the associations reported in this work are new, though some are consistent with previously detected signals. Specifically, we confirmed the putative associations between NK cell levels and variants near the *Schlafen* gene cluster, and the association of CD4+ T cells with variation in the *SH2B3/ATXN2* gene region (Ferreira et al., 2010).

In addition to associations with $p<5.26x10^{-10}$, we observed several additional signals at $p<5x10^{-8}$ (**Table S5B**) that require confirmation by further analyses. Most of them are likely to be genuine --for example, the association of a common non-synonymous variant (N1639S) in the lactase gene (*LCT*) with pDCs. It is striking that the association of two independent missense variants at this locus with leukocyte count in African Americans was recently reported (Auer et al., 2012), further supporting an unanticipated role of coding variation within this locus in the regulation of immune cell levels.

Our results also highlight the benefit of imputation and sequencing-based GWAS, both in detection of association signals and in the identification of the causal genes and variants (so-called 'fine mapping'), which is relevant for downstream functional studies. In fact, three of the 13 detected loci (*NCAM, CD4* and *HLA-E*) reached significance only after imputation. Across all loci, 20 lead variants were imputed, and two of them were not present even in the HapMap data set or in the most recent 1000 Genomes release, and thus were not directly accessible by imputation from external resources. One, rs58055840, has proxies in the 1000 Genomes panel, but the other, chr10:98088623, is not strongly correlated with other known markers. Further investigation is required to determine whether these variants are specific to Sardinians.

Functional clues from the associated variants

The 23 lead variants are located in non-coding regions, although two of them are in strong linkage disequilibrium (LD) ($r^2>0.8$) with non-synonymous coding variants (with features of variants detailed in **Table S4C**). Furthermore, seven variants fall within known elements with regulatory capacity, including repressors, enhancers and promoter elements or transcription factor binding sites (**Table S6A**). To assess functional processes and pathways through which the variants exert their effects, we selected a set of candidate genes based on physical position and biological features, and surveyed Gene Ontology (GO) terms and pathway enrichment (**Experimental Procedures**; **Table 1**). As expected, even when genes located in the *HLA* region were excluded, the overrepresented pathways

and GO categories were predominantly related to immune function (e.g., immune response, immune system process, primary immunodeficiency, hematopoietic cell lineage, antigen processing and presentation, T cell receptor complex, IgG binding, MHC protein binding and IL12-mediated signaling events) (**Table S6B**).

Overlapping associations between immune traits and diseases

After identifying immune cell associated variants we checked whether any of them correlated with known disease associations. After identifying immune cell associated variants, we systematically checked in public databases whether any of them was, or was highly correlated (r^2> 0.8), with a known disease associated variant previously reported at $p<5x10^{-8}$. We identified overlaps at 3 genetic loci: *HLA*, *IL2RA*, and *SH2B3/ATXN2* (**Table 2**; **Figure S4; Extended Experimental Procedures**). Such overlapping associations identify specific immune cell types that are unbalanced in disease status and also suggest mechanisms by which specific risk alleles might lead to disease susceptibility, as follows.

Variation downstream of the *HLA-DRA* gene decreased the levels of memory CD8+ cells not expressing the co-stimulatory molecule CD28 (CD45RA- CD28- CD8+ cells) and correlated with published associated risk alleles for ulcerative colitis, systemic sclerosis, Parkinson's disease and Hodgkin's lymphoma (Barrett et al., 2009; Enciso-Mora et al., 2010; Gorlova et al., 2011; Hamza et al., 2011).

A variant in the *IL2RA* gene region, rs61839660, was associated with a memory T cell subset expressing high CD25 levels (CD45RA- CD25hi CD4+ not Treg cells) and is also the strongest type 1 diabetes (T1D) associated variant in the region (Huang et al., 2012; Lowe et al., 2007). Moreover, association with the same immune cells was previously observed at a variant in moderate LD (r^2=0.77), which was at the time the strongest T1D-associated variant (Dendrou et al., 2009). The allele responsible for an increase in the CD45RA- CD25hi CD4+ not Treg cells reduces the risk for T1D, thus linking this specific cell type to protection against T1D. The results also suggest that anti-CD25 therapies might increase risk for T1D by reducing the number of this protective cell type. Consistent with this, clinical trials have suggested an increased risk of T1D in transplant patients treated with anti-CD25 antibody (Bayes et al., 2007; Vendrame et al., 2010).

Another overlap was seen for a variant in *ATXN2* that is highly correlated with a missense variant within the *SH2B3* gene (R262W). The W262 non-ancestral allele increases the levels of T lymphocytes and the helper CD4+ T cell subset with similar effect sizes, and it is positively associated with many autoimmune diseases (such as type 1 diabetes and celiac disease), as well as with hypertension and related pathologies (i.e., coronary heart disease and chronic kidney disease). Additionally, this variant has been associated with several endophenotypes in the general population, including platelet and eosinophil levels as well as systolic and diastolic blood pressure (Hindorff et al., 2013). *SH2B3* encodes the adaptor protein LNK, whose mouse orthologue was earlier shown to be a negative regulator of haematopoiesis, cytokine signaling and inflammation (Devalliere and Charreau, 2011). An increase of total T cells (CD3+ lymphocytes) and particularly of CD4+ T cells resulting from the W262 allele may thereby result in loss of function. Furthermore, this observation

is consistent with findings in mouse models and humans suggesting the potential efficacy of monoclonal antibodies against CD3 in T1D and other autoimmune diseases (Chatenoud, 2010).

Relevant to its function, the *SH2B3* associated variant marks an extended haplotype spanning ~200 kb (**Figure S3** panel **22A**), indicative of strong positive selection (Barreiro and Quintana-Murci, 2010). During human evolution, a lymphocytosis-associated variant may have been useful for thousands of years in resistance to pathogens, but in recent less septic environments becomes a risk factor for autoimmunity.

Other genetic variants might also be enriched by balancing selection to maintain a high degree of variation in immune cell levels in a given population, increasing the chances for survival of groups of individuals under different and often opposite environmental pressures. Among our associated variants, clear evidence of balancing selection was found in the *HLA* region (**Extended Experimental Procedures**), consistent with its key role in host defense and disease susceptibility.

In addition to coincident associations clearly satisfying stringent criteria, other overlapping signals for variants affecting both levels of specific cells and disease risk are likely genuine. For example, the allele associated with a higher level of HLA-DR+ (activated) T lymphocytes at *CIITA* is in moderate LD (r^2=0.44) with the risk allele for Celiac Disease (CD) (Trynka et al., 2011). In this case the lack of full coincidence at the same SNP or a suitable proxy may be attributable to differences in map resolution in different studies (the coverage at this locus was low in the CD study). Furthermore, our top variant is in strong LD (r^2=0.99) with a variant showing suggestive association with ulcerative colitis (McGovern et al., 2010).

Overall, the coincident associations between diseases and immune traits have special potential to reveal sites for therapeutic intervention, and indeed some of those detected had already been selected as targets for pharmaceutical therapy (**Table S6C**). It is also noteworthy that our work does not support some previous claims, largely based on functional evidence, about the involvement of specific cell type levels in specific diseases. For instance, a protective role of CD39+ activated CD4+ Tregs in various autoimmune diseases has been suggested (Chalmin et al., 2012; Fletcher et al., 2009), but no overlapping association was observed between disease and the major genetic variants affecting the quantitative regulation of this cell type.

Conclusions and prospects

As part of the dynamic mounting and control of immune reactions, our results reveal that DNA variation superimposes powerful programmed regulation on various subtypes of leukocytes. Interestingly, those showing the greatest estimated inherited control are implicated in the more sophisticated cellular functions, such as regulatory T cells, which were phylogenetically the last to evolve and are also the last to appear in ontogenesis.

A number of the genetic associations identified here explain an appreciable fraction of trait heritability and demonstrate the feasibility of genetic dissection of quantitative variation of specific immune cell types. At least three factors likely contribute to the unusually high degree of explained heritability, which contrasts sharply with typical observations in GWAS for quantitative traits, for which “missing heritability” is the norm. First, examining more restricted cell types avoids dilution and possible opposing effects in mixtures of leukocytes; this notion is consistent, for example, with findings of large effect size variants associated with fetal hemoglobin that have no detectable effects on total hemoglobin (Uda et al., 2008). Second, the large genetic effect sizes could be related to

intrinsic properties of the immune response which, confronted at the population level with an unpredictable and changing environment, must ensure optimal primed variability in the quantitative levels of immune cell types. Finally, the sequencing based approach employed provides assessment of genomic variation at an unprecedented level of resolution, except for very rare SNPs and indels.

The 13 reported loci point to specific DNA polymorphisms and putative proteins and mechanisms involved in regulation of cellular immunity. They also identify specific molecules and cell subtypes involved in a range of diseases, particularly autoimmune diseases, reflecting the dramatically shifting evolutionary balance between the optimization of effective response to pathogens and the risk of autoimmunity. Given that several association signals may have not been captured in this study due to sample size restrictions --and in most previous disease GWAS due to their restriction to common ubiquitous variants-- many more overlapping associations are likely to be forthcoming when the approach described here is extended to larger samples. These overlaps should include multiple associations for the same trait and disease, reinforcing evidence of causal relationships between them. Our survey also reveals primary candidate genes to be re-sequenced in searches for both germ line mutations in patients with selective and combined immunodeficiencies and driver somatic mutations in patients with circulating haematopoietic malignancies. Some of the observations presented here also hint at previously undocumented involvement of the immune system in maladies such as Parkinson, though rigorous testing in appropriate cohorts is required to assess these possibilities further.

For some autoimmune pathologies, the mechanistic clues involving specific cell types suggest targets but also concomitant risks for therapeutic interventions, with some drugs already in use or under clinical experimentation targeting the associated protein products for a number of loci. Further functional studies to explicate the effects of the variants on identified cell types could foster therapies aimed at controlling the numbers of those cell types to help regulate the immune system safely, preventing occurrence or lessening severity of autoimmune diseases.

EXPERIMENTAL PROCEDURES

Study population

The SardiNIA project is a longitudinal study that recruited and phenotyped 6,148 individuals, males and females, aged 14–102 y, from a cluster of four towns in the Lanusei Valley (Pilia et al., 2006), located on the central east coast of Sardinia, Italy. During clinic visits, fresh blood samples were collected and used for both DNA extraction and flow cytometric measurements. Initially, 1,629 individuals were characterized for the immune-related phenotypes described below, followed by an additional 1,241 individuals from the same cohort, to extend the sample size and validate the identified association results. Ethical permission for this study was granted by the Regional Ethics Committee (No 2009/0016600).

Flow cytometric measurements

Immunophenotyping was carried out by flow cytometry on fresh blood samples and cell phenotyping was performed within two hours after collection, to avoid any time dependent artifacts. We selected and tested a set of multiplexed fluorescent antibodies to characterize the

major leukocyte cell populations in peripheral blood, including monocytes, granulocytes, circulating dendritic cells and lymphocytes subdivided into NK, B and T cells and their subsets (**Extended Experimental Procedures**; **Figure S1**; **Table S1A**). In particular, we assessed regulatory T cells (CD25hi, CD127-), subdivided into resting, activated and cytokine-secreting non-suppressive cells (Miyara et al., 2009; Shevach, 2000). We also used the HLA-DR marker to assess the activation status of T and NK cells, and both the chemokine receptor CCR7 and the phosphatase CD45RA antigens to distinguish between naïve, central memory (CM), effector memory (EM) and terminally differentiated (TD) T cell subsets (Sallusto et al., 1999). Moreover, in selected T cell subpopulations we assessed the positivity for the ectoenzyme CD39 and the CD28 co-stimulatory antigen (Keir and Sharpe, 2005). Finally, cDCs were separated into myeloid (mDCs) and plasmacytoid (pDCs) cells and further subdivided by the expression of the adhesion molecule CD62L and the co-stimulatory ligand CD86 (Steinman and Banchereau, 2007; Ohnmacht et al., 2009).

For all cell populations, we measured both absolute counts (AC) and the proportion of each type with respect to their progenitor cell lineages, expressed as percentages of the levels of parent (%P) and grandparent (%GP) cell lineages (**Figure 1**; **Table S1B**). For example, helper CD4+ T cells were evaluated relative to CD3+ cells (parent cell population representing all T cells) and to total lymphocytes (grandparent cell population). Percentages with respect to parental and grandparental cell populations lead to more robust measures of cell levels by reducing variability in measurements resulting from sample handling or fluctuations by transient environmental factors that affect the total leukocyte counts. These percentages may also reveal association with molecular changes that alter factors involved in feedback mechanisms responsible for maintaining a balance between cells. Finally, we assessed the specific ratios of cell types that are widely clinically used and that examine the balance between T and B cells and between helper (CD4) and cytotoxic (CD8) T cells.

Overall, we examined 95 absolute counts, 94 percentages with respect to parent cells, 80 percentages with respect to grandparent cells, and 3 ratios between cell subsets (**Table S1B**).

To ensure reproducible measures over time we followed a rigorous standardization protocol (**Extended Experimental Procedures,** "Flow cytometry instrument setting and reproducibility of measurements" paragraph). Briefly: i) we daily adjusted internal parameters of FACS using standardized fluorescent beads to check and correct for laser wear and fluidic instability; and ii) we weekly validated cell counts through suitable quality control of stabilized blood samples. To directly assess reproducibility we repeated the FACS measurements in 35 participants sampled at least three months after their initial enrollment, finding overall high reproducibility (median value for all traits 0.90, mean 0.85, standard deviation 0.13) (**Table S2B**).

Heritability estimation and bivariate analysis

We estimated heritability for all inverse-normalized traits in the first 1,629 immunophenotyped individuals (comprising 211 unrelated individuals, and 1,418 subjects grouped in 249 families, leading to 567 sib-pairs, 30 half-sib pairs, 248 cousins-pairs, 609 parent-child pairs, 32 grandparent-grandchild pairs and 561 avuncular pairs for analysis), including age and gender as covariates. Furthermore, familial clustering of blood sampling (i.e., same day sampling of closely related

individuals), which could bias heritability estimates, was checked for and excluded (**Extended Experimental Procedures, "**Heritability and bivariate analysis**"** paragraph).

We also performed a bivariate analysis to estimate the phenotypic and genetic correlations between traits. In particular, for each trait pair the phenotypic correlation was computed as the Spearman coefficient, whereas the genetic correlation was estimated as the cross trait-cross individual additive genetic covariance between traits normalized by the geometric mean of the individual trait genetic variances and by the kinship coefficient of pairs of individuals. We then used a hierarchical clustering analysis that successively connected the most similar traits, based on the estimated phenotypic and genetic correlation coefficients (**Extended Experimental Procedures, "**Heritability and bivariate analysis**"** paragraph).

Genotyping and whole genome sequencing

The entire SardiNIA cohort was characterized using two Illumina custom arrays: the Cardio-MetaboChip and the ImmunoChip. These arrays were designed by international consortia to genotype regions of prior interest in metabolic and immune related traits and diseases, respectively (Cortes and Brown, 2011; Voight et al., 2012) and resulted in quality controlled 284,722 SNPs derived from both arrays. We also whole-genome sequenced 1,146 Sardinians at low pass (average 4 fold coverage) (**Extended Experimental Procedures**, "Genotyping arrays" and "Sample sequencing and variant calling" paragraphs).

Statistical and bioinformatical analyses

We performed a GWAS for each trait analyzing ~8.2 million variants assembled from the integration of the two assessed arrays, and markers imputed with the Sardinian sequencing reference panel (**Table S3; Extended Experimental Procedures,** "Genotype imputation" and "Association analyses" paragraphs) (Li et al., 2009). Association was evaluated by a variance component-based regression analysis, to account for family structure, using the same covariates as in heritability estimation (Chen and Abecasis, 2007). Traits were normalized using inverse normal transformation.

We selected all independently associated variants for each trait ($r^2<0.1$ or those remaining significant in a stepwise conditional analysis), using a significance threshold of $p<5.26\times10^{-10}$. This threshold corresponds to the standard genome-wide threshold of 5×10^{-8} after further adjustment for 95 independent tests (the number of absolute cell count measurements). While this approach is conservative given the high interdependency of cell lineages, it ensures the robustness of our findings. We successively removed poorly imputed variants, and then eliminated redundant trait-variant associations by prioritizing the most strongly associated variants at each locus and removing those in LD. We also included two suggestive associations ($5.26\times10^{-10}<p<5\times10^{-8}$) at the previously described *SH2B3/ATXN2* and *SLFN13* gene regions (Ferreira et al., 2010).

To validate findings, we measured the corresponding associated immunophenotypes in an additional 1,241 individuals from the same SardiNIA cohort and genotyped variants representing novel signals that were not supported by a directly genotyped variant ($r^2>0.85$). Variants showing an excess of discordant genotypes or less significant p-values after addition of the extended sample

were excluded from further analyses (**Table S4A,B; Extended Experimental Procedures**, "Validation of findings" paragraph).

To calculate the amount of phenotypic variance explained by genetic factors, for each trait we fitted a linear model containing age, gender and all the independent SNPs associated with that specific trait (full model), and a linear model containing only age and gender (basic model). The variance explained was calculated as the difference of the r^2-adjusted quantity observed in the full and basic models (**Table S2A**).

To prioritize candidate gene(s) at each locus, we searched for correlated expression quantitative trait loci (eQTLs), coding variants and nearby genes involved in immune-related disorders, as reported in OMIM (Online Mendelian Inheritance in Man), or implicated in immunity in previous studies (**Table S4C,D**). Bioinformatic analyses were carried out to characterize variants and genes, including co-localization with regulatory features, and their potential for pharmaceutical interest (**Table S6A,B,C**). Lastly, to assess possible impact of the detected variants on disease susceptibility, we searched for coincident associations in public repositories (**Table 2**), such as the GWAS catalog (Hindorff et al., 2013) and ImmunoBase (http://www.immunobase.org/) (**Extended Experimental Procedures)**.

A schematic overview of the overall study design is depicted in **Figure S5**.
The **Extended Experimental Procedures,** included in the Supplemental Information file, provide details about the study design, genetic and immunophenotypic data collection, and statistical and bioinformatic analyses. They are all available online on Cell Journal website.

Table 1. Twenty-three variants at the 13 associated loci

The independently associated variants for each locus are tabulated, along with the association parameters. Indicated are, from left to right, the locus number; the candidate genes potentially regulated by the variant (for each candidate gene a letter indicates the reason for inclusion: p=position; e=eQTL; c=coding; o=OMIM; b=biological candidate); the chromosomal position on hg19/GRCh37 genomic build of the lead variant and the corresponding SNP identification number (rs ID), when available; the major and minor alleles (A1 and A2) and the frequency of the major allele; the corresponding associated trait (CD8+ corresponds to the summation of CD8bright and CD8dim cells); the effect size in standard deviation units per each copy of allele A1; the standard error; the variance explained; and the p-value. The last three columns report parameters for the SNP used in the validation step: the chromosome position with the corresponding identification number; the correlation with the lead SNP and the p-value of the validation data set are listed, respectively. See also **Table S4A**.

Locus	*Candidate genes*	*topSNP (chr:position)/rsID*	*A1/A2*	*Freq A1*	*Trait*	*Effect (SE)*	*Var. Expl.*	*p-value (N=1,629)*	*SNP for validation (chr:position/ rsID)*	*r^2 with topSNP*	*Validation p-value (N=2,870)*
1	*FCGR3A(p,c,o), FCGR2C(p,o), FCGR2A(e,c,o), FCGR2B(e,o), HSPA6(e), HSPA7(e)*	chr1:161536758/ rs58055840	T/C	0.742	CD62L- myeloidcDC AC	-0.895 (0.044)	30.26	$3.73x10^{-91}$	chr1:161515326/ rs55971447	0.937	$6.83x10^{-129}$
2	*HNRPLL(p)*	chr2:38792045/ rs183949931	T/C	0.967	CD45RA- CD28- CD8br %P	0.778 (0.105)	4.05	$1.05x10^{-13}$	chr2:38792045/rs183949931	Same SNP	$1.046x10^{-20}$
2	*GALM(p,c,e), HNRPLL(b)*	chr2:38897074/ rs13011383	G/A	0.730	TD CD4+ %GP	-0.371 (0.042)	5.52	$6.05x10^{-19}$	chr2:38886041/ rs4670262	0.87	$1.26x10^{-27}$
2	*GALM(p), DHX57(e), HNRPLL(b)*	chr2:38921934/ rs7583259	G/C	0.508	CD45RA- CD28- CD8br %P	-0.548 (0.039)	15.09	$9.40x10^{-46}$	chr2:38932777/ rs4670265	0.9	$2.82x10^{-62}$
3	*CD8A(p,c,o), RMND5A(p), CD8B(b), VPS24(e)*	chr2:87014377/ rs2944254	C/T	0.810	CD4+ CD8dim AC	0.383 (0.05)	4.55	$2.52x10^{-14}$	chr2:87018547/ rs3810831	0.943	$1.3x10^{-22}$
4	*COQ2(e), PLAC8(e), HPSE(e)*	chr4:84150313/ rs4431216	T/C	0.633	CD62L- plasmacytoidcDC %P	0.337 (0.04)	5.19	$4.96x10^{-17}$	chr4:84179071/ rs7667017	0.84	$3.37x10^{-23}$
5	*HLA-E(p,c,e), HCG27(e), GNL1(c), ABCF1(e), C2(e), PSORS1C3(e), RPP21(e), TRIM39(e), ZKSCAN2(e)*	chr6:30466505/ rs117765619	G/T	0.516	CD45RA- CD8+ AC	-0.228 (0.037)	2.62	$5.24x10^{-10}$	chr6:30482993/ rs2534812	0.974	$1.34x10^{-11}$
5	*HLA-B(p,c), VARS2(e), IER3(e), ZFP57(e)*	chr6:31327382/ rs2395476	T/G	0.858	CD45RA- CD28+ CD8+ %P	0.352 (0.051)	3.21	$3.69x10^{-12}$	chr6:31327382/ rs2395476	Same SNP	$1.827x10^{-19}$
5	*HLA-DRA(p,e), BTNL2(p,c), HLA-DRB1(c,e), HLA-DQA1(e), HLA-DQB1(e), HLA-DRB5(e), HLA-DOB(e), LOC642073(e), VARS2(e), LST1(e), IER3(e), GTF2H4(e), HMGA1(e), RPL34(e)*, AOAH(e)**	chr6:32386433/ rs113534101	G/A	0.776	CD4+ CD8dim %P	-0.299(0.043)	3.07	$5.68x10^{-12}$	chr6:32383138/ rs115615758	0.97	$2.78x10^{-16}$
5	*HLA-DRA(p), LOC642073(e), HLA-DOB(e), RPL34(e)*,ARHGAP24(e)*, AOAH(e)**	chr6:32428186/ rs6923504	G/C	0.618	CD45RA- CD28- CD8+ AC	-0.249 (0.037)	3.01	$2.81x10^{-11}$	chr6:32428285/ rs6903608	0.99	$4.3x10^{-13}$

6	*IL2RA(p,o)*	chr10:6094697/ rs61839660	C/T	0.934	CD45RA- CD25hi CD4+ not Treg %P	-0.49 (0.073)	2.82	$1.85x10^{-11}$	chr10:6094697/ rs61839660	Same SNP	$5.65x10^{-23}$
6	*RBM17(p), IL2RA(p,o)*	chr10:6158412/ rs8463	A/G	0.802	CD25hi CD4+ %P	-0.294 (0.046)	2.85	$1.21x10^{-10}$	chr10:6158412/ rs8463	Same SNP	$2.02x10^{-15}$
7	*SORBS1(p), C10orf61(e), ALDH18A1(c), ENTPD1(e)*	chr10:97331924/ rs117568941	T/C	0.955	CD39+ CD8+ %GP	-0.650 (0.062)	6.68	$1.45x10^{-25}$	chr10:97331958/ rs7099430	0.969	$1.32x10^{-35}$
7	*ALDH18A1(p), ENTPD1(b)*	chr10:97393678/ rs1890187	A/G	0.975	CD39+ activated CD4+ Treg %P	-0.671 (0.073)	5.97	$5.72x10^{-20}$	chr10:97550405/ rs11188485	0.97	$2.97x10^{-32}$
7	*ENTPD1(p,e)*	chr10:97564532/ rs11517041	T/C	0.578	CD39+ activated CD4+ Treg %P	-1.113 (0.037)	60.81	$1.12x10^{-202}$	chr10:97515137/ rs3814159	0.993	$7.05x10^{-327}$
7	*ZNF518A(p), BLNK(p,o), ENTPD1(b)*	chr10:97932006/ rs117592294	C/T	0.955	CD39+ CD25hi CD4+ %P	0.497 (0.066)	4.33	$6.26x10^{-14}$	chr10:97932006/ rs117592294	Same SNP	$1.35x10^{-15}$
7	*DNTT(p), OPALIN(p), BLNK(o), ENTPD1(b)*	chr10:98088623	A/G	0.978	CD39+ CD4+ AC	-0.777 (0.094)	6.05	$1.87x10^{-16}$	chr10:98088623	Same SNP	$1.809x10^{-20}$
8	*NCAM1(b)*	chr11:112706386/ rs76771478	G/T	0.890	Lymphosum %P	0.455 (0.064)	4.51	$1.62x10^{-12}$	chr11:112707378/ rs1992842	0.96	$7.18x10^{-17}$
9	*CD4(p,e,o)*	chr12:6899181/ rs2855537	G/T	0.606	naive (CD4+ CD8+) AC	0.315 (0.048)	4.70	$5.94x10^{-11}$	chr12:6898460/ rs7956804	1	$4.77x10^{-13}$
10	*TNFSF13B(p), LIG4(o)*	chr13:108957063/ rs9520836	A/G	0.513	B cell %GP	-0.239 (0.035)	2.95	$1.45x10^{-11}$	chr13:108957063/ rs9520836	Same SNP	$1.39x10^{-14}$
11	*CIITA(p,o)*	chr16:10974355/ rs9924520	A/G	0.778	HLA DR+ T lymphocyte %P	-0.435 (0.039)	8.15	$2.20x10^{-28}$	chr16:10975311/ rs4781011	0.994	$9.29x10^{-50}$
12	*ATXN2(p), SH2B3(p,o)*	chr12:111973358/ rs597808	G/A	0.539	T lymphocyte AC	-0.195 (0.035)	2.01	$3.84x10^{-08}$	chr12:111973358/ rs597808	Same SNP	$1.87x10^{-09}$
13	*SLFN13(p), SLFN12L(p,c), CCL1(e)*	chr17:33797371/ rs9916257	T/G	0.568	NK %GP	-0.212 (0.035)	2.54	$9.78x10^{-10}$	chr17:33797371/ rs9916257	Same SNP	$4.72x10^{-20}$

**Trans eQTLs*

Table 2. Overlapping associations with complex diseases

Association statistics from the immune trait analyses are reported in the first six columns. The pathology, the disease associated variant, its best-reported p-value in public repositories and the risk allele are indicated in the 7th, 8th, 9th and 10th columns, respectively. The LD (r^2) between immune trait variant and the disease associated variant is shown in column 11th, whereas column 12th lists the risk allele coupled with the corresponding immune trait allele (and its effect). The last column indicates whether the disease was reported in GWAS Catalog (1), or ImmunoBase (2), or in PMID:23603763 (3). The disease-associated variants highlighted in boldface reach the standard genome-wide association threshold ($p<5x10^{-8}$) in public databases. See also **Figure S4**.

Gene (region)	*Immune trait*	*SNP*	*Effect allele/ Other*	*Effect (SE)*	*p-value*	*Disease*	*SNP disease*	*Best reported p-value*	*Risk allele/ Other*	*r2*	*Risk allele/ Corresponding trait allele (effect)*
***HLA* Class II (chr6p21.1)**	**CD45RA- CD28- CD8+ AC**	**rs6923504**	**G/C**	**-0.249 (0.037)**	**2.81E-11**	**Hodgkin's lymphoma**	**rs6903608**	**2.84E-50**	**G/A**	**0.99**	**G/G (decrease)**
						Systemic sclerosis	**rs3129882**	**1.89E-27**	**G/A**	**0.803**	**G/G (decrease)**
						Ulcerative colitis	**rs9268877**	**3.90E-23**	**T/C**	**0.83**	**G/G (decrease)**
						Parkinson's disease	**rs3129882**	**1.90E-10**	**G/A**	**0.803**	**G/G (decrease)**
***IL2RA* (chr10p15.1)**	**CD25hi CD4+ %P**	**rs61839660**	**C/T**	**-0.484 (0.072)**	**2.38E-11**	**Type 1 diabetes**	**rs61839660**	**5.10E-09**	**C/T**	**1**	**C/C (decrease)**
	CD45RA- CD25hi CD4+ not Treg AC	**rs61839660**	**C/T**	**-0.484 (0.072)**	**1.05E-10**	**Type 1 diabetes**	**rs61839660**	**5.10E-09**			**C/C (decrease)**
	CD45RA- CD25hi CD4+ not Treg %P	**rs61839660**	**C/T**	**-0.484 (0.072)**	**1.85E-11**	**Type 1 diabetes**	**rs61839660**	**5.10E-09**			**C/C (decrease)**
***SH2B3/ATXN2* (chr12q24.12)**	**T lymphocyte AC**	**rs597808**	**G/A**	**-0.195(0.035)**	**3.84E-08**	**Type 1 diabetes**	**rs3184504**	**2.80E-27**	**T/C**	**0.95**	**T/A (increase)**
						Celiac disease	**rs3184504**	**5.40E-21**	**T/C**		**T/A (increase)**
						Primary hypothyroidism	**rs3184504**	**2.60E-12**	**T/C**		**T/A (increase)**
						Primary sclerosing cholangitis	**rs3184504**	**5.91E-11**	**T/C**		**T/A (increase)**
						Juvenile rheumatoid arthritis	**rs3184504**	**2.60E-09**	**T/C**		**T/A (increase)**
						Rheumatoid arthritis	rs3184504	6.00E-06	T/C		T/A (increase)
						Coronary heart disease	rs3184504	6.35E-06	T/C		T/A (increase)
						Multiple sclerosis	rs3184504	6.70E-05	T/C		T/A (increase)
	CD4+ AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.66E-08**	**Type 1 diabetes**	**rs3184504**	**2.80E-27**			**T/A (increase)**
						Celiac disease	**rs3184504**	**5.40E-21**			**T/A (increase)**
						Primary hypothyroidism	**rs3184504**	**2.60E-12**			**T/A (increase)**
						Primary sclerosing cholangitis	**rs3184504**	**5.91E-11**			**T/A (increase)**
						Juvenile rheumatoid arthritis	**rs3184504**	**2.60E-09**			**T/A (increase)**
						Rheumatoid arthritis	rs3184504	6.00E-06			T/A (increase)
						Coronary heart disease	rs3184504	6.35E-06			T/A (increase)
						Multiple sclerosis	rs3184504	6.70E-05			T/A (increase)
	CD4+ not Treg AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.80E-08**	**Type 1 diabetes**	**rs3184504**	**2.80E-27**			**T/A (increase)**
						Celiac disease	**rs3184504**	**5.40E-21**			**T/A (increase)**
						Primary hypothyroidism	**rs3184504**	**2.60E-12**			**T/A (increase)**
						Primary sclerosing cholangitis	**rs3184504**	**5.91E-11**			**T/A (increase)**
						Juvenile rheumatoid arthritis	**rs3184504**	**2.60E-09**			**T/A (increase)**
						Rheumatoid arthritis	rs3184504	6.00E-06			T/A (increase)
						Coronary heart disease	rs3184504	6.35E-06			T/A (increase)
						Multiple sclerosis	rs3184504	6.70E-05			T/A (increase)
	T lymphocyte AC	**rs597808**	**G/A**	**-0.195(0.035)**	**3.84E-08**	**Celiac disease**	**rs653178**	**7.15E-21**	**C/T**	**0.96**	**C/A (increase)**
						Chronic kidney disease	**rs653178**	**3.50E-11**	**C/T**		**C/A (increase)**

						Rheumatoid arthritis	rs653178	1.50E-05	C/T		C/A (increase)
	CD4+ AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.66E-08**	**Celiac disease**	**rs653178**	**7.15E-21**			**C/A (increase)**
						Chronic kidney disease	**rs653178**	**3.50E-11**			**C/A (increase)**
						Rheumatoid arthritis	rs653178	1.50E-05			C/A (increase)
	CD4+ not Treg AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.80E-08**	**Celiac disease**	**rs653178**	**7.15E-21**			**C/A (increase)**
						Chronic kidney disease	**rs653178**	**3.50E-11**			**C/A (increase)**
						Rheumatoid arthritis	rs653178	1.50E-05			C/A (increase)
	T lymphocyte AC	**rs597808**	**G/A**	**-0.195(0.035)**	**3.84E-08**	**Primary biliary cirrhosis**	**rs11065979**	**2.87E-09**	**T/C**	**0.92**	**T/A (increase)**
	CD4+ AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.66E-08**	**Primary biliary cirrhosis**	**rs11065979**	**2.87E-09**			
	CD4+ not Treg AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.80E-08**	**Primary biliary cirrhosis**	**rs11065979**	**2.87E-09**			
	T lymphocyte AC	**rs597808**	**G/A**	**-0.195(0.035)**	**3.84E-08**	**Vitiligo**	**rs4766578**	**3.54E-18**	**T/A**	**0.96**	**T/A (increase)**
	CD4+ AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.66E-08**	**Vitiligo**	**rs4766578**	**3.54E-18**			
	CD4+ not Treg AC	**rs597808**	**G/A**	**-0.195(0.036)**	**4.80E-08**	**Vitiligo**	**rs4766578**	**3.54E-18**			
***CIITA* (chr16p13.13)**	HLA DR+ T lymphocyte AC	rs9924520	A/G	-0.425(0.042)	1.46E-23	Ulcerative colitis	rs4781011	3.23E-06	T/G	0.99	T/G (increase)
	HLA DR+ T lymphocyte %P	rs9924520	A/G	-0.435(0.039)	2.20E-28	Ulcerative colitis	rs4781011	3.23E-06	T/G		
	HLA DR+ T lymphocyte %GP	rs9924520	A/G	-0.449(0.041)	2.35E-28	Ulcerative colitis	rs4781011	3.23E-06	T/G		

Figure 1. Studied leukocyte subpopulations

Color-coded diagram of the cell types analyzed by flow cytometry with arrows depicting the hierarchical levels of separation of circulating cell populations (leukocytes) and constituent subsets of the two main arms, innate and adaptive, of the immune system. Innate cell types, which provide prompt but generic responses to aggressors, include granulocytes (yellow), monocytes (pale blue) and dendritic cells (red). Adaptive cell types, which provide highly specific responses to microbial targets and may maintain a "memory" that enables a faster and greater response to previously encountered pathogens, include B cells (magenta) and T cells (green). The natural killer cells (orange) share features of both arms of the immune system. The name and, when relevant, the identifying marker are indicated beside each population. Cells inside a light blue rectangle were phenotypically characterized with the antigen pointed to by the adjacent light blue arrow; for example, the six CD3+ subsets (CD4- CD8-, CD4+, CD4+ CD8dim, CD4+ CD8br, CD8br, and CD8dim) are shown within a blue rectangle, and were further subdivided into naïve, central memory, effector memory, and terminally differentiated cells. The red rectangle indicates that the included cell populations have been jointly analyzed for CD39, the marker indicated by the red arrow. For simplicity, 45 of the 95 analyzed cell type, described in the full text, are shown. See also **Figure S1** and **Table S1** for further details.

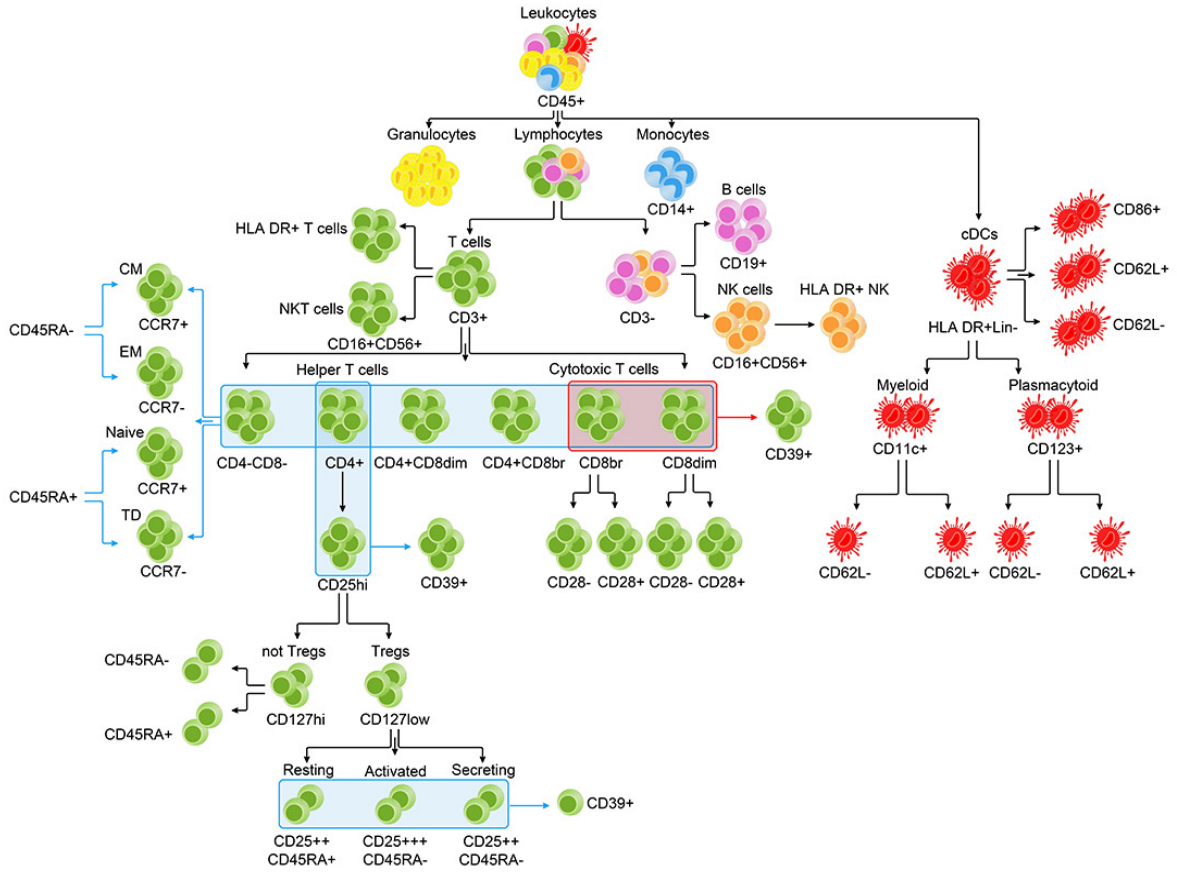


Figure 2. Phenotypic and genetic clustering

Heat map of phenotypic (lower-right triangle) and genetic (upper-left triangle) correlations for cell counts and CD4:CD8 and T:B cell ratios. Traits with a phenotypic correlation ≥0.99 were excluded (**Extended Experimental Procedures**). Genetic and phenotypic triangles follow the same trait order, dictated by the clustering of phenotypic correlations, and the dendrogram at the right reflects the clustering. Traits connected by short branches share stronger phenotypic correlation, whereas traits that join near the root of the tree are weakly correlated. Color gradations indicate correlation strength, with red indicating direct correlation (from 0 to +1) and blue inverse correlation (from 0 to -1).

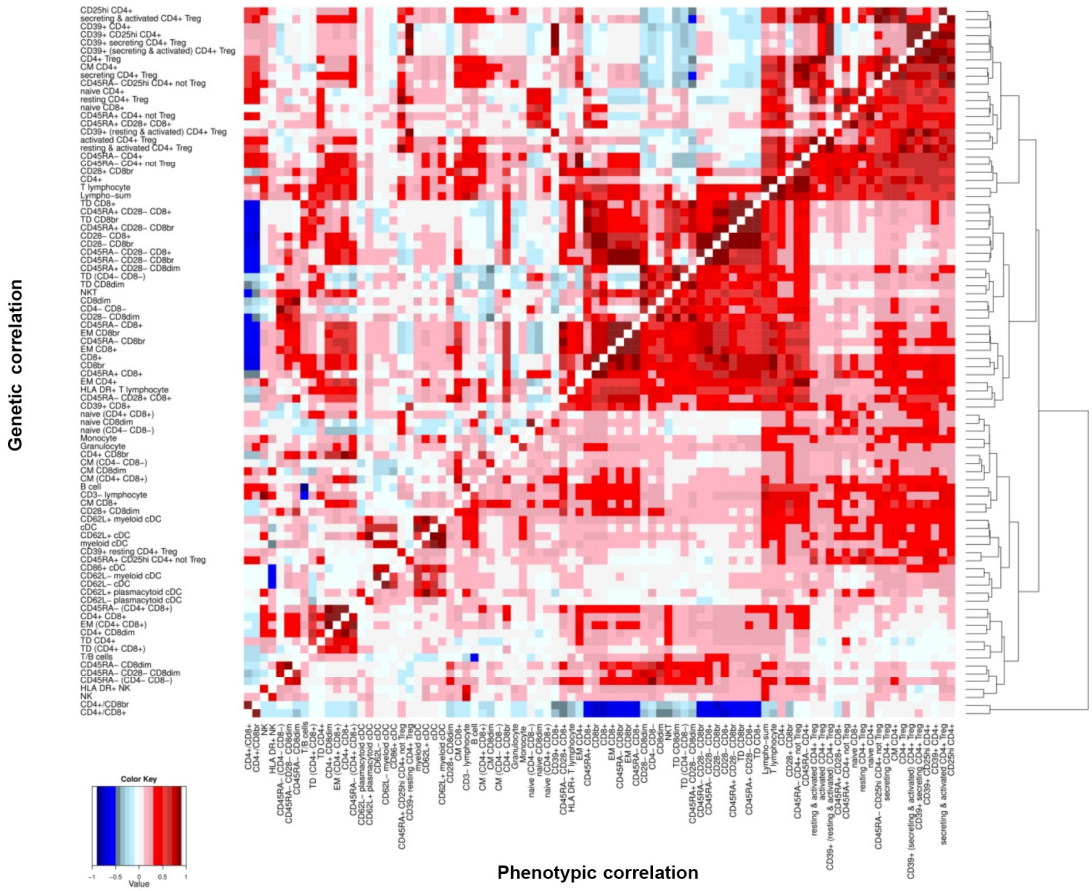


Figure 3. Manhattan plot of best p-values

For each SNP, the best p-value observed among all assessed traits is plotted on a –log10 scale (Y-axis), according to its genomic coordinates (X-axis). SNPs are colored in blue if the corresponding best p-value was directly genotyped with ImmunoChip (IC) or Cardio-MetaboChip (MC), and in gray if imputed from genomic sequencing of Sardinians. The dotted horizontal line indicates the threshold for declaring a locus genome-wide significant ($5.26x10^{-10}$). The best candidate gene is indicated nearby the peak. Loci below the significance threshold and previously described are marked with an asterisk.

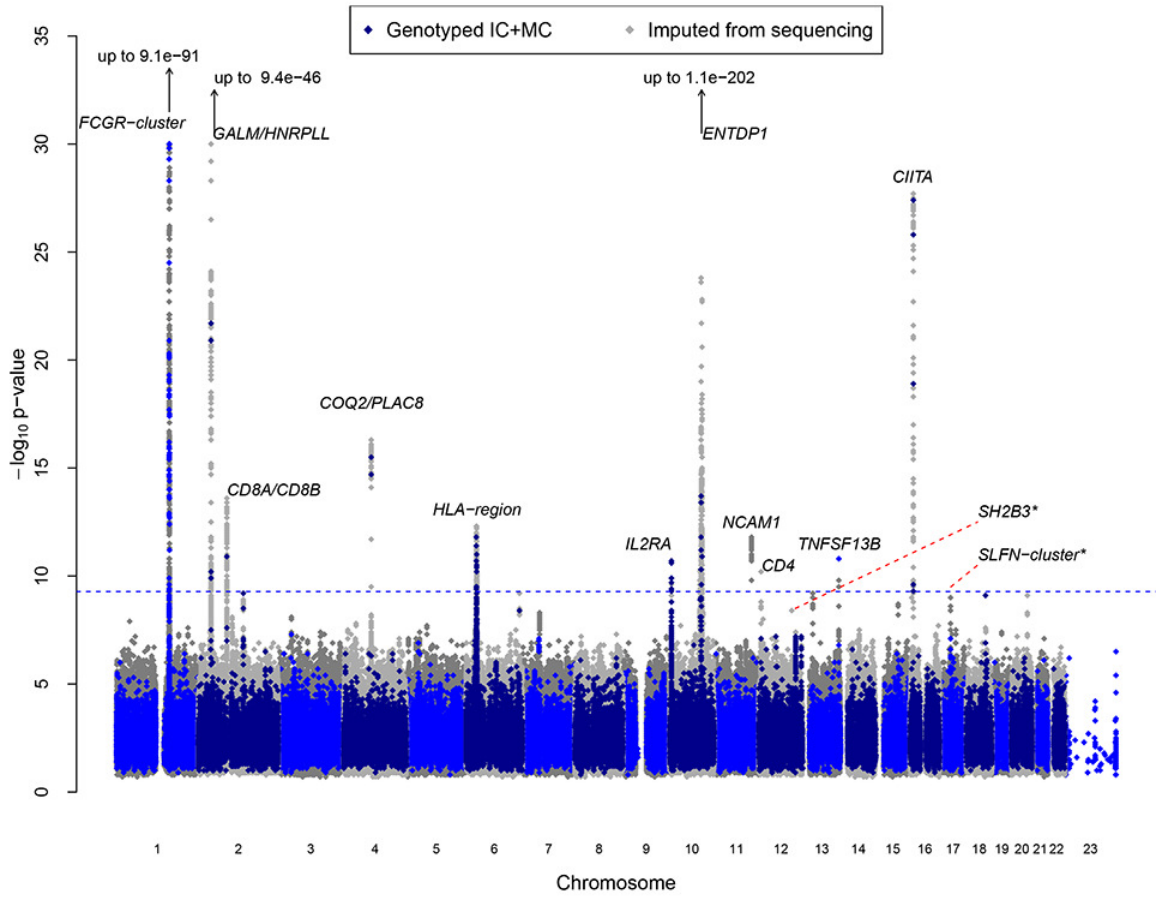


Figure 4. Regional plot and boxplot for the top signal in *ENTPD1*

These two panels represent the association in the genomic context (panel A) and in the biological context (panel B), for the most strongly associated variant at the *ENTPD1* gene. *Panel A* represents the association strength (Y axis shows the –log10 p-value) versus the genomic positions (on hg19/GRCh37 genomic build) around the most significant SNP, which is indicated with a purple circle. Other SNPs in the region are color-coded to reflect their LD with the top SNP as in the left-inset (taken from pairwise r^2 values calculated on Sardinian haplotypes), whereas symbols reflecting genomic functional annotation are indicated in the right-inset. 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 (Pruim et al., 2010). *Panel B* shows the distribution of the immunophenotypic levels within each genotype class considering the normalized trait adjusted for age and gender in relation to the 1,629 initial samples, showing the additive effect that was statistically observed. See also **Figure S3**.

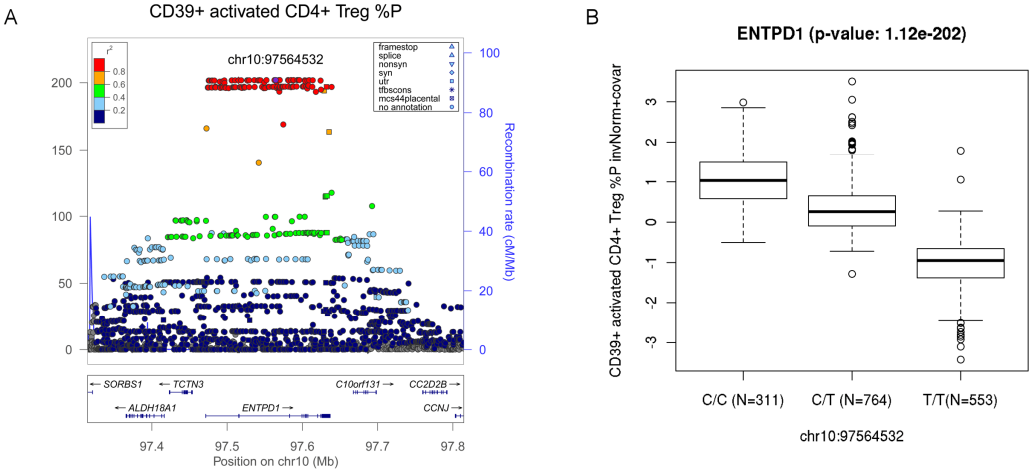
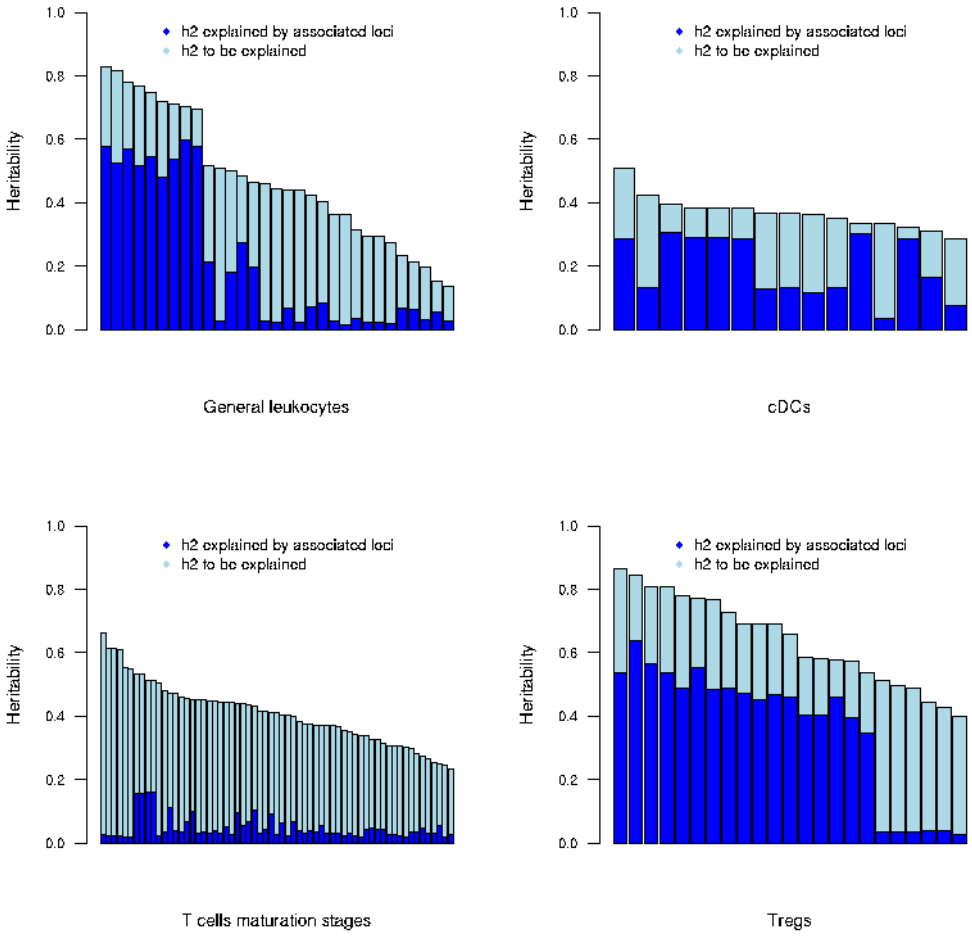


Figure 5. Proportion of heritability explained

The bar plots show the heritability of each trait (represented by a bar) for which genetic association was detected. The proportion of heritability explained by the detected loci is indicated in dark blue, while the proportion of heritability that remains to be explained is shown in light blue. Bars are grouped in their corresponding biological category as specified in **Table S1B**. See also **Table S2A.**



ACKNOWLEDGEMENTS

This work is dedicated to the memory of Prof. Antonio Cao, mentor and leader, who passed away while this manuscript was in preparation.

We thank all the volunteers who generously participated in this study and made this research possible. We thank John Todd and Marcella Devoto for critical revisions of the manuscript, Manuela Sironi and Rachele Cagliani for advice on balancing selection tests and Manuela Uda for her previous leadership on the ProgeNIA/SardiNIA study. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging, with contracts N01-AG-1-2109 and HHSN271201100005C, by Italian grants FISM 2011/R/13 “Approccio razionale per la ricerca di composti per la cura della sclerosi multipla basato sull’analisi dei target biologici individuati dagli studi di associazione sull’intero genoma in Sardegna”, FaReBio2011 “Farmaci e Reti Biotecnologiche di Qualità”, Funds MIUR/CNR for rare diseases and molecular screening, PNR-CNR Aging Program 2012-2014 and National Human Genome Research Institute grants HG005581, HG005552, HG006513 and HG007022 to G.R.A.

REFERENCES

Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M*., et al.* (2012). Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. Am J Hum Genet *91*, 794-808.

Barreiro, L.B., and Quintana-Murci L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet *11*, 17-30.

Barrett, J.C., Lee, J.C., Lees, C.W., Prescott, N.J., Anderson, C.A., Phillips, A., Wesley, E., Parnell, K., Zhang, H., Drummond, H*., et al.* (2009). Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet *41*, 1330-1334.

Bayes, B., Pastor, M.C., Lauzurica, R., Granada, M.L., Salinas, I., and Romero, R. (2007). Do anti-CD25 monoclonal antibodies potentiate posttransplant diabetes mellitus? Transplant Proc *39*, 2248-2250.

Borsellino, G., Kleinewietfeld, M., Di Mitri, D., Sternjak, A., Diamantini, A., Giometto, R., Hopner, S., Centonze, D., Bernardi, G., Dell'Acqua, M.L*., et al.* (2007). Expression of ectonucleotidase CD39 by Foxp3+ Treg cells: hydrolysis of extracellular ATP and immune suppression. Blood *110*, 1225-1232.

Buchholz, V.R., Neuenhahn, M., and Busch, D.H. (2011). CD8+ T cell differentiation in the aging immune system: until the last clone standing. Curr Opin Immunol *23*, 549-554.

Chalmin, F., Mignot, G., Bruchard, M., Chevriaux, A., Vegran, F., Hichami, A., Ladoire, S., Derangere, V., Vincent, J., Masson, D*., et al.* (2012). Stat3 and Gfi-1 transcription factors control Th17 cell immunosuppressive activity via the regulation of ectonucleotidase expression. Immunity *36*, 362-373.

Chatenoud, L. (2010). Immune therapy for type 1 diabetes mellitus-what is unique about anti-CD3 antibodies? Nat Rev Endocrinol *6*, 149-157.

Chen, W.M., and Abecasis, G.R. (2007). Family-based association tests for genomewide association scans. Am J Hum Genet *81*, 913-926.

Cortes, A., and Brown, M.A. (2011). Promise and pitfalls of the Immunochip. Arthritis Res Ther *13*, 101.

Davis, M.M. (2008). A prescription for human immunology. Immunity *29*, 835-838.

Dendrou, C.A., Plagnol, V., Fung, E., Yang, J.H., Downes, K., Cooper, J.D., Nutland, S., Coleman, G., Himsworth, M., Hardy, M*., et al.* (2009). Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. Nat Genet *41*, 1011-1015.

Devalliere, J., and Charreau, B. (2011). The adaptor Lnk (SH2B3): an emerging regulator in vascular cells and a link between immune and inflammatory signaling. Biochem Pharmacol *82*, 1391-1402.

Enciso-Mora, V., Broderick, P., Ma, Y., Jarrett, R.F., Hjalgrim, H., Hemminki, K., van den Berg, A., Olver, B., Lloyd, A., Dobbins, S.E*., et al.* (2010). A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). Nat Genet *42*, 1126-1130.

Ferreira, M.A., Mangino, M., Brumme, C.J., Zhao, Z.Z., Medland, S.E., Wright, M.J., Nyholt, D.R., Gordon, S., Campbell, M., McEvoy, B.P*., et al.* (2010). Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. Am J Hum Genet *86*, 88-92.

Fletcher, J.M., Lonergan, R., Costelloe, L., Kinsella, K., Moran, B., O'Farrelly, C., Tubridy, N., and Mills, K.H. (2009). CD39+Foxp3+ regulatory T Cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. J Immunol *183*, 7602-7610.

Gorlova, O., Martin, J.E., Rueda, B., Koeleman, B.P., Ying, J., Teruel, M., Diaz-Gallo, L.M., Broen, J.C., Vonk, M.C., Simeon, C.P*., et al.* (2011). Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. PLoS Genet *7*, e1002178.

Hamza, T.H., Zabetian, C.P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D.M., Doheny, K.F., Paschall, J., Pugh, E*., et al.* (2011). Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. Nat Genet *42*, 781-785.

Hindorff LA, M.J.E.B.I., Wise A, Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies., Available at: www.genome.gov/gwastudies. Accessed [23/05/2013].

Huang, J., Ellinghaus, D., Franke, A., Howie, B., and Li, Y. (2012). 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur J Hum Genet.

Keir, M.E., and Sharpe, A.H. (2005). The B7/CD28 costimulatory family in autoimmunity. Immunol Rev *204*, 128-143.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S*., et al.* (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J*., et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84-98.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu Rev Genomics Hum Genet *10*, 387-406.

Lowe, C.E., Cooper, J.D., Brusko, T., Walker, N.M., Smyth, D.J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M*., et al.* (2007). Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. Nat Genet *39*, 1074-1082.

Maecker, H.T., McCoy, J.P., and Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. Nat Rev Immunol *12*, 191-200.

Marrosu, M.G., Murru, R., Murru, M.R., Costa, G., Zavattari, P., Whalen, M., Cocco, E., Mancosu, C., Schirru, L., Solla, E*., et al.*(2001). Dissection of the HLA association with multiple sclerosis in the founder isolated population of Sardinia. Hum Mol Genet *10*, 2907-2916.

McGovern, D.P., Gardet, A., Torkvist, L., Goyette, P., Essers, J., Taylor, K.D., Neale, B.M., Ong, R.T., Lagace, C., Li, C*., et al.* (2010). Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet 42, 332-337.

Mackay F, Schneider P. (2009). Nat Rev Immunol. Cracking the BAFF code.9, 491-502.

Miyara, M., Yoshioka, Y., Kitoh, A., Shima, T., Wing, K., Niwa, A., Parizot, C., Taflin, C., Heike, T., Valeyre, D.*, et al.* (2009). Functional delineation and differentiation dynamics of human CD4+ T cells expressing the FoxP3 transcription factor. Immunity *30*, 899-911.

Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A.*, et al.* (2011). Multiple loci are associated with white blood cell phenotypes. PLoS Genet *7*, e1002113.

Ohnmacht, C., Pullner, A., King, S.B., Drexler, I., Meier, S., Brocker, T., and Voehringer, D. (2009). Constitutive ablation of dendritic cells breaks self-tolerance of CD4 T cells and results in spontaneous fatal autoimmunity. J Exp Med *206*, 549-559.

Okada, Y., Hirota, T., Kamatani, Y., Takahashi, A., Ohmiya, H., Kumasaka, N., Higasa, K., Yamaguchi-Kabata, Y., Hosono, N., Nalls, M.A.*, et al.* (2011). Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. PLoS Genet *7*, e1002067.

Pilia, G., Chen, W.M., Scuteri, A., Orru, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P.*, et al.* (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet *2*, e132.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336-2337.

Pulte, E.D., Broekman, M.J., Olson, K.E., Drosopoulos, J.H., Kizer, J.R., Islam, N., and Marcus, A.J. (2007). CD39/NTPDase-1 activity and expression in normal leukocytes. Thromb Res *121*, 309-317.

Sallusto, F., Lenig, D., Forster, R., Lipp, M., and Lanzavecchia, A. (1999). Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. Nature *401*, 708-712.

Sansoni, P., Vescovini, R., Fagnoni, F., Biasini, C., Zanni, F., Zanlari, L., Telera, A., Lucchini, G., Passeri, G., Monti, D.*, et al.* (2008). The immune system in extreme longevity. Exp Gerontol *43*, 61-65.

Shevach, E.M. (2000). Regulatory T cells in autoimmmunity*. Annu Rev Immunol *18*, 423-449.

Steinman, R.M., and Banchereau, J. (2007). Taking dendritic cells into medicine. Nature *449*, 419-426.

Takada, K., and Jameson, S.C. (2009). Self-class I MHC molecules support survival of naive CD8 T cells, but depress their functional sensitivity through regulation of CD8 expression levels. J Exp Med *206*, 2253-2269.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J.*, et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707-713.

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G.*, et al.* (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet *43*, 1193-1201.

Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G.*, et al.* (2008). Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc Natl Acad Sci U S A *105*, 1620-1625.

Vendrame, F., Pileggi, A., Laughlin, E., Allende, G., Martin-Pagola, A., Molano, R.D., Diamantopoulos, S., Standifer, N., Geubtner, K., Falk, B.A.*, et al.* (2010). Recurrence of type 1 diabetes after simultaneous pancreas-kidney transplantation, despite immunosuppression, is associated with autoantibodies and pathogenic autoreactive CD4 T-cells. Diabetes *59*, 947-957.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J.*, et al.* (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet *8*, e1002793.

Wing, K., and Sakaguchi, S. (2010). Regulatory T cells exert checks and balances on self tolerance and autoimmunity. Nat Immunol *11*, 7-13.

Wu, Z., Yates, A.L., Hoyne, G.F., and Goodnow, C.C. (2010). Consequences of increased CD45RA and RC isoforms for TCR signaling and peripheral T cell deficiency resulting from heterogeneous nuclear ribonucleoprotein L-like mutation. J Immunol *185*, 231-238.

Chapter 5: Genome sequencing elucidates Sardinian genetic architecture and augments GWAS findings: the examples of lipids and blood inflammatory markers

Based on:

Sidore C[*], Busonero F[*], Maschio A[*], Porcu E[*], Naitza S[*], Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, Kwong A, Ortega del Vecchyo VD, Chiang CWK, Bragg-Gresham J, Pitzalis M, Nagaraja R, Tarrier B, Brennan C, Uzzau S, Fuchsberger C, Atzeni R, Reinier F, Berutti R, Huang J, Timpson NJ, Toniolo D, Gasparini P, Malerba G, Dedoussis G, Zeggini E, Soranzo N, Jones C, Lyons R, Angius A, Kang HM, Novembre J, ***Sanna S***[#]*, Schlessinger D[#], Cucca F[#], Abecasis GR[#]*

Nature Genetics 2015. 47, 1272–1281 (2015) doi:10.1038/ng.3368 *Epub 15 Sept 2015*

[*,#] *indicate equal contributions*

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 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 AND DISCUSSION

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 >65x 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). This suggests that part of the reservoir of variants that have drifted to higher frequency in Sardinia could be especially informative for genetic association and functional studies (**Supplementary Figure 1**). The results presented here show a few clear examples.

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 relatively 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 from distinct individuals) ([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 more 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 differences in haplotype length are less marked around variants with higher frequencies, and hence shared in more than 2 heterozygous individuals (i.e. f3, f4, etc). This is evident even when comparing samples from the Lanusei valley and elsewhere in the island (**Supplementary Table 3**).

Relatedness and imputation for the Lanusei valley samples

Participants in the SardiNIA study all live in four small 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]). A total of 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 accuracy improved markedly in comparison to the imputation results based on 1000 Genomes Project Phase 3 panel, that includes 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 genetic association: the examples of lipid and inflammatory marker levels

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 $5x10^{-8}$ 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 $6.9x10^{-9}$, 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 from the known variant rs73015013 (frequency of 14%, effect -5.2 mg/dl, $p=6.4x10^{-8}$, r2<0.001). Here we report a novel association for triglycerides levels with a missense variant in *APOA5* (frequency 3% in Sardinia, effect -20.7 mg/dl, $p=1.2x10^{-12}$) (**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 only on 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, $p=4.6x10^{-9}$, 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 proteins 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 ($p=1.2x10^{-20}$) and 16.9 mg/dl lower total cholesterol levels ($p=1.2x10^{-22}$). Of note, this variant 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%). Hence, the signal in this region would have been much weaker and would likely be missed or misinterpreted. For example, the analysis after 1000 Genomes Phase 3 imputation points only to an intergenic marker (rs76053862) 122 kb away from the *beta(0)39* variant, the second most associated SNP using the Sardinian reference panel, with a much lower association signal ($p = 1.4x10^{-13}$) (**Figure 3**). Finally, two additional signals were observed for total cholesterol levels at SNP rs115048493 near genes *TMEM33* and *DCAF4L1*($p=6.94x10^{-9}$) and with HDL-c at SNP rs8092903 near *TGIF1* in females ($p=4.49x10^{-8}$) (**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 $6.9x10^{-9}$ 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 $5x10^{-8}$ threshold, with five variants in four 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 $6.9x10^{-9}$. Compared to analysis 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, $p=1.1x10^{-28}$) and ESR (12:125406340, $p=4.4x10^{-23}$) on chromosome 12, in a stretch of rare variants encompassing several genes (**Figure 4**). The lead variants were not the same but 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 $<1x10^{-15}$ (**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 that is extremely rare or absent outside Sardinia. Consequently, replication would require genetic testing in additional samples from Sardinia or in very large Southern European cohorts.

For hsCRP, we detected additional tentative 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 $5x10^{-8}$ threshold, but not the more stringent threshold for novel findings: one for ADPN, at 13:108884835 near the gene *ABHD13* $p=3.3x10^{-8}$, and another for MCP-1, at rs76135610, $p=1.8x10^{-8}$, 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 on *HBB* gene (**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 $5x10^{-6}$ 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* ($p=4.7x10^{-10}$) and adiponectin levels, and another for *PTPRH* ($p=8.3x10^{-7}$) and ESR levels. These signals, however, were not further investigated (**Table 4, Supplementary Table 10**), 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 11, 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 Table 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.

Overall, the results demonstrate the value of whole genome sequencing-based association studies in this founder population, in which 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]). On the other hand, 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.

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).

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 ([32,50]). We used known family relationships among SardiNIA study participants and the ExomePicks program (http://genome.sph.umich.edu/wiki/ExomePicks) 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 ([51]), immune disorders - ImmunoChip ([52]), and coding variation – ExomeChip, (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). 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) ([53]). 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 ([54]). 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 4.16X. A single nuclear family (two parents and one child) was sequenced to average depth >65x 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 at *ftp://ftp.1000genomes.ebi.ac.uk*) using BWA-0.5.9 ([55]), 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([56]) and verifyBamId ([57]) 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]).

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 ([58]), using 30 iterations of the haplotyping Markov chain and 400 states per iteration. Imputation used minimac software ([59]) 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 ([60]). 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 ([61]). 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 ([62]), 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 ([63]).

Rare variant analysis

We performed two regional-based tests: the Combined Multivariate and Collapsing (CMC)([64]) and the variable thresholds method (VT) ([65]). 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 $5x10^{-6}$ 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 ([66]). 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 ($p < 5 x10^{-8}$) 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)([67]). 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)([68]). 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 $6.9x10^{-09}$ and of $1.4x10^{-08}$ 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)([69–72]); ESR, MCP-1 and ADPN values were not available in those samples. In TwinsUK and ALSPAC we analysed genotypes from whole-genome sequence data([73]), 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.

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/

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.* **39,** 830–832 (2007).
2. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).
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.* **107,** 7401–7406 (2010).
4. Do, R. *et al.* Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.* **45,** 1345–1352 (2013).

5. Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum. Mol. Genet.* **21,** R1–9 (2012).
6. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455–E464 (2014).
7. Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A. & Sunyaev, S. R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 3871–3876 (2009).
8. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* **1,** 182–190 (2000).
9. Clarke, R. *et al.* Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men. *Arch. Intern. Med.* **167,** 1373–1378 (2007).
10. Pai, J. K. *et al.* Inflammatory markers and the risk of coronary heart disease in men and women. *N. Engl. J. Med.* **351,** 2599–2610 (2004).
11. Orru, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell* **155,** 242–56 (2013).
12. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).
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* **8,** e1002480 (2012).
14. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2,** e132 (2006).
15. Sanna, S. *et al.* Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nat. Genet.* **42,** 495–497 (2010).
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.* **10,** 15–17 (2009).
17. Chen, W. *et al.* Genotype calling and haplotyping in parent-offspring trios. *Genome Res.* **23,** 142–151 (2013).
18. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Res.* (2015). doi:10.1101/gr.176552.114
19. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21,** 940–951 (2011).
20. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinforma. Oxf. Engl.* **26,** 2069–2070 (2010).
21. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
22. Francalacci, P. *et al.* Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. *Am. J. Phys. Anthropol.* **121,** 270–279 (2003).
23. Francalacci, P. *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341,** 565–569 (2013).
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.* **9,** 2947–2957 (2000).
25. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human

genomes. *Nature* **491,** 56–65 (2012).
26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
27. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456,** 98–101 (2008).
28. Nelson, M. R. *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* **83,** 347–358 (2008).
29. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 11983–11988 (2011).
30. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104 (2012).
31. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10,** e1004528 (2014).
32. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81,** 913–926 (2007).
33. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).
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* **7,** e1002198 (2011).
36. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med.* **12,** 61–76 (2010).
37. Maioli, M. *et al.* Plasma lipoprotein composition, apolipoprotein(a) concentration and isoforms in beta-thalassemia. *Atherosclerosis* **131,** 127–133 (1997).
38. Maioli, M. *et al.* Plasma lipids in beta-thalassemia minor. *Atherosclerosis* **75,** 245–248 (1989).
39. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
40. Hou, S. *et al.* Genetic variant on PDGFRL associated with Behçet disease in Chinese Han populations. *Hum. Mutat.* **34,** 74–78 (2013).
41. Xu, M. *et al.* An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics* **9 Suppl 1,** S12 (2008).
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* **92,** 2147–2156 (1998).
43. Shi, X.-F. *et al.* Structural analysis of human CCR2b and primate CCR2b by molecular modeling and molecular dynamics simulation. *J. Mol. Model.* **8,** 217–222 (2002).
44. Schick, U. M. *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* **209,** 487–491 (2010).
46. Kullo, I. J. *et al.* Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am. J. Hum. Genet.* **89,** 131–138 (2011).
47. Schadt, E. E. *et al.* Mapping the Genetic Architecture of Gene Expression in Human Liver. *PLoS Biol* **6,** e107 (2008).

48. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol. Clifton NJ* **1019,** 215–236 (2013).
49. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512,** 190–193 (2014).
50. Burdick, J. T., Chen, W.-M., Abecasis, G. R. & Cheung, V. G. In silico method for inferring genotypes in pedigrees. *Nat. Genet.* **38,** 1002–1004 (2006).
51. Voight, B. F. *et al.* The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* **8,** e1002793 (2012).
52. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.* **14,** 661–673 (2013).
53. Goldstein, J. I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinforma. Oxf. Engl.* **28,** 2543–2545 (2012).
54. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–369 (2008).
55. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **26,** 589–595 (2010).
56. Li, B. *et al.* QPLOT: a quality assessment tool for next generation sequencing data. *BioMed Res. Int.* **2013,** 865181 (2013).
57. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91,** 839–848 (2012).
58. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
59. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
60. Price, A. L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4,** e236 (2008).
61. Lee, S., Zou, F. & Wright, F. A. CONVERGENCE AND PREDICTION OF PRINCIPAL COMPONENT SCORES IN HIGH-DIMENSIONAL SETTINGS. *Ann. Stat.* **38,** 3605–3629 (2010).
62. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42,** 348–54 (2009).
63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinforma. Oxf. Engl.* **26,** 2190–2191 (2010).
64. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).
65. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–838 (2010).
66. Zaitlen, N. *et al.* Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet* **9,** e1003520 (2013).
67. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet. Epidemiol.* **38,** 281–290 (2014).
68. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30,** 97–101 (2002).

69. Moayyeri, A., Hammond, C. J., Valdes, A. M. & Spector, T. D. Cohort Profile: TwinsUK and healthy ageing twin study. *Int. J. Epidemiol.* **42,** 76–85 (2013).
70. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur. J. Hum. Genet. EJHG* **21,** 659–665 (2013).
71. 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* **4,** e7554 (2009).
72. Winkelmann, B. R. *et al.* Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics* **2,** S1–73 (2001).
73. Taylor, P. N. *et al.* Whole-genome sequence-based analysis of thyroid function. *Nat. Commun.* **6,** 5681 (2015).
74. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–7 (2010).

Figure Legends

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 (evaluated in 360 chromosomes) (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.

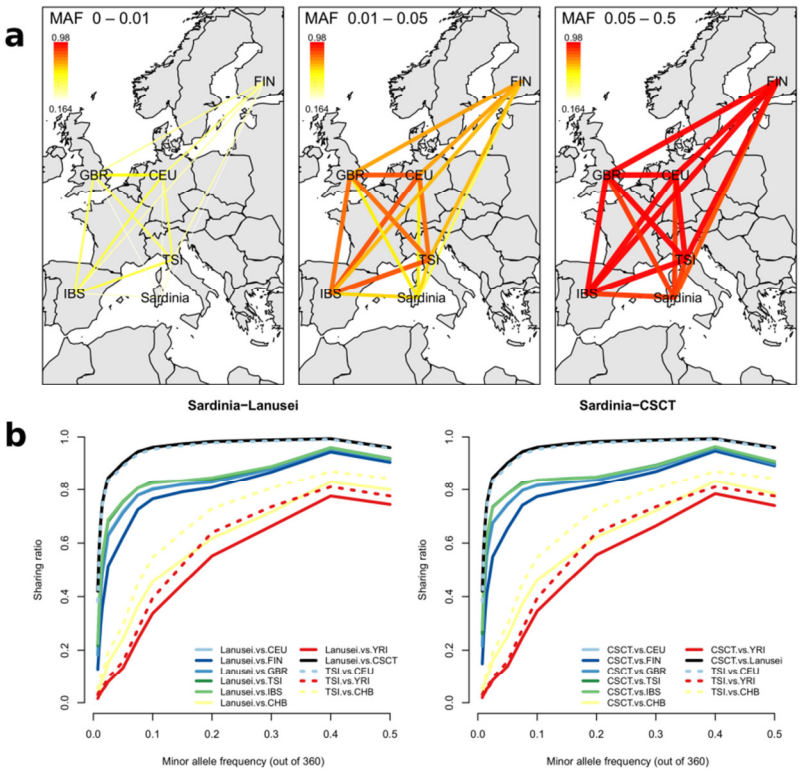


Figure 2. Length of shared haplotypes surrounding f_2 variants within Sardinians and populations in 1000 Genomes.

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 variants in each comparison. Panel c) shows the number of f_2 variants in comparisons within Sardinia (note the wider Y-axis range). SNIA: Sardinians from the Lanusei Valley.

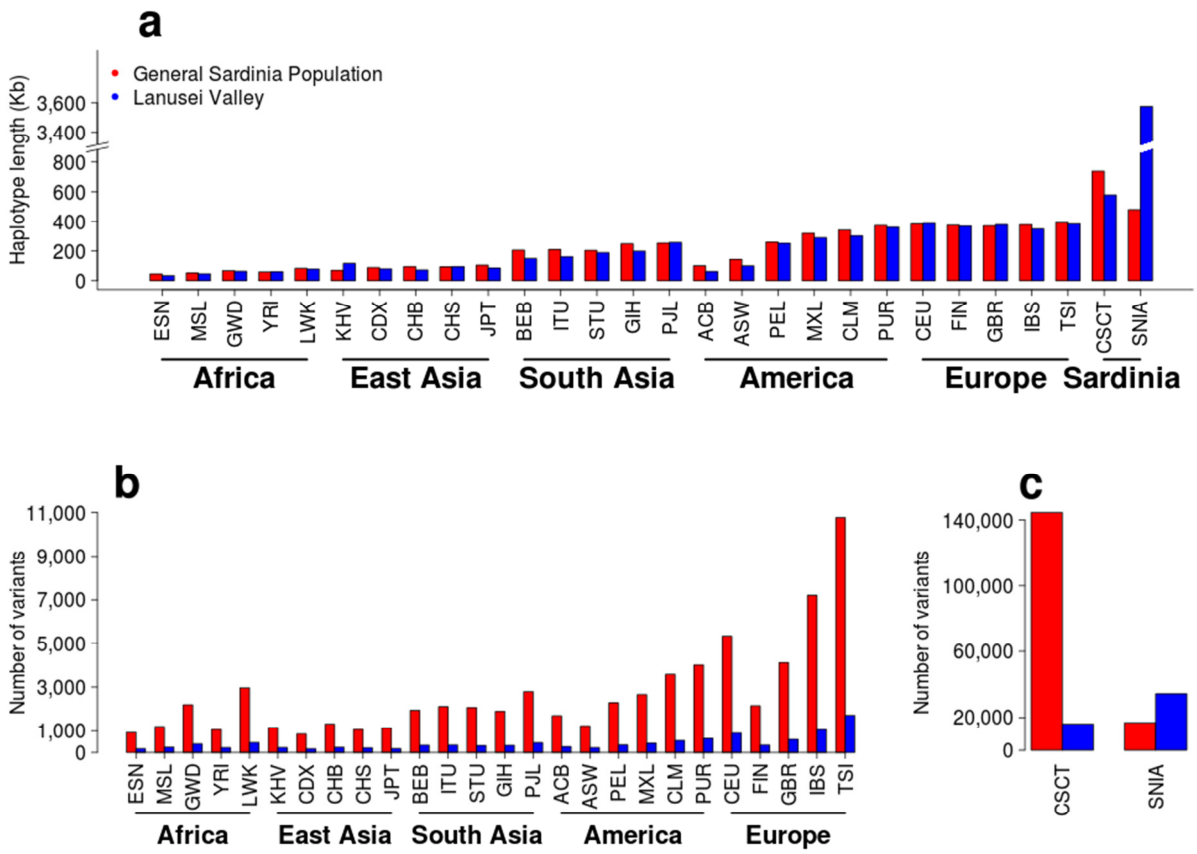


Figure 3. Regional association plots for novel lipids loci.

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 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 ([74]).

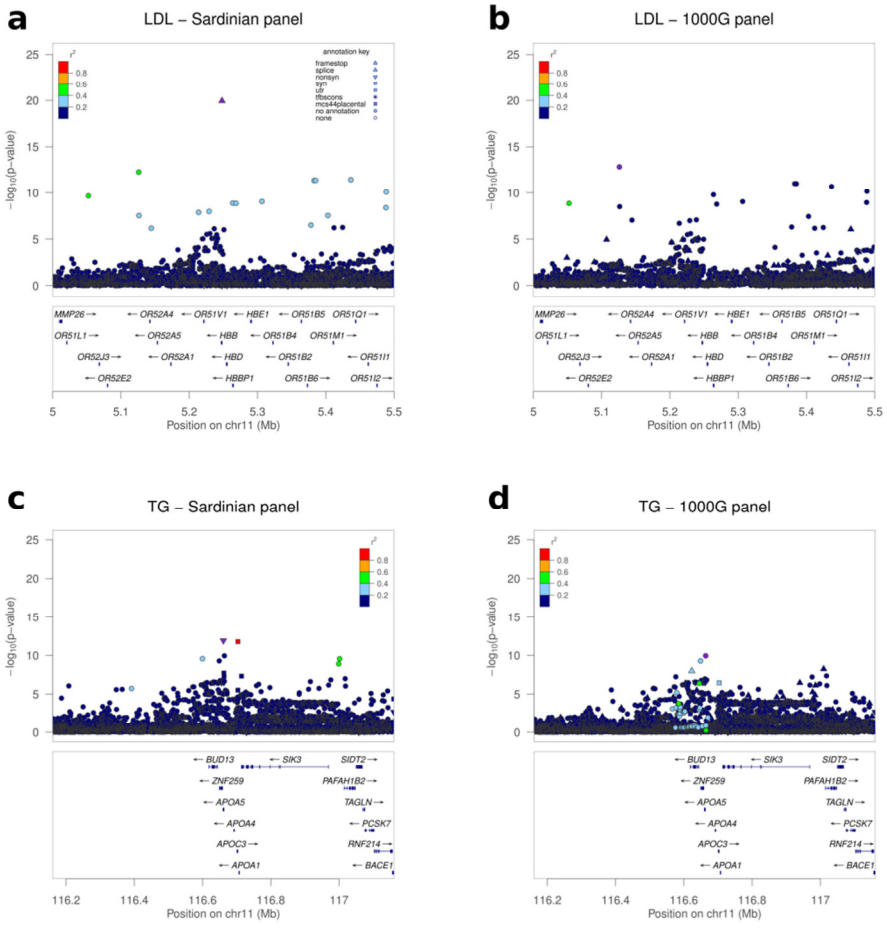


Figure 4. Regional association plot at chromosome 12 for hSCRP and ESR.

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.

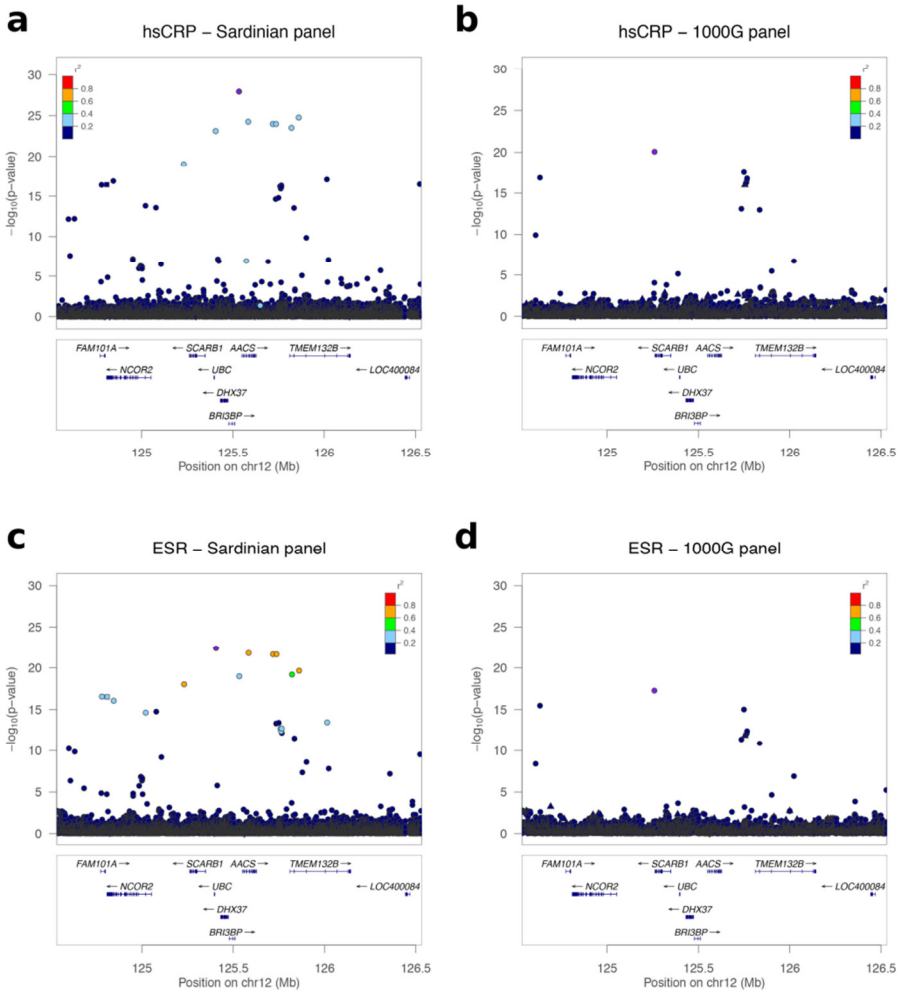


Table 1. Summary of Discovered Variants

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			*** 4.16X ***				
				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

Table 2. Summary of Lipid Association Results

The table lists association signals that reach p <$5x10^{-8}$ 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
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)**	**1.24×10^{-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

Table 3. Summary of Inflammatory Marker Association Results

The table shows the association results at that reach $p < 5x10^{-8}$ 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)	$1.35x10^{-22}$	1.6	Genotyped	Intergenic	--
ABHD13	13:108884835	N/A	A/G	0.001	-1.519 (0.275)	$3.35x10^{-08}$	0.5	0.921	3'UTR	--
hsCRP										
CRP	1:159684665	rs3091244	A/G	0.428	0.207 (0.019)	$5.28x10^{-27}$	2.0	Genotyped	Intergenic	0.249
PDGFRL	**8:17450500**	**rs73198138**	**A/G**	**0.004**	**-0.894 (0.151)**	**$3.31x10^{-09}$**	**0.6**	**0.977**	**Intronic**	**--**
HNF1A	*12:121415293* [a]	*rs7139079*	*G/A*	*0.377*	*-0.118 (0.020)*	*$2.11x10^{-09}$*	0.6	*0.998*	*Intergenic*	*0.710*
BRI3BP, AACS	**12:125533106**	**rs183233091**	**A/G**	**0.010**	**1.054 (0.094)**	**$1.09x10^{-28}$**	**2.1**	**0.941**	**Intergenic**	**--**
APOE	19:45411941	rs429358	C/T	0.073	-0.237 (0.036)	$3.78x10^{-11}$	0.7	1	Missense, C130R	0.565
ESR										
RHCE	*1:25724005* [b]	*rs630337*	*T/C*	*0.297*	*-0.109 (0.020)*	*$4.03x10^{-08}$*	*0.5*	*0.957*	*Intronic*	*0.797*
CR1	1:207684359	rs11117956	T/G	0.400	-0.153 (0.018)	$9.43x10^{-18}$	1.2	Genotyped	Intronic	0.989
HBB	11:5248004	rs11549407	A/G	0.048	-0.437 (0.042)	$1.02x10^{-25}$	1.8	0.918	Stop gained, Q40X	0.330
AACS, MIR5188	**12:125406340**	**N/A**	**G/A**	**0.007**	**1.034 (0.104)**	**$4.40x10^{-23}$**	**1.6**	**0.952**	**Intergenic**	**--**
MCP-1										
DARC,CADM3	1:159175354	rs12075	G/A	0.446	-0.405 (0.019)	$1.08x10^{-96}$	7.2	Genotyped	Missense, G44D	Same SNP
DARC,CADM3	*1:159164454* [c]	*rs2852718*	*C/T*	*0.022*	*-0.515 (0.063)*	*$3.34x10^{-16}$*	*1.1*	*0.999*	*Intronic*	*0.005*
DARC,CADM3	***1:159175494*** [d]	***rs34599082***	***T/C***	***0.037***	***-0.338 (0.049)***	***$8.23x10^{-12}$***	***0.8***	Genotyped	***Missense, R89C***	**--**
CCR2, CCR3	3:46383906	rs113403743	T/G	0.099	0.273 (0.034)	$1.47x10^{-15}$	1.1	0.997	Intergenic	0.988
CCR2	***3:46399764*** [e]	***rs200491743***	***A/T***	***0.005***	***0.799 (0.130)***	***$9.94x10^{-10}$***	***0.6***	Genotyped	***Missense, M249K***	**--**
*N4BP1, CBLN1**	16:49072490	rs76135610	T/C	0.005	0.969 (0.172)	$1.76x10^{-08}$	0.9	0.915	Intergenic	--
IL-6										
IL6R	1:154428283	rs12133641	G/A	0.255	0.118 (0.020)	$6.87x10^{-09}$	0.6	1	Intronic	0.998
ABO	9:136142355	rs643434	A/G	0.263	-0.221 (0.020)	$5.80x10^{-27}$	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)

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). Pvalues that remain significant after adjustment are marked in bold.

Gene	Chr:Start-end	#SNPs	#Pass	Burden Count	Fraction with 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)	$4.71x10^{-10}$	0.007	**$1.92x10^{-09}$**	+	$1.00x10^{-07}$	0.007	**$1.00x10^{-07}$**
MCP1													
CCR2	3:46399158-46401290	4	3	105	0.01797	0.541 (0.104)	$1.84x10^{-07}$	0.005	0.7092	+	$1.00x10^{-06}$	0.005	0.92
IFI16	1:158979950-159024668	10	8	567	0.09702	0.218 (0.046)	$2.50x10^{-06}$	0.004	0.1564	+	$1.40x^{-05}$	0.003788	0.115
ESR													
HBB	11:5247914-5248004	2	2	613	0.10318	-0.345 (0.039)	$9.77x10^{-19}$	0.013	0.015	-	$1.00x10^{-07}$	0.013	0.025
OR52H1	11:5565906-5566751	5	3	529	0.08904	-0.205 (0.042)	$1.23x10^{-06}$	0.004	0.345	-	$3.40x10^{-06}$	0.004	0.69
PTPRH	19:55693244-55716713	22	15	1152	0.19391	-0.146 (0.029)	$8.31x10^{-07}$	0.004	**$4.22x10^{-06}$**	-.	$1.18x10^{-05}$	0.0041	$1.90x10^{-05}$

.

Chapter 6: Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels

Based on:

Danjou F[*], Zoledziewska M[*], Sidore C , Steri M, Busonero F , Maschio A, Mulas A, Perseu L, Barella S, Porcu E, Pistis G, Pitzalis M, Mauro Pala M, Menzel M, Metrustry S, Spector TD, Leoni L, Angius A, Uda M, Moi P, Thein SL, Galanello R, Abecasis GR, Schlessinger D, ***Sanna S***[#]*, Cucca F*[#]

Nature Genetics 2015. In press

[*,#] *indicate equal contributions*

ABSTRACT

We report GWAS results for the levels of A1, A2 and fetal hemoglobins, analyzed for the first time concurrently. Integrating high-density array genotyping and whole-genome sequencing in a large general population cohort from Sardinia, we detected 23 associations at 10 loci. Five are due to variants at previously undetected loci: *MPHOSPH9*, *PLTP-PCIF1*, *FOG1, NFIX,* and *CCND3*. Among those at known loci, 10 are new lead variants and 4 are novel independent signals. Half of all variants also showed pleiotropic associations with different hemoglobins, which further corroborated some of the detected associations and revealed features of coordinated hemoglobin species production.

INTRODUCTION

The provision of oxygen to tissues depends on hemoglobin, requiring the coordinated expression of several globin chains that form functional tetramers. An index of the importance of hemoglobin function is the evolutionary duplication and divergence of regulation of globin gene copies to adapt to stages of development and buffer the effects of mutational loss. In particular, at birth, a switch occurs from fetal hemoglobin (HbF) toward hemoglobin A2 (HbA2) and hemoglobin A1 (HbA1), so that during adult life the hemoglobin forms comprise ~1 % HbF, ~3 % HbA2 and ~96 % HbA1. The different hemoglobins all contain alpha-globin chains, encoded by two eponymous genes on chromosome 16. Those aggregate with non-alpha-globin chains encoded, respectively, by the gamma (for HbF), delta (for HbA2) and beta-globin (for HbA1) genes in the "beta-globin gene cluster" on chromosome 11 (**Figure 1**). The molecular switch between fetal and adult hemoglobin occurs via the binding of transcription factors to regulatory DNA sequences controlling the expression of globin genes. In particular, the various genes in the beta-globin cluster are sequentially activated during ontogeny, so that time-specific expression patterns follow their genomic order[1].

Inherited disorders of hemoglobin, such as beta-thalassemia caused by mutations at the hemoglobin beta (*HBB*) locus, represent the most common monogenic disorders worldwide[2]. Prevalence is highest in areas where malaria was or remains endemic[3]. The severity of inherited hemoglobin disorders is also variable, from severe life-long transfusion-dependent anemia to mild anemia that does not require transfusion, depending on the molecular defect and genotype status as well as ameliorating variants in modifier genes. Therefore, studying the genetic regulation of hemoglobin levels might reveal new factors and mechanisms to optimize strategies for the therapy of the disorders.

The large heritable contribution to phenotypic variance of HbA2 and HbF in the general population (0.728 and 0.633 respectively; see **Methods** and previous report[4]) indicates that genetic analyses could lead to new insights. In genome-wide association studies (GWAS), two genomic regions, the beta-globin gene cluster locus and the *HBS1L-MYB* locus, have been associated at a genome-wide significant level with variations in the amount of HbA2[5], and only those loci and *BCL11A* have been associated with HbF levels[6,7]. Variants at all four loci are powerful modifiers of the severity of beta-thalassemia and sickle-cell disease[7–10]. Notably, none of the variants associated with HbA2 or HbF have been found associated with total hemoglobin, even in the largest meta-analysis of over 135,000 individuals[11]. This indicates that in analyses of total hemoglobin levels, association signals for subtypes are diluted and possibly obscured by opposite directions of effects. Currently, most of the HbF and HbA2 heritability also remains to be explained, and HbA1 variation has never been specifically assessed by GWAS at all.

A promising source to extend analyses is the founder Sardinian population, in which previous associations have been detected in a large cohort through the analysis of genotyping arrays bearing common/ubiquitous variants[7]. Here, we extend these analyses to rarer and Sardinian-specific variants inferred from whole-genome population sequencing in the same cohort (see **Supplementary Note** and **Supplementary Figure 1**). Furthermore, analyzing variants modulating HbA1, HbA2 and HbF levels concurrently in a single cohort provides a route to assess associations that overlap for different hemoglobin forms without the need to account for differences in study size, ethnic background or measurements.

RESULTS

To test for genetic associations with the levels of HbA1, HbA2 and HbF, we interrogated ~10.9 million single nucleotide polymorphisms (SNPs), genotyped or imputed in 6,602 general population volunteers of the SardiNIA longitudinal study[4] (see **Methods** and **Supplementary Table 1**).

Initial analyses showed a predominant role for the HBB:c.118C>T stop-codon mutation -- Q40X, better known as beta(0)39 mutation -- a variant common in Sardinia (rs11549407, allele frequency 4.8 %). It results in complete absence of beta-globin chain synthesis ($beta^0$) and consequent beta-thalassemia in homozygous individuals, and in a decrease of HbA1 and increase of HbA2 and HbF in heterozygous individuals (with p-values < $1.0x10^{-200}$). Because its effect has been established previously[7,12], we considered this mutation and other rarer $beta^0$-thalassemia mutations known in Sardinia as covariates (see **Methods** and **Supplementary Table 2**). The assessed individuals in the cohort include 664 healthy heterozygous carriers but no $beta^0$-thalassemia patients.

The genome-wide scan revealed 23 unique variants at 10 loci at the classical $5x10^{-08}$ threshold. Of note, 21 are significant even considering a more stringent threshold of p = $1.4x10^{-8}$, calculated based on an empirical estimate of the number of independent tests in the Sardinian genome (see Chapter 5).

Five variants are at previously undetected loci, 4 are new independent signals at known loci, and 10 refine previously described associations to new lead polymorphisms that may have functional effects (**Table 1**). Six, 14 and 8 independent genome-wide significant signals were seen for HbA1, HbA2 and HbF respectively (**Supplementary Figure 2**). Hence, some of the associated variants significantly affected more than one hemoglobin, resulting in 28 variant-trait associations (see **Table 1**, and **Supplementary Table 3**). Variants resulting from imputation and not supported by linked genotyped markers were experimentally validated (**Supplementary Table 4**)

Novel associations at new loci

Novel associations were detected for all 3 hemoglobin forms. For HbA1, we observed a signal led by chr12:123681790 (in an intron of *MPHOSPH9*), encompassing several SNPs in complete linkage disequilibrium (LD) in a region encoding several genes (see **Supplementary Figure 3**). Which gene is truly associated, and how it affects hemoglobin production, remains unclear, although among the top associated SNPs, a variant in an intron of *ARL6IP4* (chr12:123465483) falls in a highly conserved region rich in putative transcription factor binding sites and has the highest score for in-silico prediction of deleterious impact on function (CADD score)[13] as detailed in **Supplementary Table 2**. Although this association is just below the more stringent empirical threshold of significance, it is further strengthened by independent association with another hemoglobin form (HbA2, p = $5.9x10^{-5}$), as detailed in **Table 1**.

For HbA2, we identified 3 novel signals. One, rs141006889, is a missense variant located in *ZFPM1*, a gene also known as *FOG1* that encodes a cofactor of the hematopoietic transcription factors GATA1 and GATA2[14] (**Supplementary Figure 4**). The complexes formed by FOG1 and GATA proteins are essential for normal erythroid differentiation[14], as demonstrated by pathogenetic mutations that abrogate the FOG-GATA interaction to cause familial dyserythropoietic anemia and thrombocytopenia[15]. Another signal is defined by a pair of statistically indistinguishable variants, rs113267280 and rs112233623 (p-values: $1.11x10^{-29}$ and $1.29x10^{-29}$), located in *CCND3* gene, whose product, cyclin D3, is thought to be critical for erythropoiesis[16]. Knockdown of cyclin D3 correlates with reduction in the number of cell divisions during terminal erythropoiesis, thereby producing fewer and larger red blood cells[17]. These variants are also in partial LD with rs9349205 ($r^2 = 0.40$), a SNP previously associated with mean red blood cell volume and number (see **Supplementary Table 6**), which falls 160bp away from rs112233623 in the same erythroid specific enhancer functionally associated with *CCND3*[17–19]. The latter is also the associated variant with highest CADD score (see **Supplementary Table 5**).

An additional variant related to HbA2, rs59329875, was observed for the first time in this study. It is situated between *PLTP,* which has been associated with several plasma lipoprotein and triglyceride levels[20–23], and *PCIF1,* which is thought to negatively regulate gene expression by RNA polymerase II[24].

As for HbF, we identified one new variant associated with its level: rs183437571, located on chromosome 19 in an intron of *NFIX,* which encodes a CCAAT-binding transcription factor. This variant is just below the empirical significance threshold of $p = 1.4x10^{-8}$ but is supported by considerable biological evidence implicating the gene and the surrounding region in hemoglobin regulation. Specifically, rs183437571 falls in a CpG region that is differentially methylated in fetal and adult red blood cell progenitors[25]. In mice, *Nfix* was recently identified as one of the regulatory factors with relatively restricted expression in hematopoietic stem cells,[26] and required for the survival of hematopoietic stem and progenitor cells during stress hematopoiesis[27]. Intriguingly, *NFIX* is situated in a region of ~300 Kb that encompasses a number of genes involved in erythropoiesis (*DNASE2* and *KLF1*)[28–32] or otherwise associated with red blood cell traits, including mean corpuscular hemoglobin (*SYCE2*, *FARSA* and *CALR*)[11] (**Supplementary Figure 5** and **Supplementary Table 6**). *KLF1* is a particularly interesting candidate gene[32,33], but mutations observed in previous studies[34] were not found and the gene itself is situated in an LD block distinct from our association signal. However, long distance regulatory interactions remain a possibility.

Of the 5 novel signals, the discovery of chr12:123681790 for HbA1, rs141006889 for HbA2, and rs183437571 for HbF were strongly influenced by the assessment of variants from Sardinian whole-genome sequencing. Specifically, chr12:123681790 was missing in 1000 Genomes phase III[35], and using this public reference panel the signal was misplaced to another variant ~1Mb away; rs141006889 was included in the design of one genotyping array (ExomeChip) after it was identified through our sequencing effort, but is currently not detected in sequenced 1000 Genomes samples; and rs183437571 was poorly imputed with 1000 Genomes phase III, with a resulting signal that was not genome-wide significant (see **Table 1** and **Supplementary Table 7**).

Overall, the amount of variance explained by markers associated at the genome-wide level (**Table 1**) account for a fraction of the estimated genetic component of each trait (from 46 % for HbA1 to 68 % for HbA2, see **Methods**), supporting inheritance models that include small effect size and/or rare variants. For instance, 21 additional genes with suggestive significance signals ($p<1.x10^{-04}$, minor allele frequency [MAF] > 0.5 %) were related to genome-wide significant loci listed here, either in the scientific literature (Pubmed before 2006) or by expression levels (Human Expression Atlas[36]) or Gene Ontology[37] categories, using GRAIL

software[38] (see **Supplementary Note** and **Supplementary Table 8**). Four of the suggestive signals most strongly linked to genome-wide association findings were located in *NFE2,* which encodes Erythroid Nuclear Factor 2[39]; *ADGB,* which encodes a recently discovered globin of unknown physiological function[40]; and *SPTB* and *ANK1,* both of which encode proteins affecting the stability of erythrocyte membranes[41].

To test for replication of the associations at new loci detected in Sardinia, we used the largest independent sample reported to date, which measured HbA2 and HbF as well as F-cells (see **Methods**) in 4,131 individuals from the TwinsUK cohort enrolled from the United Kingdom (UK) general population[42]. For two loci, both associated with HbA2, we successfully replicated the association seen in Sardinia. In particular, we observed a p-value of 6.98×10^{-06} for rs59329875 in the *PLTP-PCIF1* intergenic region (MAF of 0.18) and a p-value of 1.73×10^{-04} for rs113267280 in *CCND3* (MAF of 0.01). The rarity of other variants precluded replication. The *MPHOSPH9* and *FOG1* variants associated with HbA1 and HbA2, respectively, are missing in publicly available imputation panels (as detailed above), and rs183437571 in *NFIX* associated with HbF was imputed as monomorphic in the TwinsUK cohort (see **Table 2** and **Methods**).

Fine mapping at known loci

The integration of whole-genome sequence variants in the scan was also instrumental to refine signals at previously known loci, either identifying a better lead variant or indicating novel independent signals. Specifically, as detailed below, we refined the association within the alpha and beta-globin gene clusters with all 3 hemoglobins; the association of the *HBS1L-MYB* intergenic region with HbA2 and HbF; and the association of the *BCL11A* gene with HbF.

Associations within the beta-globin gene cluster were intricate. As reported above, the strongest modifier in this region is the *HBB* beta(0)39 variant, acting on all 3 hemoglobin types (see **Figure 1**, **Methods** and **Supplementary Table 2**). Multiple additional independent signals were observed in conditional analyses for HbA2 and HbF, but they were distinct for each hemoglobin type, highlighting different regulatory patterns within the beta-globin gene cluster. Specifically, for HbA2, we confirmed 2 known independent associations at missense mutations in the *HBD* gene (rs35152987 and rs35406175, the latter perfectly tagged by our lead signal, see **Supplementary Table 2**). In addition, we identified 3 novel independent signals (rs12793110, rs11036338 and rs7936823) within a block of LD around the *HBB* gene, confirming a controlling role of this region in HbA2 production[5] (see **Figure 1** and **Supplementary Figure 4**). For HbF levels, 2 new independent signals were detected in a separate LD-block of the beta-globin gene cluster (see **Figure 1** and **Supplementary Figure 5**). The first, situated in an intron of the *HBE1* gene (rs67385638), remained associated even when taking into account 43 other variants in the beta-globin gene cluster associated with hemoglobin variation (see **Supplementary Note**). The second was located in a cyclic AMP response element upstream from *HBG2* (rs2855122) already implicated in drug-mediated HbF induction by butyrate[43] : different features of this marker make it a strong candidate for fetal to adult hemoglobin switching modulation (see **Supplementary Note**).

At the alpha-globin gene cluster, 2 variants were associated with HbA1 and 3 with HbA2, of which one affected both traits (**Table 1** and **Figure 1**). All results at this locus were corrected for any effect of the most frequent alpha-globin gene deletion present in Sardinia (NG_000006.1:g.34164_37967del3804, known as –a3.7 deletion type I), directly genotyped in a subset of the volunteers and imputed for the rest of the cohort

(see **Methods**). This deletion was associated at the genome-wide level with both HbA1 and HbA2 and only nominally with HbF (see **Table 1** and **Supplementary Table 2**). The most strongly associated signals (rs570013781 and rs141494605) were situated within the *NPRL3* and *HBM* genes, affecting HbA1 and HbA2 respectively. *NPRL3* contains several hypersensitive sites involved in the regulation of alpha-globin gene. *HBM* encodes a globin member of the avian alpha-D family[44] and its expression is highly regulated in human erythroid cells, although the protein has not been detected in human erythroid tissues. These observations suggest a possible regulatory function for which high-level protein expression is not required[44]. An independent variant associated with HbA1 and HbA2 (chr16:391593) was observed within the *AXIN1* gene, in which a further independent SNP (rs148706947) was found associated with HbA2 alone (**Supplementary Figure 3** and **Supplementary Figure 4**).

We also examined variants in the *HBS1L-MYB* intergenic region known to be associated with HbF and HbA2 levels[5]. We confirmed the role of the known variant (rs66650371, a TAC deletion) on the expression of both forms of hemoglobin[45,46] (see **Supplementary Note**). A further novel independent signal for HbF was found at rs11754265 in an intron of *HBS1L*, which has been shown to be a much stronger eQTL than rs66650371 for *HBS1L* and the neighboring *ALDH8A1* in monocytes[47].

In line with previous studies[6–8,48,49] the second intron of *BCL11A* gave multiple signals associated with HbF levels. They are explicable by the joint action of variants in each of two independent groups of statistically indistinguishable SNPs: one group formed by rs4671393, rs766432 and rs1427407, with p-values between $2.6x10^{-130}$ and $5.6x10^{-129}$, and the other by rs13019832 and rs7606173, with p-values of $6.1x10^{-33}$ and $9.1x10^{-33}$ in our cohort. The most likely causal candidate in the first group is rs1427407, a variant already associated with HbF in other population cohorts and functionally associated with *BCL11A* regulation[50]. In the second group we can instead point to rs13019832, which shows the highest functional CADD score (**Supplementary Table 5**). This variant has also been correlated, in adipose tissue, with the methylation of a CpG site (cg23678058) in a region that is functionally associated with *BCL11A* expression[51] and shows evidence of an effect on GATA-1 binding in peripheral blood-derived erythroblasts[52,53].

Pleiotropic effects

Among our 23 lead variants, 6 were associated (at least with p<0.01) with a second hemoglobin type, and another 6 were associated with all 3 (including beta(0)39 and –a3.7 deletion type I) (**Figure 1** and **Table 1**). Overall, all but 3 pleiotropic variants modulate different hemoglobins in the same manner, i.e., with the same allele increasing the levels of all associated hemoglobins. The 3 exceptions include the beta(0)39 variant, which decreases HbA1 while increasing HbA2 and HbF, and 2 SNPs mapping in the beta-globin gene cluster, both affecting HbA2 and HbF but in opposite directions (**Figure 1** and **Table 1**). In addition, many of the additional suggestive signals are associated with more than one hemoglobin type, increasing the likelihood that they are true signals (see **Methods**). In fact, 14 of these variants – all sharing effects on HbA1 and HbA2, but none with HbF – showed between-trait combined p-values that were genome-wide significant (**Supplementary Table 9**) and hint at additional pathways of potential interest in hemoglobin dynamics.

In general, the extended number of genetic variants showing joint association with HbA1 and HbA2 rather than HbF is consistent with high correlations of levels of adult hemoglobins HbA1 and HbA2 but only partial correlations of these hemoglobin forms with levels of HbF (see **Methods**).

Given the central role of hemoglobin in providing oxygen to the body tissues and the substantial fraction of total body cells accounted for by circulating red cells, factors impacting hemoglobin production and red cell count unsurprisingly have pleiotropic effects on other non-hematological traits. This is exemplified by the strong impact of the major beta(0)39 mutation on cholesterol and LDL-cholesterol. Here we extended the analysis for this mutation to 69 non-hematological quantitative traits selected from among those assessed in the SardiNIA cohort[4] (see **Supplementary Note**). We found the variant also significantly associated with increased total white blood cell counts ($p = 3x10^{-7}$) -- with the major contribution coming from neutrophil counts ($p = 1x10^{-6}$) -- and platelet counts ($p = 9x10^{-5}$) (see **Supplementary Table 10)**

DISCUSSION

We provide evidence for 23 associated variants at 10 loci influencing the levels of one or more of the 3 hemoglobin species measurable in post-natal life. Our results are based on a cohort from the Sardinian founder population that is much larger than previously described GWAS for HbF and HbA2 and interrogates a high resolution genetic map, based on population sequencing that expands the assessed spectrum of allelic variants 10-fold compared to previous studies. The finding that 2 of the 5 newly reported loci were not detectable without using the SardiNIA reference panel, and the others were misplaced (**Table 1** and **Supplementary Table 7**), further highlights how large-scale sequencing efforts in this founder population can reveal functionally relevant variants that may be very rare and hence missed in other populations.

For the same reasons, however, replication of results for such variants or translation of findings directly to other populations is difficult. For example, the other currently reported sample of comparable size, from the United Kingdom, could provide replication only for the two variants present there. Similar limitations will likely be found in other GWAS designed to detect effects of rare and founder variants. However, additional corroboration of our findings for such variants comes from their independent associations with other hemoglobin species and hematological traits in Sardinians, and also from the biological function of the genes involved. For instance, variant chr12:123681790 within *MPHOSP9*, associated with HbA1, also shows suggestive evidence of association with HbA2. The variant in *FOG1,* very rare in Europeans (MAF 0.4 %), is a missense variant in a gene implicated in erythropoiesis; and the variant in *NFIX*, absent in other European populations, falls within a cluster of genes involved in erythropoiesis and in a CpG region differentially methylated in fetal and adult red blood cell progenitors[25].

By carrying out GWAS for HbA1, HbA2 and HbF assessed for the first time in the same individuals, we see a wide range of pleiotropic effects of variants across the 3 hemoglobin types (**Table 1**). Strikingly, HbA2 harbors more than half of the loci discovered here (see

Figure 2), with many pleiotropic effects on HbA1 and some on HbF. Thus, although it has a minor role in the transport of oxygen to tissues[54], variations in HbA2 participate in pathways that regulate the levels of the other hemoglobins active in postnatal life.

The direction of pleiotropic effects among the different hemoglobin types provides some additional clues to mechanism. Within the alpha-globin gene cluster, in agreement with the presence of alpha-globin chains in HbA1, HbA2 and HbF, all variants affecting more than one hemoglobin showed the same direction of effect for all. The regulation of globin chains from the beta-globin gene cluster, however, is more complicated. It involves variants with the same direction of effect for all hemoglobins (rs7936823) and other variants most likely involved in switching mechanisms that affect fetal and adult hemoglobins in opposite directions (rs2855122). Still other variants change the kinetics of competition among non-alpha globin chains; for example, the beta(0)39 mutation decreases beta-globin levels and thereby increases the availability of alpha-globin chains to combine with delta and gamma-globins, leading to higher levels of HbA2 and HbF.

Variants influencing only 2 forms of hemoglobin acted mainly in the same direction and never jointly affected HbA1 and HbF. As for variants shared only between HbA2 and HbF, they can be attributed to specific cis-regulatory mechanisms in the beta-globin gene cluster (rs12793110 and rs7944544) or to loci with a role in erythroid differentiation (*CCND3* and *MYB*). By contrast, variants shared between HbA2 and HbA1 were either trans-acting (in *MPHOSPH9*) or localized in the alpha-globin gene cluster but with effect sizes probably too small to impact HbF production. Consistent with the latter possibility, the –a3.7 deletion type I, which has strong genome-wide significant effects on HbA1 and HbA2, had much smaller, only suggestive, effects on HbF (see **Supplementary Table 2**).

Our analyses also detected broader pleiotropic impacts, most strikingly for the beta(0)39 variant. In addition to effects on LDL-c described in the **companion paper**, we report for the first time that beta(0)39 is also significantly associated with increased total counts of white blood cells (and some subsets) as well as platelet counts. This suggests that in heterozygous carriers this variant drives a broader increase in bone marrow-derived blood cells. Speculatively, some of these, such as augmented leukocyte and neutrophil counts, may have provided protection against pathogens other than malaria, thus increasing selection for the balanced polymorphism.

The detected variants provide candidate modifiers influencing the clinical status of patients with monogenic hemoglobin disorders. For example, we carried out a preliminary analysis of a small sample of 306 beta-thalassemia patients homozygous for the beta(0)39 stop codon mutation but showing very great heterogeneity in disease presentation and course. In addition to those described previously[7–10], some variants detected in this study showed possible effects as modifiers of disease severity (see **Supplementary Note**). However, the potential of these variants to help predict disease severity remains tentative without studies of larger sample sets. Nevertheless, the variants already add to the candidate targets for therapeutic intervention in the widely prevalent inherited beta-thalassemia and other hemoglobinopathies[2].

FIGURES AND TABLES

Figure 1. Association at the globin clusters.
Schematic representation of association results in the genomic context of the beta-globin (panel a) and alpha-globin (panel b) gene clusters. For each hemoglobin, the markers associated are positioned with + or – corresponding to an increase or decrease in the corresponding trait by the effective allele (as in **Table 1**). Symbol is larger if the marker is associated at genome-wide level or smaller if it results from the analysis of pleiotropic effects. The beta(0)39 mutation and –a3.7 type I deletion as well as relevant genes and the locus control region hypersensitivity sites (HS) are indicated. Finally, at the bottom of each panel is represented the linkage disequilibrium (r^2) profile for the region in Sardinia, with colors ranging from high (red), to intermediate (green), and low (blue).

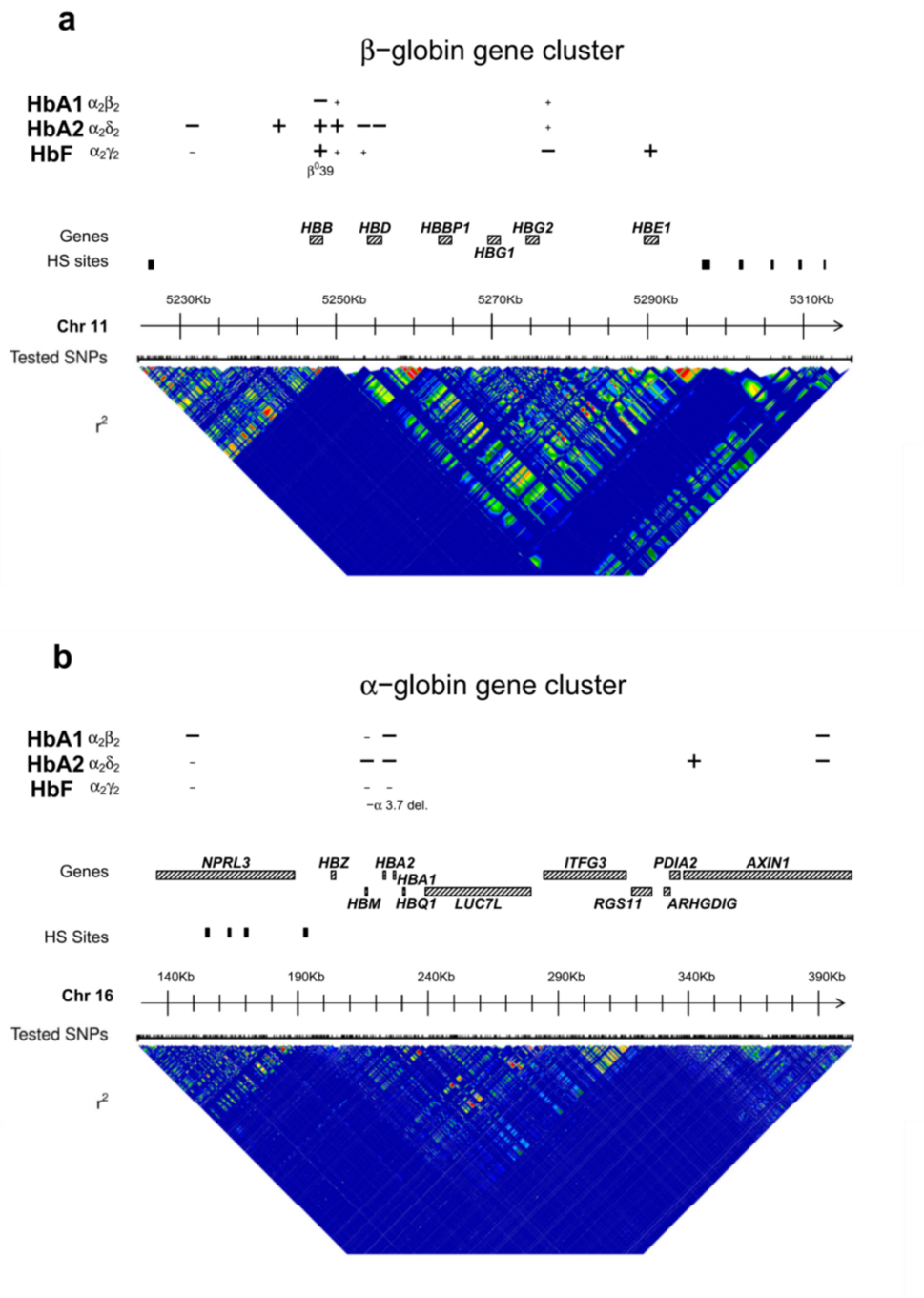


Figure 2. Diagram of genome-wide associated loci.

Representation of genome-wide significant findings on hemoglobin levels in relation to their contribution to the phenotypic variation (variance explained, panel a) or to their individual impact (effect size, panel b). At each step, the length of the black bar represents the magnitude of variance explained (panel a) or effect size (panel b) for each trait, locus, gene and variant. The bars are connected by colored bands to their sub-components (loci for each trait, genes for each locus, variants for each gene). Three colors (yellow, green and blue) represent the 3 hemoglobin forms (HbA1, HbA2 and HbF respectively), and for loci or genes affecting more than one hemoglobin: gray combines HbA1 and HbA2, cyan combines HbA2 and HbF, and light gray represents effects common to all 3 hemoglobin forms. Each panel is drawn to show loci in order of their importance, i.e. from the largest to smallest amount of explained phenotypic variance (panel a) or effect size (panel b). The variance explained by each locus was calculated fitting a regression model including all variants at that locus, while the effect size for a locus is the sum of effect sizes of all variants in that locus (**Supplementary Table 3** reports effect sizes for such joint models). For variants associated with more than one trait the maximum value is used. Markers are reported as chromosome : position when an rs ID was not available; and when an intergenic region is involved instead of a single gene, we show nearby genes within brackets.

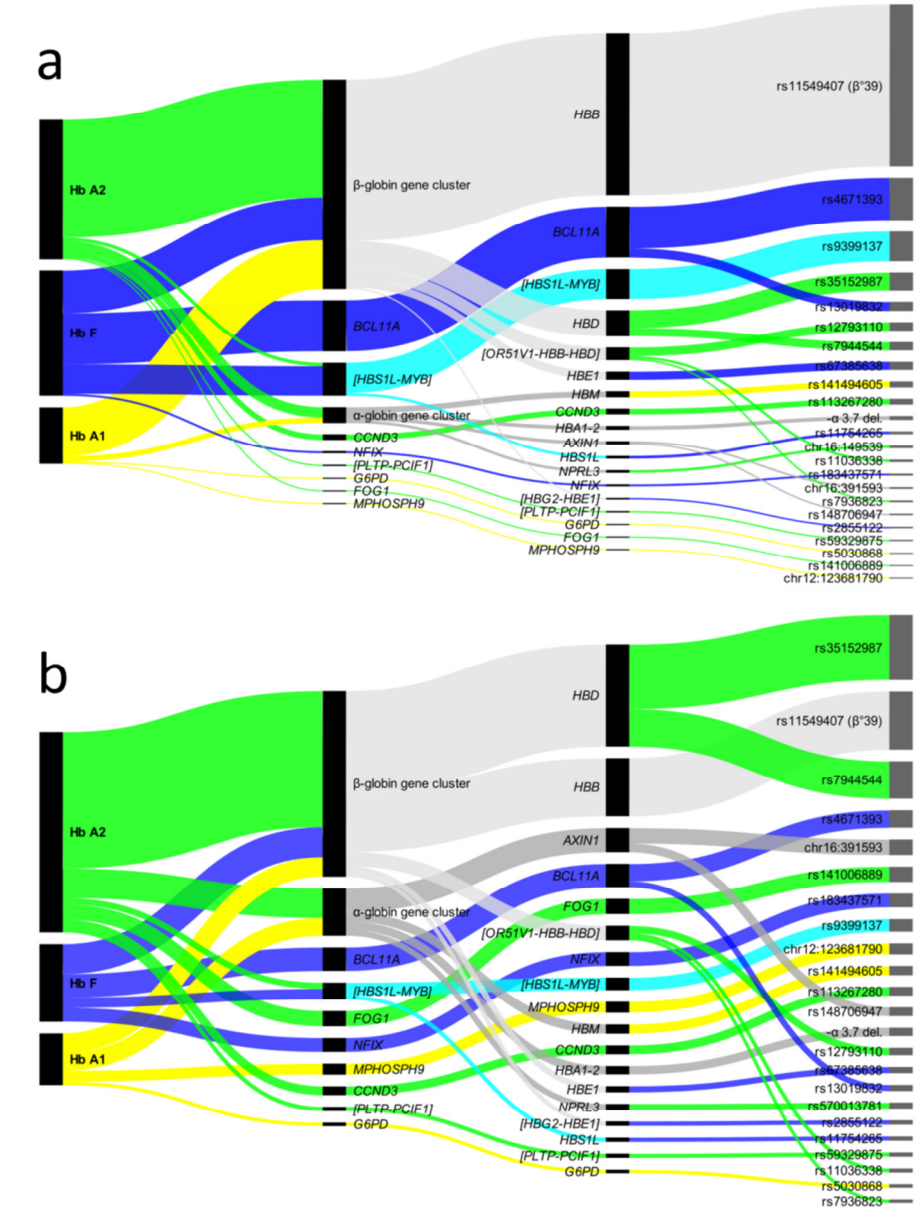


Table 1. Most significant independent association results from single variant tests for hemoglobin A1, A2 and fetal.

The table shows the most significant association results (all results are corrected for beta^0 mutations observed in the *HBB* gene, and results on the alpha-globin gene cluster are adjusted for the -a3.7 deletion type 1, see **Methods**). Novel signals are shown in bold while variants refining previously reported signals are in italic. At each locus, we indicated the chromosome and genomic position (hg19 build), the rs ID when available, the effect allele tested for association (EA) and the other allele at the SNP (OA), the imputation accuracy (RSQR), the SNP effect allele frequency (EAF) and the regression coefficients. We then indicated whether the SNP is also linked the other hemoglobin forms ($p < 0.01$), and indicated the direction of the effect allele (+ for increasing the levels of Hb, - for decreasing). The candidate genes likely to be modulated by the lead SNP are also reported along with their inclusion criteria, as described in **Methods** (p = position, c = coding, e = eQTL, o = OMIM, b = biological). Where "Alpha-globin gene cluster" is mentioned we refer to *NPRL3*, *HBZ*, *HBQ1*, *HBA1*, *HBA2* and *HBM* genes; while for "Beta-globin gene cluster" we refer to *HBB*, *HBD*, *HBBP1*, *HBG1*, *HBG2* and *HBE1* genes. Association coefficients for males and females are reported in **Supplementary Table 11**.

									Shared effects		
Traits (units) and loci #	Candidate genes	chr:position	rsID from dbsnp142	Alleles (EA/OA)	RSQR	EAF	Effect (StdErr)	p-value	HbA1	HbA2	HbF
HbA1 (g/dl)											
locus1 [1]	Alpha-globin gene cluster(p,o,b); *MPG*(p)	*16:149539* [1,4]	rs570013781	*A/G*	*0.98*	*0.136*	*-0.1995 (0.023)*	*5.86x10^{-18}*	-	-	-
	Alpha-globin gene cluster (p,o,b); *AXIN1*(p)	**16:391593** [1,3,5] **(cond.)**	-	**T/C**	**0.94**	**0.012**	**-0.4028 (0.058)**	**3.28x10^{-12}**	**-**	**-**	
locus2	*FAM3A*(p); *G6PD*(p,c,o,b); *IKBKG*(p)	*X:153762634* [4]	*rs5030868*	*A/G*	*Genotyped*	*0.085*	*-0.1256 (0.019)*	*2.78x10^{-11}*	-		
locus3 [2]	*MPHOSPH9*(p)	**12:123681790** [2]	-	**A/C**	**0.96**	**0.010**	**-0.3606 (0.064)**	**1.68x10^{-08}**	**-**	**-**	
HbA2											
locus1 [4] *(%)*	Beta-globin gene cluster(p,o,b); *HBD*(c)	*11:5255582* [4]	*rs35152987*	*A/C*	*Genotyped*	*0.004*	*-2.182 (0.109)*	*4.35x10^{-86}*		-	
	Beta-globin gene cluster (p,o,b); *HBD*(c)	*11:5251849* [4] *(cond.)*	*rs7944544*	*T/G*	*0.98*	*0.005*	*-1.26 (0.097)*	*3.90x10^{-38}*		-	*+*
	Beta-globin gene cluster (p,o,b); *HBB*(c); *HBG1/HBG2*(e); *OR51V1*(p)	*11:5231565* [4] *(cond.)*	*rs12793110*	*T/C*	*1.00*	*0.181*	*-0.2408 (0.019)*	*5.75x10^{-36}*		-	-
	Beta-globin gene cluster (p,o,b); *OR51V1*(p)	*11:5242698* [4] *(cond.)*	*rs11036338*	*C/G*	*0.99*	*0.381*	*0.1282 (0.017)*	*2.03x10^{-14}*		*+*	

	Beta-globin gene cluster (p,o,b); *HBG1/HBG2(e)*	*11:5250168 [4] (cond.)*	*rs7936823*	*G/A*	*0.96*	*0.466*	*0.1117 (0.015)*	*5.00x10^{-13}*	*+*	*+*	*+*
locus2 [1,3,5] (g/dl)	Alpha-globin gene cluster (p,o,b); *HBM* (c); *LUC7L(p)*	**16:216593 [1,3]**	**rs141494605**	**C/T**	**0.97**	**0.149**	**-0.3080 (0.025)**	**3.94x10^{-35}**	**-**	**-**	**-**
	Alpha-globin gene cluster (p,o,b); *AXIN1*(p)	**16:391593 [1,3,5] (cond.)**	**-**	**T/C**	**0.94**	**0.012**	**-0.5112 (0.063)**	**6.48x10^{-16}**	**-**	**-**	
	Alpha-globin gene cluster (p,o,b); *ARHGDIG*(p); *AXIN1*(p); *ITFG3*(p); *PDIA2*(p); *RGS11*(p)	**16:342218 [1,3,5] (cond.)**	**rs148706947**	**T/C**	**0.93**	**0.021**	**0.2892 (0.051)**	**1.04x10^{-08}**		**+**	
locus3 [2] (%)	*CCND3*(p,b)	**6:41952511 [2]**	**rs113267280**	**G/T**	**0.99**	**0.101**	**0.2923 (0.026)**	**1.11x10^{-29}**		**+**	**+**
locus4 (%)	*MYB*(b)	6:135418916	rs7776054	G/A	Genotyped	0.210	0.1762 (0.020)	3.71x10^{-19}		+	+
locus5 [2] (%)	*CTSA*(p); *PCIF1*(p,c); *PLTP*(p,e); *MMP9*(e); *TNNC2*(e)	**20:44547672 [2]**	**rs59329875**	**C/T**	**1.00**	**0.134**	**-0.1399 (0.024)**	**3.64x10^{-09}**		**-**	
locus6 [2] (%)	*FOG1*(p,b,c); *C16orf85*(p)	**16:88601281 [2]**	**rs141006889**	**G/A**	**Genotyped**	**0.007**	**-0.5074 (0.087)**	**5.33x10^{-09}**		**-**	
HbF (g/dl)											
locus1	*BCL11A*(p,o,b)	2:60720951	rs4671393	A/G	1.00	0.136	0.578 (0.023)	2.60x10^{-130}			+
	BCL11A(p,o,b)	*2:60710571 [4] (cond.)*	*rs13019832*	*A/G*	*1.00*	*0.484*	*-0.2024 (0.017)*	*9.12x10^{-33}*			-
locus2	*MYB*(b)	6:135419018	rs9399137	C/T	Genotyped	0.205	0.4202 (0.020)	1.09x10^{-93}		+	+
	HBS1L(p,c,e); *ALDH8A1*(e)	**6:135356216 [3] (cond.)**	**rs11754265**	**C/G**	**1.00**	**0.367**	**-0.1421 (0.021)**	**5.04x10^{-12}**			**-**
locus3 [4]	Beta-globin gene cluster (p,o,b); *HBG1/HBG2*(e)	*11:5290370 [4]*	*rs67385638*	*G/C*	*1.00*	*0.236*	*0.2038 (0.019)*	*1.09x10^{-25}*			*+*
	Beta-globin gene cluster (p,o,b); *HBG1/HBG2*(e)	*11:5277236 [4] (cond.)*	*rs2855122*	*C/T*	*1.00*	*0.395*	*-0.1458 (0.022)*	*2.57x10^{-11}*	*+*	*+*	-
locus4 [2,5]	*NFIX*(p)	**19:13121899 [2,5]**	**rs183437571**	**T/C**	**0.97**	**0.010**	**0.4607 (0.081)**	**1.61x10^{-08}**			**+**

[1] = association results locally corrected for the –a3.7 deletion type I (NG_000006.1:g.34164_37967del3804) (see **Supplementary Note**); [2] = first time associated to the trait and in a novel locus; [3] = first time associated to the trait in a previously reported locus; [4] = signal refining a previously reported signal; [5] = result not found using the 1000 Genomes reference panel ; cond. = obtained by conditional analysis on variants reported on the upper rows for the considered locus.

Table 2. Replication of novel loci.

The table describes association in the TwinsUK cohort (N = 4,131 individuals). For each SNP, we indicated the associated hemoglobin tested, the number of samples analysed, the imputation accuracy according to the IMPUTE-INFO metric, the effect allele tested for association (EA) and the other allele at the SNP (OA), the SNP effect allele frequency (EAF) and the regression coefficients. The last column explains the reason for the SNPs not being tested.

Traits (units) and loci # from Table 1	*SNP*	*Candidate genes*	*INFO score*	*Alleles (EA/OA)*	*EAF*	*Effect (StdErr)*	*p-value*	*Notes*
HbA1 (g/dl)								
locus3	*chr12:123681790*	*MPHOSP9*	-	-	-	-	-	Not imputable because absent in 1000 Genomes; at the moment, Sardinian specific.
HbA2 (%)								
locus3	*rs113267280*	*CCND3*	0.843	G/T	0.011	0.442 (0.118)	$1.73x10^{-04}$	.
locus5	*rs59329875*	*PLPT-PCIF1*	0.994	C/T	0.185	0.132 (0.029)	$6.98x10^{-06}$	
locus6	*rs141006889*	*FOG1*	-	-	-	-	-	Not imputable because absent in 1000 Genomes; detected in the NHLBI GO Exome Sequencing Project (ESP).
HbF (%)								
locus4	*rs183437571*	*NFIX*	0.294	T/C	0.000	-	-	Imputed as monomorphic in TwinsUK cohort.

METHODS

Sample description

The population studied here includes 6,921 individuals, representing > 60 % of the adult population of 4 villages in the Lanusei Valley in Sardinia, Italy. They are part of the SardiNIA project, a longitudinal study including genetic and phenotypic data of 1,257 multigenerational families with more than 37,000 relative pairs. Details of phenotype assessments for these samples have been published previously[4]. All participants gave informed consent to study protocols, which were approved by the institutional review board of the University of Cagliari, the National Institute on Aging, and the University of Michigan.

For whole-genome sequencing, we selected 1,122 individuals from the SardiNIA study and 998 individuals enrolled in case–control studies of Multiple sclerosis and Type I Diabetes in Sardinia. Genomes were sequenced to an average coverage of 4.16-fold. Details on sequencing protocol, data process and variant calling can be found elsewhere[55] and in the **companion paper**. The 2,120 sequenced samples consist of 695 complete and incomplete trios; to avoid over-representation of rare haplotypes during imputation process we considered only parents for each trio – totaling 1,488 samples – to build our reference panel[55] (see **companion paper** for details).

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 and Imputation

The 4 micro-arrays used for genotyping the entire SardiNIA cohort were the Illumina® Infinium HumanExome BeadChip, ImmunoChip, Cardio-MetaboChip and HumanOmniExpress BeadChip. Genotyping was carried out according to manufacturer protocols at the SardiNIA Project Laboratory (Lanusei, Italy), at the Technological Center - Porto Conte Ricerche (Alghero, Italy) and at the National Institute on Aging Intramural Research Program Laboratory of Genetics (Baltimore, MD). Genotypes were called using GenomeStudio (version 1.9.4) and refined using Zcall (version 3)[56]. We applied standard per sample quality control filters to remove samples with low call rates or for which reported relationships and/or gender disagreed with genetic data. Details on quality controls were described elsewhere[55]. Altogether, 890,542 autosomal markers and 16,325 X-linked markers were genotyped across SardiNIA study samples. We selected for phasing and imputation only the 6,602 samples for which all 4 arrays were successfully genotyped.

Genotypes were phased using MACH software[57], using 30 iterations of the haplotyping Markov chain and 400 states per iteration. We performed imputation using Minimac software[58] and a reference panel including haplotypes of 1,488 Sardinian whole-genomes[55] (see **companion paper**). Variants with estimated imputation quality (RSQR) <= 0.3 or <0.8 were discarded if the estimated MAF was >= 1 % or between 0.5 % and 1 % respectively; variants with MAF < 0.5 % were kept only if genotyped. RSQR thresholds for rare and low frequency variants were more stringent than those proposed for other traits[55] as they led to better genomic control parameters (1.001, 0.993 and 0.985 for HbA1, A2 and fetal, respectively). We also performed imputation

using the 1000 Genomes Project Phase III (version 5)[59] haplotype set, and used the same thresholds to discard variants. Genomic control parameters for 1000 Genomes imputation were 1.050, 0.997 and 0.984 for HbA1, A2 and fetal, respectively.

Association analysis

We performed association analyses of all 3 hemoglobins in grams per deciliter (g/dl) as well as percentage (%) for HbA2 and HbF. HbA2 (%) and HbF (%) were directly measured from high-performance liquid chromatography, while HbA1 (g/dl), HbA2 (g/dl) and HbF (g/dl) were derived from total hemoglobin measured by Coulter counter. As expected, measurements in % and g/dl were highly correlated for HbF (Spearman's Rho = 0.99) and for HbA2 (Rho = 0.85). HbA1 (%) was not considered for genetic association because it was too highly correlated with both HbA2 (%) and HbF (%) as a consequence of their derivation formula (Rho = -0.803 and -0.757, respectively, $p < 1x10^{-20}$). Considering only non-carriers of beta^0-mutations, HbA1 (g/dl) was highly correlated with HbA2 (g/dl) (Rho = 0.662, $p < 1x10^{-20}$) and poorly with HbF (g/dl) (Rho = -0.055, $p = 3.44x10^{-5}$). Likewise, HbA2 and HbF were weakly positively correlated as percentage measures (Rho = 0.108, $p = 4.08x10^{-16}$) and even less as g/dl (Rho = 0.066, $p = 5.81x10^{-5}$), consistent with previous findings[5]. Measurements were available for a subset of 6,305 individuals; descriptive statistics are reported in **Supplementary Table 1**. Association results were considered genome-wide significant when p-value was less than $5x10^{-08}$, however we also noted in the text variants that would not meet a threshold of $1.4x10^{-8}$ we introduce for sequencing based GWAS carried out in Sardinians for variants with MAF > 0.5 % (see **companion paper**).

Before association analyses, traits were normalized using inverse normal transformation; for HbF we also removed outliers with values above 5 %. Analyses were adjusted for age, age^2, and gender as well as for the presence of at least one of the 3 beta^0 mutations (beta(0)39 (rs11549407), HBB:c.20delA (rs63749819) and HBB:c.315+1G>A (rs33945777)), all directly genotyped or sequenced (see Characterization of beta0 mutations paragraph). Regression coefficients for beta(0)39 – the most common in Sardinia with 10.3 % of carriers – are reported in the **Supplementary Table 2**.

Association was performed using the q.emmax test in EPACTS[60], which implements a linear mixed model procedure to correct for cryptic relatedness and population stratification by incorporating a genomic-based kinship matrix. Associations reported in the table refer to the best p-value obtained with either percentage or original units for HbA2 and HbF. Notably, HbF signals always resulted in lower p-values considering g/dl, whereas for HbA2 analysis, this was only the case for rs141494605. All loci passed the genome-wide significance threshold of $p<5x10^{-08}$ for both % and g/dl except for rs59329875, which was genome-wide significant only for the HbA2 measure reported in **Table 1**.

To identify independent signals we performed regional conditional analysis, using forward selection procedure adding, at each step, the most associated variant as covariate in the model. In this sequential analysis, we tested only SNPs lying in a region of 2Mb centered on the lead variant. The same genome-wide significance threshold used for primary signals was also considered for independent signals. For loci where different independent signals were found, we

also report model parameters of jointly associated variants in **Supplementary Table 3.** Finally, the lead variants and their surrogates ($r^2 > 0.90$) were annotated using Combined Annotation Dependent Depletion (CADD) score[13] and reported in **Supplementary Table 5.**

Heritability and variance explained

We estimated heritability for the 3 hemoglobins using Merlin-regress[61] on the same sample used for the GWAS study. Estimates for normalized levels of hemoglobins were respectively 0.520 for HbA1 (g/dl), 0.728 for HbA2 (%) (0.700 for g/dl) and 0.633 for HbF (%) (0.624 for g/dl). We then calculated for each hemoglobin form the proportion of phenotypic variance explained by the associated lead variants. We measured that as the difference of R^2-adjusted observed between the full and the basic model, where the basic model includes only phenotypic covariates (age, age^2 and gender) and the full model also includes all the independent SNPs associated with the specific trait. R^2-adjusted values were calculated using a linear mixed model procedure from lmekin() function in the "Kinship" R package[62]. Estimates were 0.240 for HbA1 (g/dl), 0.492 for HbA2 (%) and 0.383 for HbF (%).

Characterization of beta^0 mutations

For the present study we designed a Taqman custom assay for the HBB:c.118C>T nonsense mutation (rs11549407, also known as beta(0)39), and genotyped 6,602 samples. Comparison of Taqman genotypes and imputation results (rs11549407, RSQR = 0.92) produced an overall concordance of 98.8 %. Also, we further sequenced all samples discordant between red blood cell index-based diagnosis (using MCV, MCH, HbF % and HbA2 %) and Taqman genotypes, using Sanger sequencing to determine any additional beta-globin mutations different from beta(0)39, thus identifying 3 carriers for the HBB:c.20delA (rs63749819) and one for the HBB:c.315+1G>A (rs33945777) mutations.

Characterization of the deletion at the alpha-globin gene cluster

In Sardinia 3 variants are known to be mainly responsible for alpha-thalassemia: SNPs rs111033603 and rs41474145, and the deletion NG_000006.1:g.34164_37967del3804; the latter, known as the –a3.7 deletion type I, is by far the most common[63]. We did not observe the rarer rs111033603 or rs41474145 in our sequencing effort. To establish genotypes at the deletion site in the full cohort, we used an inference strategy combined with experimental data. Specifically, we first characterized the structural variant by PCR in 260 unrelated sequenced individuals randomly selected in the SardiNIA cohort. We calculated the relative coverage of the deleted region in the whole-genome sequenced samples by considering the ratio of read count in the potentially deleted region (223,450 to 226,953 bp – excluding 150 bp boundaries) with read count in the nearby region not subject to deletion (227,254 to 230,757 bp). We then identified coverage ratio thresholds that best predicted PCR genotypes at the deletion and used these thresholds to infer genotypes for the 2,120 sequenced individuals. We then inserted genotypes in

the Sardinian reference panel and imputed the deletion on the total SardiNIA cohort. To assess accuracy of imputation we considered the best guess genotypes and searched for Mendelian errors in families. The observed rate was 0.58 % over 1,193 parent-offspring pairs, consistent with high imputation precision. Association results reported in the manuscript at this locus are corrected for the inferred –a3.7 deletion type I dosages.

Variants validation

We validated all variants that showed genome-wide significant p-values in the primary or conditional analysis that were not directly genotyped or had no surrogates ($r^2 > 0.90$) that were directly genotyped. We did not validate variant rs13019832 at *BCL11A* for HbF, which was highly linked with findings of previous reports (rs7606173)[48,50]. Validation was performed using Sanger sequencing or Taqman, depending on variant frequency, for 5 variants. We selected for each variant all individuals carrying the minor allele (heterozygous and homozygous) plus a random subset of subjects homozygous for the other allele (in all, 3,084 subjects were genotyped), except for rs141494605 and chr16:391593, for which we specifically selected worse imputation dosages (borderline RSQR). In addition, for rs17525396, we used independent genotypes available for a subset of the cohort[64], derived from Affymetrix 6.0 (see **Supplementary Table 4**).

Replication of variant effects

Replication was performed in the TwinsUK cohort[42]. Genotyping was performed using a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M), and imputation performed using the IMPUTE software package (v2) and 1,000 Genomes haplotypes released on 16 Jun 2014-- Phase I integrated variant set release[35,65]. Details on quality controls are provided as **Supplementary Note**. HbA2 levels and HbF percentage were obtained by HPLC, and F-cells were enumerated after intracellular HbF staining and subsequent flow cytometry[66]. Measurements were available in 4,131 samples. Association analyses were performed with merlin-offline package in Merlin, to account for relatedness[61]. To be consistent with analyses performed in the SardiNIA study, age, age squared and gender were used as covariates and the traits transformed using quantile normalization.

Selection of candidate genes

At each locus, we defined a list of genes to be considered as plausible candidates if they satisfied one of the following: 1) genes that were +/- 25Kb of the lead SNP, indicated (p) in **Table 1**; 2) genes with exonic variants (frame-shift, stop-codon, non-synonymous and synonymous) along with splice-site and 5'/3' UTR variants in LD (r2≥0.8) with the lead SNP (c); 3) genes whose expression was modulated by the SNP itself or by an eQTL in LD ($r^2 \geq 0.8$) with the top SNP (e); 4) genes with clear biological function connected to the traits (b); or 5) genes harboring variants responsible for which Mendelian diseases, as reported in OMIM (o). Candidate genes from eQTL data were searched using an automatized pipeline querying 16 eQTL public repositories[47,67–81], including the Pritchard eQTL browser; only top SNP eQTLs or any SNP with FDR < 0.05 were considered.

Pleiotropy and gene connections analysis

To characterize genome-wide significant results and to identify suggestively significant ones, we searched for effects shared between the different hemoglobin forms as well as evidence of connections between both. Specifically, for genome-wide significant markers, we simply reported the effect direction for all traits with $p < 0.01$ when a marker is associated at genome-wide level for one trait (see **Table 1**). To identify candidates with suggestive p-values between 1.00×10^{-04} and 5.00×10^{-08}, we selected among these:

- markers with MAF > 0.5 % and showing 2-trait combined p-values < 5×10^{-08}; p-values were combined using inverse variance weighted meta-analysis, as implemented in Metal software[82];
- markers falling in or nearby genes that demonstrated evidence of connections with genome-wide significant loci, either in Pubmed (using the 2006 data set to avoid confounding by subsequent GWAS discoveries), or in Human Expression Atlas[36] and Gene Ontology[37] databases using GRAIL[38] and considering genes reported with multiple hypothesis corrected p-values < 0.05.

Using these criteria, we identified 21 further genes with biological connections to genome-wide significant loci reported in **Supplementary Table 8** and 14 variants with combined p-values between 2.08×10^{-08} and 1.18×10^{-11}, reported in **Supplementary Table 9**.

URLs

SardiNIA project: https://sardinia.irp.nia.nih.gov
1000 Genomes project: http://www.1000genomes.org
HumanExome BeadChip design: http://genome.sph.umich.edu/wiki/Exome_Chip_Design
ImmunoChip, Cardio-MetaboChip and HumanOmniExpress BeadChip: http://www.illumina.com
GenomeStudio software: http://www.illumina.com/applications/microarrays/microarray-software/genomestudio.html
MACH software: http://csg.sph.umich.edu/abecasis/MACH
Minimac software: http://genome.sph.umich.edu/wiki/Minimac
Zcall software: https://github.com/jigold/zCall
IMPUTE v2 software: http://mathgen.stats.ox.ac.uk/impute/impute_v2.1.0.html
Merlin (including Merlin-regress and Merlin-offline): http://csg.sph.umich.edu/abecasis/merlin
Epacts software: http://genome.sph.umich.edu/wiki/EPACTS
Metal software: http://csg.sph.umich.edu/abecasis/metal
GWAS Catalog: http://www.genome.gov/gwastudies
Grail software: https://www.broadinstitute.org/mpg/grail
Gene Ontology: http://geneontology.org
Human Expression Atlas: http://symatlas.gnf.org
Pritchard eQTL browser: http://eqtl.uchicago.edu

REFERENCES

1. Sankaran, V. G., Xu, J. & Orkin, S. H. Advances in the understanding of haemoglobin switching. *Br. J. Haematol.* **149,** 181–194 (2010).
2. Modell, B. & Darlison, M. Global epidemiology of haemoglobin disorders and derived service indicators. *Bull. World Health Organ.* **86,** 480–487 (2008).
3. Malaria Genomic Epidemiology Network & Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. *Nat. Genet.* **46,** 1197–1204 (2014).
4. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2,** e132 (2006).
5. Menzel, S., Garner, C., Rooks, H., Spector, T. D. & Thein, S. L. HbA2 levels in normal adults are influenced by two distinct genetic mechanisms. *Br. J. Haematol.* **160,** 101–105 (2013).
6. Bae, H. T. *et al.* Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood* **120,** 1961–1962 (2012).
7. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 1620–1625 %U http://www.ncbi.nlm.nih.gov/pubmed/18245381 (2008).
8. Lettre, G. *et al.* DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 11869–11874 (2008).
9. Danjou, F. *et al.* Genetic modifiers of β-thalassemia and clinical severity as assessed by age at first transfusion. *Haematologica* **97,** 989–993 (2012).
10. Danjou, F. *et al.* A genetic score for the prediction of beta-thalassemia severity. *Haematologica* haematol.2014.113886 (2014). doi:10.3324/haematol.2014.113886
11. Van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–375 (2012).
12. Trecartin, R. F. *et al.* beta zero thalassemia in Sardinia is caused by a nonsense mutation. *J. Clin. Invest.* **68,** 1012–1017 (1981).
13. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
14. Freson, K. *et al.* Molecular cloning and characterization of the GATA1 cofactor human FOG1 and assessment of its binding to GATA1 proteins carrying D218 substitutions. *Hum. Genet.* **112,** 42–49 (2003).
15. Nichols, K. E. *et al.* Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. *Nat. Genet.* **24,** 266–270 (2000).
16. Kozar, K. *et al.* Mouse development and cell proliferation in the absence of D-cyclins. *Cell* **118,** 477–491 (2004).
17. Sankaran, V. G. *et al.* Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. *Genes Dev.* **26,** 2075–2087 (2012).
18. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).
19. Kamatani, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* **42,** 210–215 (2010).
20. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat.*

Genet. **41,** 56–65 (2009).
21. Jarvik, G. P. *et al.* Genetic and nongenetic sources of variation in phospholipid transfer protein activity. *J. Lipid Res.* **51,** 983–990 (2010).
22. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7,** e1001300 (2011).
23. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44,** 269–276 (2012).
24. Hirose, Y. *et al.* Human phosphorylated CTD-interacting protein, PCIF1, negatively modulates gene expression by RNA polymerase II. *Biochem. Biophys. Res. Commun.* **369,** 449–455 (2008).
25. Lessard, S., Beaudoin, M., Benkirane, K. & Lettre, G. Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. *Genome Med.* **7,** 1 (2015).
26. Riddell, J. *et al.* Reprogramming Committed Murine Blood Cells to Induced Hematopoietic Stem Cells with Defined Factors. *Cell* **157,** 549–564 (2014).
27. Holmfeldt, P. *et al.* Nfix is a novel regulator of murine hematopoietic stem and progenitor cell survival. *Blood* **122,** 2987–2996 (2013).
28. Kawane, K. *et al.* Requirement of DNase II for definitive erythropoiesis in the mouse fetal liver. *Science* **292,** 1546–1549 (2001).
29. Porcu, S. *et al.* Klf1 affects DNase II-alpha expression in the central macrophage of a fetal liver erythroblastic island: a non-cell-autonomous role in definitive erythropoiesis. *Mol. Cell. Biol.* **31,** 4144–4154 (2011).
30. Zhou, D., Liu, K., Sun, C.-W., Pawlik, K. M. & Townes, T. M. KLF1 regulates BCL11A expression and [gamma]- to [beta]-globin gene switching. *Nat Genet* **42,** 742–744 (2010).
31. Siatecka, M. & Bieker, J. J. The multifunctional role of EKLF/KLF1 during erythropoiesis. *Blood* **118,** 2044–2054 (2011).
32. Satta, S. *et al.* Compound heterozygosity for KLF1 mutations associated with remarkable increase of fetal hemoglobin and red cell protoporphyrin. *Haematologica* **96,** 767–770 (2011).
33. Borg, J. *et al.* Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat. Genet.* **42,** 801–805 (2010).
34. Perseu, L. *et al.* KLF1 gene mutations cause borderline HbA2. *Blood* **118,** 4454–4458 (2011).
35. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
36. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–6067 (2004).
37. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
38. Raychaudhuri, S. *et al.* Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLoS Genet* **5,** e1000534 (2009).
39. Andrews, N. C. The NF-E2 transcription factor. *Int. J. Biochem. Cell Biol.* **30,** 429–432 (1998).
40. Hoogewijs, D. *et al.* Androglobin: a chimeric globin in metazoans that is preferentially expressed in Mammalian testes. *Mol. Biol. Evol.* **29,** 1105–1114 (2012).
41. Iolascon, A., Perrotta, S. & Stewart, G. W. Red blood cell membrane defects. *Rev. Clin. Exp. Hematol.* **7,** 22–56 (2003).
42. Moayyeri, A., Hammond, C. J., Valdes, A. M. & Spector, T. D. Cohort Profile: TwinsUK and

Healthy Ageing Twin Study. *Int. J. Epidemiol.* **42,** 76–85 (2013).
43. Sangerman, J. *et al.* Mechanism for fetal hemoglobin induction by histone deacetylase inhibitors involves gamma-globin activation by CREB1 and ATF-2. *Blood* **108,** 3590–3599 (2006).
44. Goh, S.-H. *et al.* A newly discovered human alpha-globin gene. *Blood* **106,** 1466–1472 (2005).
45. Farrell, J. J. *et al.* A 3-bp deletion in the HBS1L-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* **117,** 4935–4945 (2011).
46. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124,** 1699–1710 (2014).
47. Zeller, T. *et al.* Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. *PLoS ONE* **5,** e10693 (2010).
48. Bhatnagar, P. *et al.* Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J. Hum. Genet.* **56,** 316–323 (2011).
49. Bauer, D. E. & Orkin, S. H. Update on fetal hemoglobin gene regulation in hemoglobinopathies. *Curr. Opin. Pediatr.* (2010). doi:10.1097/MOP.0b013e3283420fd0
50. Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342,** 253–257 (2013).
51. Grundberg, E. *et al.* Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. *Am. J. Hum. Genet.* **93,** 876–890 (2013).
52. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–770 (2014).
53. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41,** D56–63 (2013).
54. Steinberg, M. H. & Adams, J. G. Hemoglobin A2: origin, evolution, and aftermath. *Blood* **78,** 2165–2177 (1991).
55. 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. EJHG* (2014). doi:10.1038/ejhg.2014.216
56. Goldstein, J. I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinforma. Oxf. Engl.* **28,** 2543–2545 (2012).
57. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
58. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
59. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
60. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).
61. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30,** 97–101 (2002).
62. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria, 2013). at <http://www.R-project.org/>

63. Origa, R. *et al.* Complexity of the alpha-globin genotypes identified with thalassemia screening in Sardinia. *Blood Cells. Mol. Dis.* **52,** 46–49 (2014).
64. 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.* **8,** e1002480 (2012).
65. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5,** e1000529 (2009).
66. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39,** 1197–1199 %U http://proxy.library.upenn.edu:5567/pubmed/17767159 (2007).
67. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39,** 1494–1499 (2007).
68. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39,** 1217–1224 (2007).
69. Veyrieras, J.-B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet.* **4,** e1000214 (2008).
70. Dimas, A. S. *et al.* Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* **325,** 1246–1250 (2009).
71. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
72. Fehrmann, R. S. N. Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genet.* **7,** (2011).
73. Innocenti, F. *et al.* Identification, Replication, and Functional Fine-Mapping of Expression Quantitative Trait Loci in Primary Human Liver Tissue. *PLoS Genet.* **7,** e1002078 (2011).
74. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and Common Regulatory Variation in Population-Scale Sequenced Human Genomes. *PLoS Genet.* **7,** e1002144 (2011).
75. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).
76. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13,** R7 (2012).
77. Wright, F. A., Shabalin, A. A. & Rusyn, I. Computational tools for discovery and interpretation of expression quantitative trait loci. *Pharmacogenomics* **13,** 343–352 (2012).
78. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
79. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
80. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24,** 14–24 (2014).
81. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
82. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinforma. Oxf. Engl.* **26,** 2190–2191 (2010).

Chapter 7: Major height reducing variants and selection for short stature on the island of Sardinia

Based on:

Zoledziewska M, Sidore C*, Chiang CWK*, **Sanna S***, Mulas A, Steri M, Busonero F, Marcus JH, Marongiu M, Maschio A, Ortega Del Vecchyo D, Floris M, Meloni A, Delitala A, Concas MP, Murgia F, Biino G, Vaccargiu S, Nagaraja R, Lohmueller KE, UK10K consortium, Timpson NJ, Soranzo N, Tachmazidou I, Dedoussis G, Zeggini E, The Understanding Society Scientific Group, Uzzau S, Jones C, Lyons R, Angius A, Abecasis GR#, Novembre J#, Schlessinger D#, Cucca F#.*

Nature Genetics 47, 1352–1356 (2015) doi:10.1038/ng.3403 *Epub 15 Sept 2015*

**,# indicate equal contributions*

ABSTRACT

We report sequencing-based whole-genome association analyses to evaluate the impact of rare and founder variants on stature in a cohort of 6,307 Sardinian islanders. We identified two variants with large effects. One is a rare stop codon in the *GHR* gene, relatively frequent in Sardinia (0.87% vs <0.01% elsewhere), which in homozygosity causes the short stature Laron syndrome. We find that it reduces height in heterozygotes by an average of 4.2 cm (-0.64 s.d). The other variant, in the imprinted *KCNQ1* gene (MAF = 7.7% vs <1% elsewhere) reduces height by an average of 1.83 cm (-0.31 s.d.) when maternally inherited. Additionally, polygenic scores indicate that known height-decreasing alleles are at systematically higher frequency in Sardinians than would be expected by genetic drift. The findings are consistent with selection toward shorter stature in Sardinia and a suggestive human example of the proposed "island effect" reducing the size of large mammals.

Human height is a canonical complex trait, under tight genetic control with heritability of 80-90% (1,2). Although rare variants with strong effects have been reported in families with monogenic forms of dwarfism or gigantism, the ~700 reported variants affecting height - which explain only about 16% of the observed heritability - are typically common alleles with modest effect sizes (average <0.3 cm) ([3,4]). Little is known about the impact of rare and founder variants on stature at a population level and whether they contribute to variation in height between populations. The founder Sardinian population is especially suitable to assess the impact of such variants. Although most of the common genetic variants present elsewhere in Europe also exist in Sardinia, the isolated island population is enriched for numerous variants that are very rare or absent elsewhere ([5]) and were not included in the commercial genotyping arrays or multi-population sequencing panels that are commonly used to characterize genetic variants through imputation ([6]).

We therefore used whole genome sequencing to investigate height in a large sample of Sardinians, who, with an average male stature of 168.5 cm ([7]), are among the shortest European populations.

We used whole genome sequencing (~4x) of 2,120 Sardinians to construct a reference panel of ~17.6 million SNPs (**Supplementary Fig. 1a,b**) and carry out a genome wide association study (GWAS) for height. After stringent quality controls and imputation using a scaffold of 890,542 genotyped SNPs, 11,826,948 SNPs were assessed in 6,307 participants in the SardiNIA study, from villages in the Lanusei valley ([1]). The GWAS found two signals strongly associated with stature, one located in the *GHR* (5p12) and the other in the *KCNQ1* (11p15.5) genes, which encode the growth hormone receptor and a voltage-gated potassium channel, respectively (**Supplementary Fig. 1c**). Notably, their joint effect in the SardiNIA cohort is as large as that contributed jointly by the top 10 height associated alleles assessed in the GIANT meta-analysis([4]) and by the top 5 when using the effect sizes observed in the replication set.

The first of these signals is rs121909358 ($p=1.07x10^{-10}$, effect -0.64 s.d. corresponding to -4.2 cm, **Fig.1a, Supplementary Fig. 2**). The height-reducing T allele is found on a single haplotype (**Supplementary a Fig. 3**). It creates a loss of function termination codon (R61X) in *GHR*. The variant and its association with height would not have been detected without imputation from

the Sardinian sequencing panel (imputation accuracy, RSQR=0.94, validated by direct genotyping) ([6]), as the variant is extremely rare outside of Sardinia (frequency <1/60,000, ExAC Browser, **URLs**).

Homozygosity for this stop codon variant is one of several mutations in *GHR* known to cause Laron syndrome (LS) (OMIM#262500); a rare autosomal recessive condition characterized by primary growth hormone insensitivity. Since the initial description ([8]), more than 250 LS cases have been reported (Orphanet, **URLs**), with the majority of patients identified in Maghrebi-Sephardic Jewish groups ([9]) and an isolated population of Spanish descent in Ecuador ([10]). The global estimated prevalence of LS is 1-9 per million (Orphanet, **URLs**) suggesting world-wide carrier frequencies of less than 0.01%. In contrast, we observed an unexpectedly high frequency of 0.87% for the R61X variant among 1,481 unrelated individuals from the SardiNIA cohort. Consistent with this frequency, 1 homozygous affected LS individual has been observed among the 10,721 inhabitants of the 4 villages in the Lanusei valley. The association of R61X with height was replicated in an independent Sardinian cohort of 5,314 individuals from an additional 6 villages (**Supplementary Note**), though its frequency and the effect size are estimated to be smaller (MAF= 0.46% in 857 unrelated individuals, $p_{one\text{-}tail}$ =0.015, effect -0.31 s.d., corresponding to -1.89 cm).

Our results extend to the general population the evidence that *GHR* mutations affect height of heterozygous carriers (**Supplementary Table 1**, [11,12]). In addition, 30% of the carriers from the SardiNIA study also showed limited elbow extension, which is very rare in unaffected individuals but characteristic of LS patients due to underdevelopment of the muscular system and an abnormal degree of humerus rotation (**Supplementary Table 2,** [8]). Interestingly, among 2,120 sequenced Sardinians, we also found instances of two additional rare variants described to cause LS in Southern European and South American populations (**Supplementary Note, Supplementary Table 3**); however those variants were at frequencies too low in the SardiNIA cohort (MAF<0.003) to assess phenotypic effects in heterozygotes.

The second GWAS signal in *KCNQ1* (**Fig. 1b**) is complicated by the fact that it falls in a known tissue-specific imprinted gene cluster. Indeed, we found striking evidence that the association with short stature is maternally inherited (**Fig. 1, Table 1**), with the strongest maternal effects at rs150199504 (MAF= 7.7%, $p=5.6\times10^{-9}$, maternal effect -0.315 s.d., corresponding to -1.83 cm), and no significant paternal effect (p=0.95) (**Table 1, Supplementary Fig. 2**). By directly typing one of the top associated variants, rs2075870, which also showed a modest albeit significant association with decreased height in ~90,000 individuals of European origin ([13]), we confirmed the association in the independent Sardinian cohort ($p=3.6\times10^{-4}$ for the maternal effect -0.22 s.d., corresponding to -1.17 cm and p=0.1 for paternal effect). The association signal spans 48Kb encompassing rs2075870 and 4 additional variants in LD with rs150199504 (pvalue $<1\times10^{-6}$, $r^2>$ 0.7) (**Fig. 1, Table 1**) making it difficult to identify the causal variant (s).

However, we found that differences in allele frequencies and LD patterns among the variants in Sardinia compared to other populations provided a route to prioritize the list for the responsible variant(s) (**Fig. 2**). Remarkably, among the SNPs in LD in Sardinia, we could exclude rs2075870, rs67004488, rs149658560 and rs12790610 as causal based on their frequencies, LD patterns and results from GWAS in other populations. In particular, these variants are common (MAF ~10%), in

LD with each other (r^2>0.3) in South Asia, and yet no association of rs2075870 with height has been observed there ([13]). By contrast, among our core associated SNPs, the top variants rs150199504 and rs143840904 are in lower LD with rs2075870 and much rarer in South Asia (r^2<0.3 and MAF <1.2 % and <2.6% respectively) (**Fig. 2d**) and thus association with height could be missed if they are not directly typed in very large sample sets. Hence, rs143840904 and especially our lead variant rs150199504 are plausible causal candidates.

To further assess their candidacy, we directly tested the 6 core associated variants in 19,053 individuals from 6 GWAS European cohorts, among which we expect more resolving power than in Sardinia due to lower LD in the region (**Fig. 2b, 2c**). Among the 5 variants that passed quality checks, rs150199504 was again the most significantly associated and had the strongest effect in these samples as well ($p=2.82x10^{-4}$, effect -0.243 s.d) – even though it was the rarest of the five (MAF = 0.89 %). To a lesser extent significant association was also seen for rs143840904 ($p=1.23x10^{-3}$, effect -0.145 s.d.), but was not observed for the 3 other variants (**Supplementary Table 4**). Interestingly, in a reciprocal conditional analysis, the effect of rs143840904 was completely accounted for by rs150199504 (p=0.24, effect -0.06 s.d.). By contrast residual association remained at rs150199504 after conditioning on rs143840904 (p=0.06, effect -0.172 s.d.). This further genetic evidence supports rs150199504 as the main driver of the association with decreased height at this locus. Suggestively, rs150199504 (and rs143840904) fall in a differentially methylated region (ENCODE, **URLs**), hinting at a possible effect on expression.

The maternal effect we observed for *KCNQ1* on height is consistent with the established monoallelic expression of maternal alleles at this imprinted locus ([14]). Furthermore, the observation that translocations and inversions disrupting the function of *KCNQ1* result in Beckwith-Wiedemann gigantism ([15]) suggests that, by inference, the short stature alleles reported here result in a gain of function.

KCNQ1 variation has been implicated in several other traits, including platelet aggregation, electrocardiographic measures and type 2 diabetes, with the latter also influenced by parent of origin effects ([16–20]). Those associations were, however, all completely independent of any of the 6 top *KCNQ1* associated variants considered here (r^2<0.08). Furthermore, the 6 variants showed no significant association with any of 193 traits measured in the SardiNIA study participants (data not shown)([1,21]).

To evaluate the overall impact of known variants on the average short stature observed in Sardinia relative to other populations and to test the possibility that short stature might be selected for in this island population, we used polygenic height scores. These scores measure the total frequency of height-changing alleles in a population, weighing each allele by its effect size. A general North-to-South gradient for height in Europe due to directional selection has been reported ([22,23]) with Sardinia as a significant outlier among the Human Genome Diversity Panel European populations (**URLs**). Consistent with these studies, we observed a significantly lower polygenic height score in Sardinia compared to other European populations examined in the 1000 Genomes project, including the Southern European Tuscans and Spanish (**Fig. 3).** Adding our *KCNQ1* and *GHR* variants to the previously described 691 alleles ([4]), the polygenic score of Sardinians decreased by 3.8%. Overall, Sardinian scores are lower than would be expected compared to other European populations ($p=1.62x10^{-6}$, -5.9cm relative to CEU, 1.6% average

increase in frequency for height decreasing alleles), even when calibrating for genome-wide patterns of differentiation due to genetic drift, suggesting that selection has played a role in decreasing height in Sardinia. The differences in height explained by the polygenic score are in accord with the observed ~10 cm of phenotypic differences between Sardinians and the other European populations.

We have also considered the possibility that Sardinians might have an additional contribution of reduced height due to the expression of recessively acting height-decreasing alleles exposed due to founder effects. However, the impact of elevated homozygosity among Sardinians on height appears to be small (0.129 s.d.) relative to the effects predicted by the polygenic score (0.910 s.d.) (**Supplementary Note**).

An example of low frequency allele affecting height was recently reported from the Icelandic population ([24]). However, our findings demonstrate for the first time that part of the missing heritability of human height can be attributable to rare variants involved in monogenic disorders, as shown by *GHR*, as well as by variants common in isolated populations but rare elsewhere, as exemplified by *KCNQ1*. Indeed, a shift toward higher frequencies for variants with large size effects observed in Sardinia ([6,25]) – and in this case the powerful height-decreasing variants -- allowed us to detect, in a cohort of thousands of participants, associations that were missed in GWAS and meta-analyses of hundreds of thousands of individuals.

Intriguingly, the increased frequencies of height-decreasing alleles at *GHR* and *KCNQ1,* and especially the polygenic height scores in this population, are also consistent with the long-standing observation of an "island effect" in which many large animals become adaptively smaller on islands relative to their mainland counterparts ([26]). The extinct Sardinian mammoth (*Mammuthus lamarmorae*) and deer (*Megaloceros cazioti*) are two examples ([27]). One complication to assess this in in humans is that selection for decreased height likely began prior to the peopling of Sardinia among the early European farmer lineage ([28]) that is thought to have initially colonized the island([29]), and Sardinians might have simply retained short stature that evolved earlier. However, we observe lower polygenic height scores in Sardinia even when compared with other populations with high proportions of early European Neolithic ancestry (Tuscans and Spanish)([30]). Thus, selection for decreased height likely continued and was particularly strong in the lineage leading to modern Sardinians. One conjecture is that crop yields or other nutritional sources were limited in the restricted island environment, but exactly why selection for decreased height was acting among the Neolithic ancestors of the Sardinians, and likely intensified after the occupation of the island, remains an open and interesting question.

URLs

HGDP: http://www.hagsc.org/hgdp/index.html

OMIM: http://www.omim.org/

ExAC Browser: http://exac.broadinstitute.org

SardiNIA project home page: https://sardinia.irp.nia.nih.gov/

EPACTS: http://genome.sph.umich.edu/wiki/EPACTS

ENCODE: https://www.encodeproject.org/

GWAS Catalog: https://www.ebi.ac.uk/gwas/

Acknowledgments

We thank all the volunteers who generously participated in this study and made this research possible. All participants provided informed consent and studies were approved by the Local Research Ethic Committees (No 2009/0016600). This study was funded in part by the National Institutes of Health (National Institute on Aging, National Heart Lung and Blood Institute, and National Human Genome Research Institute). This research was supported by National Human Genome Research Institute grants HG005581, HG005552, HG006513, HG007089, 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 grant FaReBio2011 "Farmaci e Reti Biotecnologiche di Qualità"; PB05 InterOmics MIUR Flagship Project; by NIH NRSA postdoctoral fellowship (F32GM106656) to C.W.K.C; by UC MEXUS-CONACYT doctoral fellowship 213627 to D.O.D.V; by Italian Ministry of Education, University and Research (MIUR) no.5571/DSPAR/2002. The HELIC study was funded by the Wellcome Trust (098051) and the European Research Council (ERC-2011-StG 280559-SEPI). The TEENAGE study has been supported by the Wellcome Trust (098051), European Union (European Social Fund—ESF) and Greek national funds through the Operational Programme 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF)—Research Funding Programme: Heracleitus II, investing in knowledge society through the European Social Fund. The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/. This study makes use of data generated by the UK10K Consortium, derived from samples from UK10K_COHORTS_TWINSUK (The TwinsUK Cohort) and UK10K_COHORT_ALSPAC (the Avon Longitudinal Study of Parents and Children). A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. We thank Jeremy Berg for scripts and suggestions on the polygenic score analysis.

Competing financial interests

The authors declare no competing financial interests.

REFERENCES

1. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. 2, e132 (2006).

2. Silventoinen, K., Kaprio, J., Lahelma, E. & Koskenvuo, M. Relative effect of genetic and environmental factors on body height: differences across birth cohorts among Finnish men and women. Am. J. Public Health 90, 627–630 (2000).

3. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–8 (2010).

4. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46, 1173–1186 (2014).

5. Francalacci, P. et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341, 565–569 (2013).

6. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nature Genetics (2015) 47, 1272–1281 (2015) doi:10.1038/ng.3368 *Epub 15 Sept 2015*

7. Arcaleni, E. Secular trend and regional differences in the stature of Italians. J. Anthropol. Sci. Riv. Antropol. JASS Ist. Ital. Antropol. 90, 233–237 (2012).

8. Laron, Z. Laron Syndrome - From Man to Mouse - Lessons from Clinical and Experimental Experience. (Springer-Verlag, 2011). at <http://www.springer.com/us/book/9783642111822>

9. Laron, Z. The syndrome of familial dwarfism and high plasma immunoreactive human growth hormone. Birth Defects Orig. Artic. Ser. 10, 231–238 (1974).

10. Rosenbloom, A. L., Guevara Aguirre, J., Rosenfeld, R. G. & Fielder, P. J. The little women of Loja--growth hormone-receptor deficiency in an inbred population of southern Ecuador. N. Engl. J. Med. 323, 1367–1374 (1990).

11. Laron, Z., Klinger, B., Erster, B. & Silbergeld, A. Serum GH binding protein activities identifies the heterozygous carriers for Laron type dwarfism. Acta Endocrinol. (Copenh.) 121, 603–608 (1989).

12. Guevara-Aguirre, J. et al. Effects of heterozygosity for the E180 splice mutation causing growth hormone receptor deficiency in Ecuador on IGF-I, IGFBP-3, and stature. Growth Horm. IGF Res. Off. J. Growth Horm. Res. Soc. Int. IGF Res. Soc. 17, 261–264 (2007).

13. Lanktree, M. B. et al. Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. Am. J. Hum. Genet. 88, 6–18 (2011).

14. Lee, M. P. et al. Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. Proc. Natl. Acad. Sci. U. S. A. 96, 5203–5208 (1999).

15. Soejima, H. & Higashimoto, K. Epigenetic and genetic alterations of the imprinting disorder Beckwith-Wiedemann syndrome and related disorders. J. Hum. Genet. 58, 402–409 (2013).

16. Horikoshi, M. et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nat. Genet. 45, 76–82 (2013).

17. Johnson, A. D. et al. Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. Nat. Genet. 42, 608–613 (2010).

18. Newton-Cheh, C. et al. Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat. Genet. 41, 399–406 (2009).

19. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. 42, 579–589 (2010).

20. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. Nature 462, 868–874 (2009).

21. Orru, V. et al. Genetic variants regulating immune cell levels in health and disease. Cell 155, 242–56 (2013).

22. Turchin, M. C. et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. 44, 1015–1019 (2012).

23. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. PLoS Genet. 10, e1004412 (2014).

24. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat. Genet. 46, 294–298 (2014).

25. Danjou, F. et al. Whole genome sequencing-based gwas in sardinia explicates genetic regulation of hemoglobin levels and clinical consequences. Nature Genetics 47, 1264–1271 (2015) doi:10.1038/ng.3307 *Epub 15 Sept 2015*.

26. Millien, V. Morphological evolution is accelerated among island mammals. PLoS Biol. 4, e321 (2006).

27. van der Geer, A., Lyras, G., de Vos, J. & Dermitzakis, M. in Evolution of Island Mammals 103–130 (Wiley-Blackwell, 2010). at <http://onlinelibrary.wiley.com/doi/10.1002/9781444323986.ch9/summary>

28. Mathieson, I. et al. Eight thousand years of natural selection in Europe. bioRxiv 016477 (2015). doi:10.1101/016477

29. Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413 (2014).

30. Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature advance online publication, (2015).

FIGURES AND LEGENDS

Figure 1. Regional association plots for *GHR* and *KCNQ1* locus. a) *GHR* locus, the Y axis shows the association strength (–log10 pvalue) versus the genomic positions (hg19/GRCh37) around the most significant SNP (purple). Other SNPs in the region are color-coded to reflect their LD with the top SNP. Symbols reflect genomic functional annotation. Genes and the position of exons are shown below. b) Regional plot at the *KCNQ1* locus for the paternal and maternal effects respectively. The position of GWAS catalog SNPs (**URLs**) with the corresponding traits and the position of exons in the *KCNQ1* region are indicated below.

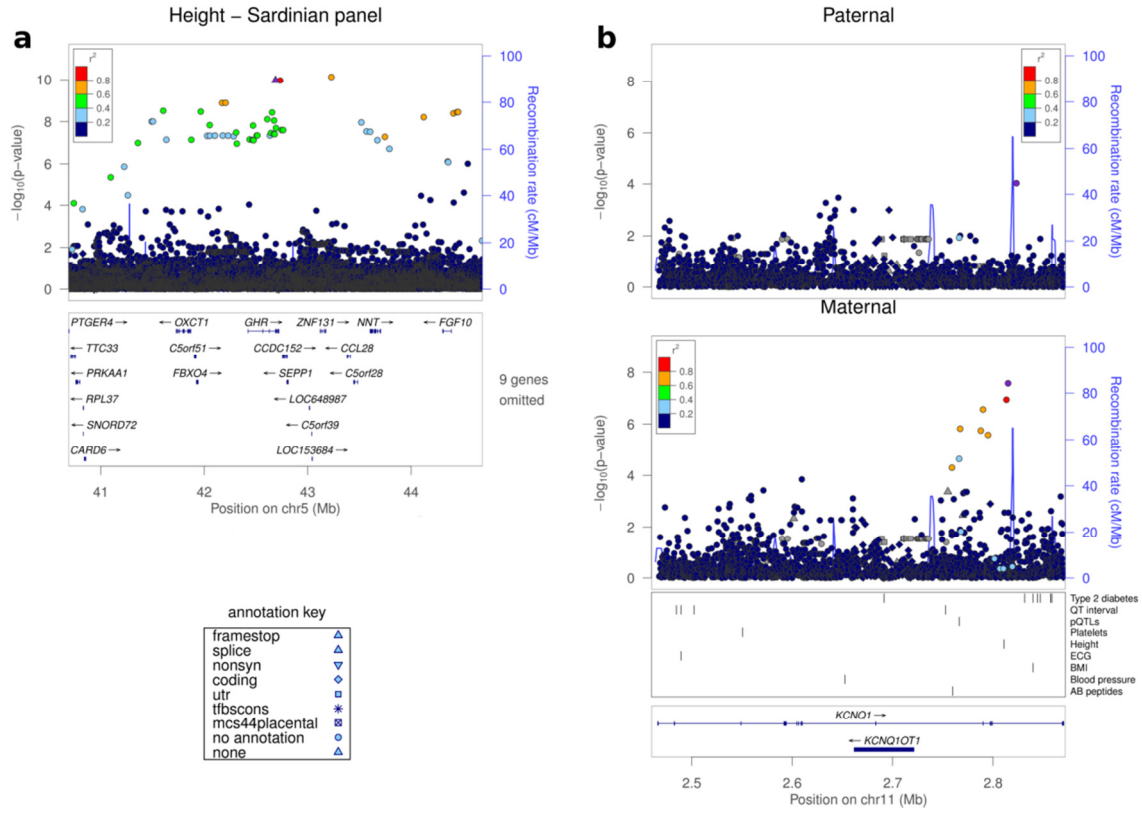


Figure 2. Worldwide frequency and LD pattern for the six top *KCNQ1* SNPs. The figure illustrates the frequency (upper panel) and the pairwise LD matrix (lower panel) for the six top SNPs associated in Sardinia at the *KCNQ1* locus. Data are presented for 4 populations: a) Sardinia, b) Southern Europe, c) Northern Europe, d) South Asia. Matrix cells are colored according to the LD value: green if r^2>=0.7; yellow if 0.3<= r^2 <0.7 ; red if r^2<0.3.

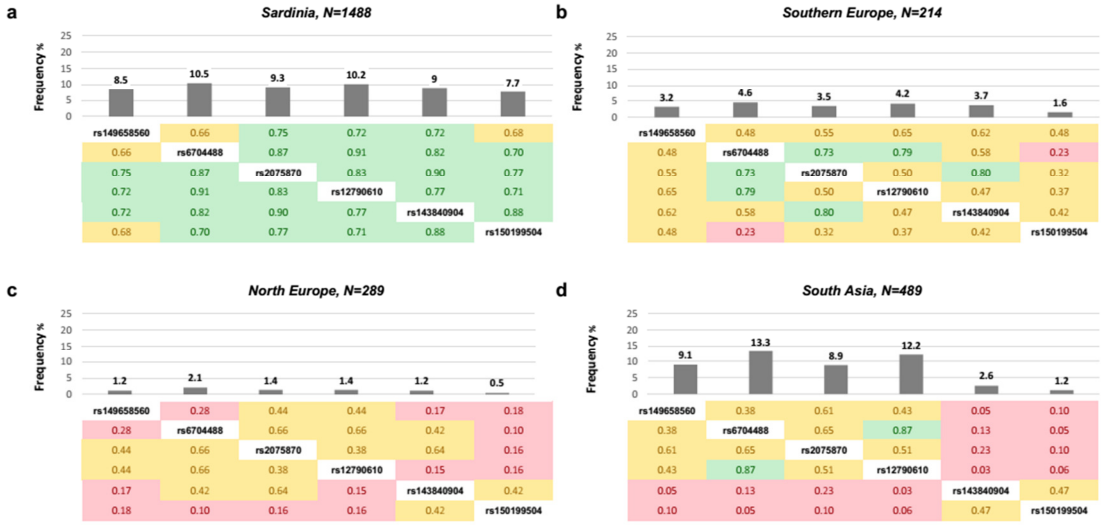


Figure 3. Polygenic score analysis for height. Polygenic score based on the 2 top associated variants (rs121909358 and rs150199504) and the 691 height loci from GIANT for which the effect size in Sardinia and allele frequencies in 1000 Genomes phase 3 data are available. The black circles indicate the scale for display of p-values according to circle size. Abbreviations: SDI: SardiNIA cohort; IBS, TSI, GBR, CEU, and FIN: 1000 Genomes populations.

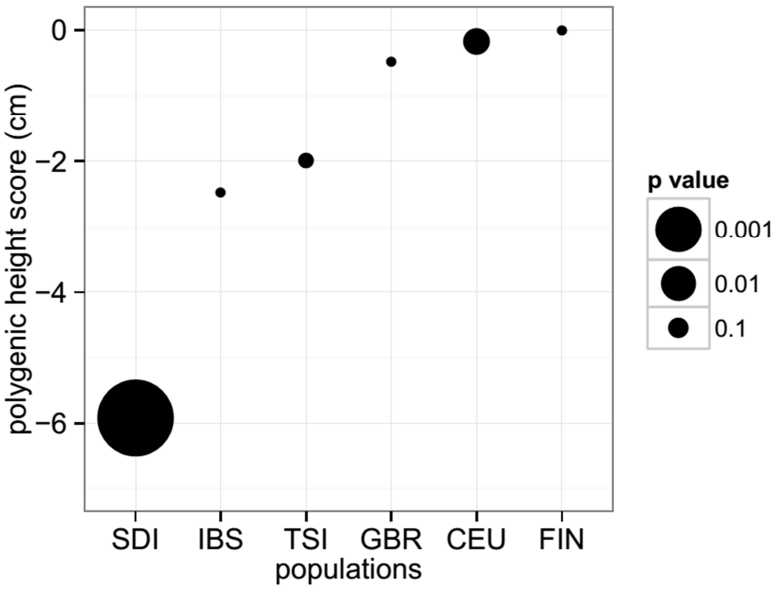


Table 1. Parental of origin effects at *KCNQ1*. The table summarizes the strongest results for the parental of origin association test at the *KCNQ1* locus (defined as pvalue<$1x10^{-6}$ in either the maternal or paternal tests for the assessed 500Kb region). At each SNP, we report in the column N the number of informative transmissions used (see Methods) and the association parameters obtained evaluating the minor allele i) without considering parent of origin, ii) when maternally inherited, and iii) paternally inherited. The last column reports the pvalue for heterogeneity between estimated paternal and maternal effects.

					Both		Maternal		Paternal		
rs ID	**Chr:Position**	**Minor Allele/ Other**	**MAF**	**N**	**Effect (StdErr)**	**Pvalue**	**Effect (StdErr)**	**pvalue**	**Effect (StdErr)**	**pvalue**	**Heterogenity pvalue**
rs150199504	11:2814960	G/C	0.083	5059	-0.168 (0.039)	$1.84x10^{-5}$	-0.315 (0.054)	$5.56x10^{-9}$	-0.0032 (0.050)	0.9488	2.46x10-5
rs143840904	11:2813322	T/C	0.094	5041	-0.152 (0.038)	$4.58x10^{-5}$	-0.274 (0.050)	$3.92x10^{-8}$	+0.0021 (0.049)	0.9653	$7.55x10^{-5}$
rs2075870	11:2790019	A/G	0.094	5044	-0.158 (0.038)	$2.65x10^{-5}$	-0.273 (0.051)	$6.97x10^{-8}$	-0.0172 (0.048)	0.793	0.0002
rs149658560	11:2767262	A/G	0.076	5050	-0.161 (0.042)	$1.01x10^{-4}$	-0.297 (0.058)	$2.93x10^{-7}$	-0.0121 (0.052)	0.8183	0.0003
rs12790610	11:2794998	G/A	0.095	5014	-0.165 (0.037)	$1.02x10^{-5}$	-0.258 (0.051)	$4.73x10^{-7}$	-0.044 (0.048)	0.3531	0.0023
rs67004488	11:2787804	G/A	0.104	5026	-0.157 (0.036)	$1.2x10^{-6}$	-0.244 (0.049)	$5.21x10^{-7}$	-0.040 (0.047)	0.3875	0.0024

Online Methods

Research subjects.

All individuals included in the study were of Sardinian origin and participate in a longitudinal study of age-related quantitative traits on the island (SardiNIA, **URLs**). The study involves four villages: Lanusei, Ilbono, Elini end Arzana, located in the Lanusei Valley([1,21,31]). 6,148 volunteers have been described before ([1]) and an additional 773 individuals have been enrolled during the follow up stage of the project ([6]). 6,602 individuals had complete genotyping data. For analyses, we only included measurements for individuals at age >20 years, and also discarded 4 subjects with Morquio Syndrome (OMIM *607939), leading to a total of 6,307 samples.

All participants provided informed consent and studies were approved by the Local Research Ethic Committees (No 2009/0016600).

Genotyping methods, low-pass sample sequencing, variant calling, genotype imputation and GWAS analysis.

All SardiNIA individuals were typed with four Illumina Infinium arrays. Low pass sequencing, variant calling, genotype imputation and GWAS analysis was conducted as previously described ([31]).

GWA analysis.

For our GWAS we tested association for the 11,826,948 imputed or genotyped variants that passed quality control filters [MACH r^2>0.3 for MAF>=0.01, r^2>=0.6 for MAF<0.01 ([31])], assuming an additive model of inheritance and adjusting for age, age squared and gender as covariates and applying the inverse normal transformation to the residuals. Association was performed using EMMAX ([32]) as implemented in the software EPACTS (**URLs**), which accounts for relatedness and population structure using an empirical kinship matrix derived from genotype data. The genomic control inflation factor was λ=0.989, indicating no inflation of results.

Validation of imputation results by genotyping.

GWAS identified three loci significantly associated with stature: the *GHR* gene, with top variant rs121909358; the *KCNQ1* gene, with 6 variants in LD (**Table 1**); and the *SMURF2* gene, with top variant rs143051029.

We validated imputation of rs121909358 genotypes by directly genotyping 2,818 samples with a TaqMan assay. Concordance between imputation and validation was 99.89%. At *KCNQ1*, two leading variants, rs67004488 and rs2075870, were present on the Cardio-Metabo Illumina chip, so that validation was not necessary. The third association at rs143051029 was evaluated with standard Sanger sequencing. We selected 96 samples for sequencing, including 4 imputed homozygotes, 22 imputed heterozygotes with uncertain allele dosages and 70 randomly selected samples. The variant, located in a complex region, did not pass validation due to the high mismatch rate (34.4%) between imputed genotypes and those validated by Sanger sequencing and was not further considered in analyses.

Conditional analysis.

We conducted standard conditional analyses using EPACTS software for the two identified regions by including the top variants as covariates. We examined the 1Mb region around the top SNPs (rs121909358

for *GHR* and rs150199504 for *KCNQ1*). In both cases, the top variant completely explained the association at the two loci; none of the SNPs in the region passed the significance threshold after Bonferroni correction. The variant chr5:43229441, 540Kb away, from rs121909358, was fully explained by the effect of rs121909358 (p after conditional = 0.1).

Replication cohorts.

We replicated findings in an independent cohort of 5,314 Sardinians and 19,053 non-Sardinian European samples. Details on genotyping and analyses are described in Supplementary.

Characterization of the associated region on chromosome 5.

To visualize the haplotypes carrying the Laron variant (**Supplementary Fig. 3**), we interrogated ±3Mb surrounding chr5:42689036 in 11 sequenced unrelated carriers of rs121909358. The analysis was performed using SelScan ([33]) and included 9,526 SNPs with MAF >5% in Sardinia.

Parent-of-origin effects.

For SNPs in the *KCNQ1* locus, we estimated parental origin of alleles for all individuals using Merlin (*--best* option)([34]). We then considered two separate variables, one for the maternal (G_m) and one for the paternal (G_p) allele, coded as 1 if the corresponding transmitted allele was the minor allele at the SNP, and 0 otherwise. Missing values were assigned to founders and other individuals for whom parental origin could not be defined unambiguously. Of consequence variables G_m and G_p were non-missing for 5,026 SardiNIA individuals and 4,666 OGP individuals. Two linear models were then used:

$Y \sim \beta_0 + \beta_1 G_m + \beta^T C$
$Y \sim \beta_0 + \beta_2 G_p + \beta^T C$

where Y denotes trait and C, other covariates. As both the SardiNIA and OGP studies consists of large families, the transmissions evaluated by G_p and G_m are not independent. We therefore tested the null hypothesis $\beta_1 \neq 0$ (for model 1) and $\beta_2 \neq 0$ (for model 2) by fitting a mixed linear regression model that accounts for familiar relatedness (*lmekin()* and *kinship()* functions in the coxme and kinship R packages). In the models, we used the same covariates and trait normalization procedure as in the GWAS analysis. We then assessed the hypothesis of heterogeneity of effects, $\beta_1 \neq \beta_2$, using Cochran's Q statistic. The test was carried out for all SNPs in the *KCNQ1* gene, and on SNP rs2075870 in the OGP cohort.

Population-level height polygenic score calculation and evaluation.

In the population genetic analyses, we focused on a subset of 1,081 unrelated sequenced individuals (**Supplementary Note**).

To investigate whether height-decreasing loci have been under selection in Sardinia, for each population *m*, we calculated the polygenic height score as

$$Z_m = 2\sum_{l=1}^{L} \beta_l p_{ml}$$

where β_l is the effect size of the height-increasing allele *l* and p_{ml} is the frequency of allele *l* in population *m*. To avoid biases and to ensure uniformity of the source of effect size estimates, we used the effect size

estimates from the Sardinian dataset regardless of whether the variant is significantly associated with height in this dataset. We first calculated the polygenic height score (Z_m) based on the 691 height loci identified by the GIANT consortium ([4]) with effect sizes estimated in the Sardinian dataset and then added the two top variants reported, totaling 693 height alleles. To test if there were a signature of polygenic adaptation on height in Sardinia, we adopted a framework developed by Berg and Coop ([23]), which builds a multivariate normal model based on matched, presumably neutral variants, to account for relationships among populations (**Fig. 3**). Populations with extreme polygenic scores relative to the expectation (pvalue = 0.01) are likely to have undergone selection. To construct a null distribution of frequencies needed for the multivariate normal framework, we obtained for each of the height loci all variants in the 1000 Genomes phase 3 European data with minor allele count +/- 10 counts (~ 1% in frequency), B score ([35]) +/- 50 units, and local recombination rates +/- 0.5 cM/Mb. A random subset of 509,386 SNPs, representing 10% of the union of the matched SNPs, were then used as a set of matched SNPs for the analysis. Of note, we also repeated the calculation using effect sizes estimated by the GIANT consortium as well as using only a subset of 162 SNPs that are not subject to population stratification ([22]) (**Supplementary Fig. 4**).

Methods references

31. 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

32. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42, 348–54 (2009).

33. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. 31, 2824–2827 (2014).

34. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 30, 97–101 (2002).

35. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5, e1000471 (2009).

Part III: Improving GWAS studies with population-specific reference panels

Chapter 8: Population specific imputation panels as a general tool to enhance genetic discoveries

Serena Sanna[1] and Cisca Wijmenga[2]

1. Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, Italy

2. Department of Genetics, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands

After the completion of the Human Genome Project and the advent of genome-wide association studies (GWAS), thousands to million of markers across the genome were evaluated for several complex traits and diseases. Those studies revealed that, with only few exceptions, the genetic component of complex traits and diseases is fractioned into multiple variations of moderate or small impact (effect size) rather than a few with large effect size. The number of DNA variants contributing to the polygenic liability of a disease or trait's variation is unknown *a priori*, and it can be extremely large even for highly heritably traits. For example, in the past decade GWAS studies have identified up to 697 independent single nucleotide polymorphisms (SNPs) that influence human height, but their global contribution explains only 20% of the estimated heritability [*Wood et al, Nature Genetics 2014*]. Therefore many more associated variants have yet to be found for human height. The scenario is similar for other complex traits: despite the many SNPs identified, a substantial fraction of the heritability remains unexplained [*Manolio et al, Nature 2009*]. What is the cause of this "missing heritability"? Previous genome-wide association studies have assessed thousands of individuals, but they focused mostly on common SNPs (minor allele frequency (MAF) in the population >5%) which were experimentally derived by commercial genotyping arrays or statistically inferred with genotype imputation methods and the HapMap Project [*Li et al, Annu Rev Genomics Hum Genet. 2009*]. Low frequent (MAF<5%) and rare variations (MAF<1%), as well as other types of genetic alterations, have therefore been largely unexplored.

In principle, known variants in this frequency range can be assessed by custom genotyping arrays. However, as custom arrays can only include a limited number of variants, one needs to focus on a specific subset. For example, the Illumina ExomeChip custom array was designed to assess ~200,000 variants in coding regions, while the ImmunoChip and Cardio-MetaboChip were set up to study a similar number of variants in genes associated or potentially involved in immune or cardio-metabolic traits, respectively. The cost of such arrays is still affordable and allowed the characterization of hundreds of thousands of individuals. It has to be noted that those arrays are limited not only in the number of variants that can be tested, but are also limited to those variants that are known by the time of the array's design.

Identification of novel variants and assessment of the full variation present on a genome is possible by whole-genome sequencing. However efficient detection of rare and low frequency variants requires sequencing hundreds to thousands of individuals of which the cost is still prohibitively high. An alternative cost-effective approach is to sequence a subset from a study sample that incorporates a maximal number of variants (i.e. founders individuals), and use their haplotypes to impute the missing genotypes in the other study samples with the genotype imputation approach [*Li et al, Annu Rev Genomics Hum Genet.*

2009]. The benefits are higher for homogenous populations, especially in isolates, as there are fewer haplotypes to be tracked down, but there is clear evidence of successful designs in "open" populations. An additional route that only has a computation cost is to use publicly available sequencing data, such as the 1000 Genomes Project.

The 1000 Genomes Project

The goal of the 1000 Genomes Project was to find most genetic variants that have frequencies of at least 1% in the populations studied, by sequencing the whole-genome of 2,500 individuals from 26 different populations among five continental groups—Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR) and South Asia (SAS). The effort is an extension of the HapMap project which was created to catalogue common genetic variation (MAF>5%). The completion of the 1000 Genomes Project was announced recently [*1000 Genomes Consortium, Nature 2015*], but several phases of interim data have been released starting from 2009. At each release, the data was promptly formatted to be used with the most commonly used genotype imputation software (MACH/minimac, http://genome.sph.umich.edu/wiki/Minimac; IMPUTE/IMPUTE2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html; Beagle https://faculty.washington.edu/browning/beagle/b3.html), and to date there are already hundreds of publications that have benefitted from this resource in genome-wide association studies (GWAS catalog, www.genome.gov/gwastudies). The 1000 Genomes catalogued 84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes [*1000 Genomes Consortium, Nature 2015; Sudmant et al, Nature 2015]*. This is a massive repository of variations, and represents a unique resource for studies that cannot afford whole-genome sequencing of their samples of interest. However, large population-specific reference panel are expected to be useful for genotype accuracy, as only ~500 individuals per ancestry were sequenced and therefore low frequent and rare variation may still be missed or poorly represented. Evidence of successful results with population-specific reference panels are discussed below.

Population-specific reference panels in isolates

The first example of successful population-specific reference panel design in the era of whole-genome sequencing has been seen in the Icelandic population. In 2011, the deCode group (http://www.decode.com/) identified a novel susceptibility locus for sick sinus syndrome (SSS) by combining both public and population specific whole-genome sequences with GWAS-genotyping array data of >38,000 Icelandic individuals [*Holm et al. Nature Genet 2011*]. A first GWAS for SSS with 792 cases and 37,592 controls was carried out to assess 7.2 million SNPs, directly genotyped or imputed from either HapMap or 1000 Genomes pilot 1. The results highlighted a novel locus on chromosome 14q11 with three statistically indistinguishable variants, at which the minor allele frequency was low (1–2.6%), and the minor allele was the risk allele. To refine this association, 7 SSS cases, enriched for carriers of the detected variants, and 80 controls were whole-genome sequenced at 10× depth on average, and then the ~11 million detected variants were imputed into the full GWAS data set. The increased genomic resolution allowed to narrow the signal at 14q11 and point to a missense variant, c.2161C>T, located in exon 18 of the *MYH6* gene, encoding the alpha heavy chain subunit of cardiac myosin. The signal was confirmed by direct genotyping and replication in an additional cohort, independent from the GWAS study. The c.2161C>T

variant was neither present in the available 1000 Genomes Project data nor in the HapMap samples and neither in 121,390 chromosomes inspected by the Exome Aggregation Consortium (http://exac.broadinstitute.org/); therefore this association could not be found in other populations or even in the Icelanders without their population-specific sequencing panel (**Table 1**).

The same group of scientists have continued to sequence Icelandic individuals and recently published novel genetic discoveries with the use of an expanded population-specific reference panel. A total of 2,636 Icelanders were whole-genome sequenced at a medium depth of 20×, leading to the characterization of 20 million SNPs and 1.5 million indels which were subsequently imputed in 104,220 Icelanders genotyped for 676,913 autosomal SNPs using Illumina chip arrays [*Gudbjartsson D. et al Nature Genet 2015*]. With this extended data set the authors carried out several GWAS for different traits and diseases as well as different model of inheritance (additive, recessive, parental-of-origin specific). They identified a novel association at *MYL4*, where a rare c.234delC frameshift deletion (MAF=0.65%) strongly increases the risk of early-onset atrial fibrillation; a novel association with gallstone disease and liver function at two rare (MAF=0.22% and 0.21%) coding variants in the *ABCB4* gene (p.Gly622Glu and p.Leu445Glyfs*22); and a novel association between thyroid stimulating hormone (TSH) levels and a rare variant (rs139242164, MAF=0.44%) in the *GNAS* gene, when the minor allele was maternally inherited. The latter variant was likely missed in previous European studies because parent-of-origin effects are not assessed in standard GWAS; the variant is also more frequent in other European populations (MAF 1-2%, based on 1000 Genomes). By contrast, the first and third variants are absent from the most recent 1000 Genomes release and Exome Aggregation Consortium (ExAC, http://exac.broadinstitute.org/), and the second (p.Gly622Glu) has been seen in only 1 chromosome among the 121,354 assessed by the ExAC (**Table 1**). Therefore and again, the results could have been achieved only by means of the population-specific sequencing panel.

Recently, large scale sequencing reference panels have been reported for another large population isolate: Sardinians. By low-pass whole-genome sequencing of >2,000 individuals and integration with genotyping arrays in > 6,000 individuals, scientists from the SardiNIA cohort have documented the relevance of a population-specific panel not only as average increase in imputation accuracy [*Pistis et al, Eur J Hum Genet 2014*] but also as an effective tool for genetic discoveries [*Orrù V et al, Cell 2013; Sidore C et al, Nature Genet 2015; Danjou et al, Nature Genet 2015; Zoledwieska et al, Nature Genet 2015*]. The group reported sequencing-based GWAS analysis for 285 quantitative traits (272 immune related traits, 12 blood markers and height). Among the novel signals, many would have been missed without the Sardinian sequencing panel because they were either i) absent in the 1000 Genomes panel or ii) too rare to be well imputed. In fact, when repeating the analyses using 1000 Genomes for imputation, such associations were either below genome-wide significance or misplaced to a nearby variant (**Table 1**). For example, Sidore and colleagues [*Sidore C et al, Nat Genet 2015]* reported an association with triglyceride levels and a missense variant, R282S, in a known gene, *APOA5*; the frequency is 2.5% in Sardinia but is currently absent from 1000 Genomes and to date it has been found only on two among 120,520 chromosomes characterized in the Exome Aggregation Consortium. This variant would therefore be missed without the population-specific reference panel. There were two other examples of alleles that rose in frequency in Sardinia and for which imputation was aided by the customized haplotype reference set. One is a stop codon variant (Q40X) in the *HBB* gene [*Sidore et al, Nature Genet 2015]* which was shown to be associated with LDL-cholesterol levels. The minor allele has been seen in one chromosome among the 5,008 sequenced in the 1000 Genomes Phase 3 project but is instead fairly common in Sardinia (MAF=4.8%) due to selection forces. The imputation with 1000 Genomes panel was very inaccurate (imputation accuracy RSQR=0.31) and the association at this locus was misplaced to an intronic marker (rs76053862) located 122 kb away from the

coding SNP. The second is a variant near the *CCND3* gene, which showed association with HbA2, an adult form of hemoglobin [*Danjou et al, Nature Genet 2015]*. The variant, rs113267280, has a MAF of 1% in Europe, but there was a 10-fold increase in frequency in Sardinia and therefore 1000 Genomes imputation misplaced the association to an intronic SNP located 202Kb away. Because this variant is relatively common in populations outside Sardinia, replication of this finding was possible in an independent cohort from UK.

There was another interesting finding in the SardiNIA sequencing based GWAS that demonstrated the value of a large scale population-specific reference panel to efficiently detect and accurately impute rare and low frequency variants. In an association scan for height variation, authors identified a rare variant (encoding p.Arg61*, MAF=0.87%) which creates a loss-of-function termination codon in the *GHR* gene [*Zoledwieska et al, Nature Genet 2015]*. The imputation accuracy was high (RSQR=0.94) and estimated genotypes were highly concordant when consequently validated with experimental methods (99.89%). This variant is one of the mutations in *GHR* known to cause Laron syndrome (Online Mendelian Inheritance in Man (OMIM), #262500), a rare autosomal recessive condition characterized by primary growth hormone insensitivity. Its frequency is extremely low outside Sardinia (currently seen only in 2 chromosomes among 121,388 assessed by the ExAC) and replication was only possible in an independent Sardinian population cohort.

Population-specific reference panels are being created also for genetic isolates in a broader sense, such as the Ashkenazi Jews (AJ), identified as Jewish individuals of Central- and Eastern European ancestry in the United States. Genetic analyses of recent AJ history highlighted a narrow population bottleneck of only hundreds of individuals in late medieval times, followed by rapid expansion, suggesting that whole-genome sequencing of a limited number of samples representing diversity in the settlers group could catalogue nearly all founder variants. Carmi and colleagues [*Carmi S et al, Nature Commun. 2014*] reported a population specific (AJ) reference panel set up with 128 high depth genomes (>50x) and they estimated that the panel improves imputation accuracy for AJ SNP arrays by 28%. Imputation with this panel lowers the number of wrongly imputed non-reference variants with MAF <1% by 2.7-fold, with the improvement remaining at 1.5–2-fold at higher frequencies. These results motivate using a population-matched, rather than a merely continent matched, reference panel, even for the closely related AJ and European populations.

Population-specific reference panels in open populations

Non-isolate populations of European origin can be relatively well imputed with publicly available panels such as 1000 Genomes. However, because the degree of ancestry matching between the genotyped sample (to be imputed in) and the reference haplotype panel (e.g., 1000 Genomes) as well as the number of individuals in the reference haplotype panel are both key ingredients for genotype imputation, large population-specific reference panels can further increase imputation accuracy. This is especially true for low-frequency and rare variants. The number of individuals in the panel is important because imputation accuracy is in part a function of how many copies of a variant exist in the haplotype panel. If only one copy exists (i.e., a singleton), that variant will likely be difficult to impute accurately. One simple way to increase the number of copies of a variant is to increase the number of individuals in the haplotype panel. Furthermore, by sequencing thousands of individuals there is an increased chance to detect rare sites that are missed or poorly represented in 1000 Genomes Project and that can be associated to phenotypes variation. Several sequencing efforts are currently ongoing in this direction for many populations, and results reported so far are encouraging. For example, Vrieze and colleagues carried out whole-genome

sequencing at moderate-high depth (10x on average) on 1,325 individuals of European origin living in Minnesota [*Vrieze et al, Psychophysiology 2014*]. They identified 27.1 million autosomal variants, of which 21.3 Million have MAF <5% in the samples studied. Using this reference panel for imputation in a cohort of 6,610 Minnesotans accuracy increased by 36% compared to that observed with 1000 Genomes [*Vrieze et al, Psychophysiology 2014; Pistis et al Eur J Hum Genet 2014].*

A very large reference panel has been created for the British population by the UK10K consortium. A total of 3,781 British individuals were sequenced at low depth (average 7x), and when compared to two large-scale European sequencing repositories, the effort led to the discovery of over 24M novel single nucleotide variants, of which 99% had MAF <1% [*The UK10K Consortium, Nature 2015*]. When using this British-specific reference panel, imputation accuracy in British cohorts increased at all frequency ranges, and further increases when combining the panel with 1000 Genomes haplotypes [*Huang et al, Nature Commun. 2015*]. For example, for variants with MAF between 0.5% and 1%, the average imputation Rsq is 0.477 with 1000 Genomes, and increases to 0.573 and 0.702 when using the UK10K and UK10K+1000Genomes panels, respectively. Using this combined reference panel in a GWAS of >9,000 British individuals for lipids and inflammatory levels, two novel variants were identified. The first was a low-frequency intronic variant (MAF 2.6%) in *ADIPOQ* associated with decreased adiponectin levels, and the second was a rare splice variant (rs138326449) in *APOC3* [*Timpson, et al. Nature Commun. 2014*] associated with triglycerides. The minor allele at the splice site was seen in 1% of the British chromosomes, while it appears very rare (<0.5%) in other European populations. The variant was previously found to be associated with triglycerides in a study that combined genotyping data with exome-sequencing and used a specific statistical test that aggregates rare mutations in one unique score to improve power [*The TG and HDL Working Group of the Exome Sequencing Project, NHLBI. N. Engl. J. Med. 2014*]. It could not have been found using standard single variant tests outside of the British population and without the population specific panel.

A population-specific reference panel was also built for the Dutch population within the Genome of the Netherlands (GoNL) Project [*Francioli et al, Nature Genet 2014*]. By whole-genome sequencing 769 individuals in 250 families at ~13× coverage, the project built a resource of 1,000 independent haploid genomes as representative of a small (41,543-km2), densely populated (>17 million inhabitants) country in northwestern Europe. The project discovered 20.4 million SNPs in addition to 1.2 million biallelic indels (<20 bp in length) and 27,500 larger deletions (>20 bp in length). Of the SNPs, 6.2 million were common (MAF>5%), 4.0 million are low frequency (MAF=0.5–5%), and 10.2 million are rare (MAF<0.5%). Relatively to dbSNP (release 137) and the 1000 Genomes Project Phase 1 and HapMap CEU panels, GoNL identified 7.6 million novel sites of which the majority are very rare (MAF < 0.5%), including 5.8 million singletons. The panel improved imputation accuracy when inferring missing sites in a Dutch samples set, and for low frequent and rare variants there was a further gain when incorporating the 1000 Genomes sequence data, albeit smaller than that estimated for the British population. Specifically, the average imputation accuracy was 0.65 for variants with MAF 0.5-1%, and increased to 0.75 and 0.77 when using the GoNL and GoNL+1000KG, respectively [*Francioli et al, Nature Genet 2014; Deelen et al, Eur J Hum Genet 2014*]. The GoNL reference panel was used to impute nine large Dutch biobanks (~35,000 samples) and perform association analyses on blood lipid levels [*van Leeuwen et al, Nature Comm 2015*]. The nine cohorts were imputed and analyzed independently and the statistics were meta-analyzed. The results highlighted five novel signals at four loci, of which three have an increased frequency in GoNL compared with 1000 Genomes, suggesting that there may have been genetic drift in the Dutch population for these loci. The most interesting is a rare missense variant, which is 3.65-fold more frequent in the Dutch compared to other European populations (frequency 3.4% vs 0.5% in 1000 Genomes non-CEU samples)(**Table 1**). The

GoNL imputation panel was therefore useful to accurately estimate genotypes at this site. The association was seen with both LDL-C and total cholesterol and points to the *ABCA6* gene (ATP-binding cassette, subfamily A (ABC1), member 6). The mutation changes the amino acid cysteine into arginine at position 1359 (Cys1359Arg) and is predicted to be damaging for the structure and function of the protein [*van Leeuwen et al, Nature Comm 2015*].

The findings from GoNL and UK10K suggest that efforts with next-generation sequencing to build population-specific imputation panels will enhance discovery of clinically relevant findings even in open populations.

Other community resources

It has been clear that large and ancestry-matched reference panels lead to better accuracy and more discoveries: many research groups are sequencing hundreds to thousands individuals to create a study-specific panel that well matches with the samples in a cohort of interest. It is also clear that for the rare and very rare sites there is always an improvement in combining sequencing data from other, even diverse, populations, for example by combining the population-specific panel with 1000 Genomes [*Huang et al, Nature Commun. 2015; Deelen et al, Eur J Hum Genet 2014; Vrieze et al, Psychophysiology 2014*]. Those observations motivated a community-wide effort to create a unified reference panel across diverse populations: the Haplotype Reference Consortium (HRC, http://www.haplotype-reference-consortium.org/). The HRC will create the largest reference panel for imputation by collaborating with all single research groups that are carrying out whole-genome sequencing, including GoNL, UK10K and SardiNIA**.** The first release of the HRC panel includes the 1000 Genomes Project Phase 3 data as well as additional ~ 30,000 samples, mostly of European origin. In the future, the reference panel will increase in size and include samples from a more diverse set of world-wide populations. Free imputation servers will allow anyone to use the full haplotype reference panel to impute missing genotypes in their data: users will be able to upload genotype data to the server, imputation will be carried out remotely on the server, and the imputed data will then be made available to the user (https://imputation.sanger.ac.uk/; https://imputationserver.sph.umich.edu/). This is a valuable resource especially for medium and small size laboratories that do not have sufficient expertise and computational capacity to carry out imputation, and it represents a step toward a responsible sharing of genomic research data.

Table 1. Novel loci detected with population-specific reference panels

The table shows associations whose discovery was enhanced, or only detectable, with population-specific reference panels. For each variant, we list the associated trait, the rs number (when available, or chromosome and position or substitution as given in the original manuscript), the population where it was discovered, the minor allele frequency (MAF) in other Europeans according to 1000 Genomes estimates, and the MAF in the ExAC browser among all samples studied (only for variants in a coding region). hsCRP= high sensitivity C-reactive protein; ESR= erythrocytes sedimentation rate.

Associated trait	**Variant (Gene)**	**Discovery Population**	**MAF in the discovery population**	**MAF in other Europeans (1000Genomes)**	**MAF in ExAC (if coding)**
Sick Sinus syndrome	c.2161C>T/p.Arg721Trp (*MYH6*)	Icelanders	0.38%	absent	absent
Early-onset atrial fibrillation	c.234delC/p.Cys78Trpfs*29 (*MYL4*)	Icelanders	0.65%	absent	absent
Gallstone disease and liver function	p.Gly622Glu (*ABCB4*)	Icelanders	0.22%	absent	0.00082%
Gallstone disease and liver function	p.Leu445Glyfs*22 (*ABCB4*)	Icelanders	0.21%	absent	absent
CD39+ CD4+ cells count	10:98088623 (near *ENTPD1*)	Sardinians	3.2%	absent	---
CD62L- myeloid dendritic cells count	rs58055840 (near *FCGR3A*)	Sardinians	26%	14%	---
LDL cholesterol, total cholesterol	rs11549407 (*HBB*)	Sardinians	4.8%	absent	0.04%
Triglycerides	11:116661101 (*APOA5*)	Sardinians	2.5%	absent	0.0016%
hsCRP, ESR	12:125406240 (near *AACS*)	Sardinians	0.7%	absent	---
Hemoglobin A1	12:123681790 (near *MPHOSPH9*)	Sardinians	1%	absent	---
Hemoglobin A2	rs113267280 (near *CCND3*)	Sardinians	10%	1%	---
Hemoglobin A2	rs141006889 (*FOG1*)	Sardinians	0.7%	absent	0.06%
Fetal Hemoglobin	rs183437571 (near *NFIX*)	Sardinians	1%	absent	---
Height	rs121909358 (*GHR*)	Sardinians	0.87%	absent	0.0016%
Triglycerides	rs138326449 (*APOC3*)	British	1%	0.5%	0.14%
LDL cholesterol	rs77542162 (*ABCA6*)	Dutch	3.4%	0.5%	1.08%

References

1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing*. **Nature** 467, 1061–1073 (2010)

1000 Genomes Project Consortium. *A global reference for human genetic variation.* **Nature** 526, 68–74 (01 October 2015) doi:10.1038/nature15393

Carmi S, et al. *Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins.* **Nat Commun.** 2014 Sep 9;5:4835. doi: 10.1038/ncomms5835.

Danjou F, et al. *Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels.* **Nat Genet**. 47, 1264–1271 (2015) doi:10.1038/ng.3307 Epub 15 Sept 2015

Deelen P, et al. *Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands*. **Eur J Hum Genet**. 2014 Nov;22(11):1321-6. doi: 10.1038/ejhg.2014.19. Epub 2014 Jun 4.

Francioli LC, et al. *Whole-genome sequence variation, population structure and demographic history of the Dutch population. Genome of the Netherlands Consortium*. **Nat Genet.** 2014 Aug;46(8):818-25. doi: 10.1038/ng.3021. Epub 2014 Jun 29.

Gudbjartsson D, et al. *Large-scale whole-genome sequencing of the Icelandic population.* **Nat Genet**. 2015 May;47(5):435-44. doi: 10.1038/ng.3247.

Holm H, et al. *A rare variant in MYH6 is associated with high risk of sick sinus syndrome.* **Nat Genet.** 2011 Mar 6;43(4):316-20. doi: 10.1038/ng.781.

Huang J, et al. *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel*. **Nature Commun.** 6, 8111 (2015).

Li Y, Willer C, Sanna S, Abecasis G. *Genotype imputation.* **Annu Rev Genomics Hum Genet.** 2009;10:387-406. doi: 10.1146/annurev.genom.9.081307.164242. Review.PMID:19715440 | PMCID:PMC2925172

Manolio TA, et al. *Finding the missing heritability of complex diseases.* **Nature**. 2009 Oct 8;461(7265):747-53. doi: 10.1038/nature08494.

Orrù V, et al. *Genetic variants regulating immune cell levels in health and disease.* **Cell**. 2013 Sep 26;155(1):242-56. doi:10.1016/j.cell.2013.08.041. PubMed PMID: 24074872. *(Chapter 4 in this thesis)*

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. 2015 Jul;23(7):975-83. doi: 10.1038/ejhg.2014.216. Epub 2014 Oct 8

Sidore C, et al. *Genome sequencing elucidates Sardinian genetic architecture and augments GWAS findings: the examples of lipids and blood inflammatory markers*. **Nature Genetics** 47, 1272–1281 (2015) doi:10.1038/ng.3368 Epub 15 Sept 2015

Sudmant PH, et al. *An integrated map of structural variation in 2,504 human genomes.* **Nature** http://dx.doi.org/10.1038/nature15394

The TG andHDL Working Group of the Exome Sequencing Project, NationalHeart, Lung, and Blood Institute. *Loss-of-function mutations in APOC3, triglycerides, and coronary disease*. **N. Engl. J. Med**. 371, 22–31 (2014).

Timpson NJ, et al. *A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans.* **Nature Commun**. 5, 4871 (2014)

UK10K Consortium, et al. *The UK10K project identifies rare variants in health and disease.* **Nature**. 2015 Oct 1;526(7571):82-90. doi: 10.1038/nature14962. Epub 2015 Sep 14.

van Leeuwen EM, et al. *Population-specific genotype imputations using minimac or IMPUTE2.* **Nat Protoc**. 2015 Sep;10(9):1285-96. doi: 10.1038/nprot.2015.077. Epub 2015 Jul 30.

Vrieze SI, et al. *In search of rare variants: preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes.* **Psychophysiology**. 2014 Dec;51(12):1309-20. doi: 10.1111/psyp.12350.

Wood, A.R. et al, *Defining the role of common variation in the genomic and biological architecture of adult human height*. **Nat Genet.** 2014 Nov;46(11):1173-86. doi: 10.1038/ng.3097.

Zoledziewska M, et al. *Height-reducing variants and selection for short stature in Sardinia.* **Nat Genet.** 47, 1352–1356 (2015) doi:10.1038/ng.3403 Epub 15 Sept 2015

Chapter 9: Conclusions and future prospects

The studies described in this thesis can be divided in two major parts. In the first part (Chapter 2 and 3), we showed the advantages of combining genotyping with whole-genome sequencing by genotype imputation in the isolate of Sardinia. We firstly investigated benefits of this approach for detecting traits-associated rare and low frequent variants in the worst case scenario: sequencing only the exons of already established loci in a small subset of individuals. Given the success of this proof-of-concept study, we carried out whole-genome sequencing of a larger subset of individuals and evaluated the benefit of using this population-specific reference set of haplotypes instead of the publicly available 1000 Genomes data. We showed that this approach can greatly enhance the genetic information content for genome-wide association studies (GWAS) in a cost-effective manner. We then assessed for association with several, clinically diverse traits all the 17 million discovered and imputed variants in up to 6,602 individuals. The results are illustrated in the second part of the thesis (Chapter 4 to 7). Our association analyses not only revealed novel loci for all the traits analyzed, but also highlighted several key messages. First, population isolates are an ideal setting to study rare variants. In fact, most of the novel signals identified in the SardiNIA cohort were more frequent in Sardinians than in other European populations; this demonstrates that isolation and genetic drift confer an intrinsic enhanced power for association of rare variants. We also showed that replication of signals, which was considered a standard step in GWAS of common variants, is not trivial when the risk allele frequency drops. In fact, when the variant is very rare outside the population being studied, very large sample sizes are required to provide sufficient power for replication. Alternatively, one could assess the findings in an independent, sufficiently large sample selected from the same population – but if this sample doesn't exists, one may have to recollect, genotype and phenotype samples for each phenotype of interest. This is impractical considering the cost and time required. Therefore other factors have to be considered to evaluate if findings are genuine, for example examining consistency of association at other variants according to linkage disequilibrium patterns, use of more stringent significance levels, evaluation of robustness of estimates with statistical resampling methods, or integration of functional data.

Genetic studies in isolates also present disadvantages: in fact extended linkage disequilibrium (LD) may limit the resolution of the association signal and make it difficult to identify the causal variant in associated loci. This phenomenon is even more relevant when the variant is rare, as the haplotype where the rare allele arose is usually young and underwent very few recombination events. We showed that association signals may encompass large regions, even 4-5Mb. In such cases, trans-ethnic analyses and integration of functional data is required to dissect the association curve, as the LD would not be disrupted within the same population unless the sample size can be increased dramatically. Isolated populations are also limited in the number of rare variants that can be assessed for association. In fact, variants that were not present in the initial pool of founder haplotypes or that were lost in subsequent generations are absent in present-day chromosomes.

We have described in details the results obtained in Sardinia when coupling whole-genome sequencing and genotyping data from the same population with genotype imputation. We then described in the last

Chapter similar ongoing efforts and illustrate results reported to date, including recent findings from the Genome of the Netherlands Project, the UK10K consortium and the deCode group. The finding clearly demonstrate that population-specific reference panel enhanche genetic discoveries also in open populations. We expect more discoveries as integration of sequencing data will be the norm in all existing GWAS cohorts and meta-analyses.

Within the SardiNIA study, several steps are currently ongoing or can be made to improve the efficiency of the panel. While more individuals are been sequenced to further expand the spectrum of rare variations assessable, variant calling for short insertions and deletions have been incorporated in the standard pipeline for sequencing analysis to be consequently imputed in the GWAS cohort and analyzed against the existing phenotypes. Published GWAS in the cohort were indeed strictly limited to SNPs detected in the Sardinian sequenced genomes and all other types of variants were assessed only by imputation with 1000 Genomes. Future improvements of the pipeline could incorporate algorithms to call other types of structural variants. Finally, owing to the family design of the sequenced samples, *de novo* mutations could, in principle, be identified. It has been shown that the rate of *de novo* mutations over the genome is highly dependent on father's age at conception and therefore it could be related to diseases such as schizophrenia and autisms [*Kong A, et al, Rate of de novo mutations and the importance of father's age to disease risk. Nature 2012*]. The impact of de-novo mutations in other common diseases and in complex traits variation is still largely unexplored and sequencing in large family studies can finally shed lights on it. In the Sardinian sequencing effort, the average coverage (4x) is lower than what is reported for other whole-genome sequencing studies that investigated *de novo* events, as the Genomes of the Netherlands Project (13x) or deCode (30x), therefore appropriate modifications have to be found to adapt the statistical modeling to the specific study design. For example, the homogeneity of the population and the presence of more structured families rather than trios (more generations or more sibs in a family) can bring in additional information on haplotype definition and can facilitate the discrimination of Mendelian inconsistencies from sequencing errors and *de novo* mutations.

The cost of whole-genome sequencing will probably drop in the next years until a point where sequencing will replace genotyping arrays. Meanwhile, genotype imputation will play a crucial role in genetic studies. In fact, with the need of more samples and the implementation of biobanks we are likely to scale sample size of quantitative traits GWAS to a factor of 100. Sooner than the drop in sequencing cost, several hundreds of thousands or even millions of samples will be available from the same country, genotyped with GWAS arrays and to be imputed with population-specific panels or large scale community resources for imputation (for example the Haplotype Reference Consortium). Collaborations across centers and research groups within and across countries will be essential for efficient use and implementation of biobanks.

Because not only the GWAS data sets but also reference panels are going to increase in size, special efforts are being made in improving current methodological approaches for imputation that cannot feasibly scale to very large samples. For example, in a preliminary overview of minimac version 4, computational time required to impute large chromosomes scales from days to minutes for panels of current size, making feasible its application to reference panels with 100,000s of samples [*Sayantan Das, ASHG 2015*, *PgmNr 1278: Minimac4: A next generation Imputation Tool for Mega Reference Panels.]*

Future improvements are expected not only from the genetic level. Advances in technologies have made it possible to measure at fine-scale many components of a human body, from count of specific immune cells to expression of genes in a variety of tissues, to small-molecule metabolite profiling, only to name a few. GWAS studies are therefore starting to assess biological mechanisms in addition to classical biomarkers.

From the phenotypic perspective, biobanks will provide a centralized database of many possibly measurable clinical and biochemical features of the same individual, and there is need to fully use those data. While most of the GWAS studies have so far searched for genetic association in a two dimensional space (one phenotype versus genotype), we envisage that more power in complex traits mapping will come from changing standardly used statistical methods to assess variations in a multidimensional system, and therefore looking for association of a genotype with multiple phenotypes at the same time, or multiple genotypes with one phenotype, or a combination of those.

Appendix

Chapters 2-7 are based on academic, peer-reviewed publications. Supplementary Information files for each chapther are available online, as indicated below.

Supplementary Files for Chapter 2 can be downloaded from:

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002198#s5

Supplementary Files for Chapter 3 can be downloaded from:

http://www.nature.com/ejhg/journal/v23/n7/suppinfo/ejhg2014216s1.html

Supplementary Files for Chapter 4 can be downloaded from:

http://www.sciencedirect.com/science/article/pii/S0092867413010726

Supplementary Files for Chapter 5 can be downloaded from:

http://www.nature.com/ng/journal/v47/n11/full/ng.3368.html#supplementary-information

Supplementary Files for Chapter 6 can be downloaded from:

http://www.nature.com/ng/journal/v47/n11/full/ng.3307.html#supplementary-information

Supplementary Files for Chapter 7 can be downloaded from:

http://www.nature.com/ng/journal/v47/n11/full/ng.3403.html#supplementary-information

Summary

Genome-wide association studies (GWASs) have initiated an era of gene discoveries characterized by findings that are robust and reproducible in independent studies. They were firstly implemented around ten years ago with methods and technology available at the time; technology and statistical methods have continued to improve by modeling needs and supporting challenges highlighted by the approach and results.

The first GWASs assessed ~100,000–2,000,000 common single nucleotide polymorphisms (SNPs) characterized with genotyping arrays or imputed with genotype imputation, an approach to statistically infer variants that are not directly genotyped. For this purpose, GWASs have intensively used data from the HapMap Project, which consisted on a reference set of 270 individuals characterized for ~2 million common polymorphisms (HapMap phase II). The use of HapMap in genotype imputation has been mostly replaced now with the 1000 Genomes reference set. This panel consists of ~2500 individuals for whom the full spectrum of variation in a genome (SNPs, indels or structural changes) have been characterized with next-generation sequencing machines. Currently, GWASs can assess millions of common and rare variants by direct whole-genome sequencing or, again, by genotype imputation. Notably, it was unclear until recently whether the genotype imputation approach would have been efficient to infer rare genotypes and therefore whether novel variants discovered with sequences in reference sets would have been effectively assessable. In fact, the genotype imputation approach is based on the idea that shared stretches of chromosomes can be found even in unrelated individuals, so that each persons' pair of haplotypes can be reconstructed as a mosaic of small pieces of haplotypes from a reference set. Therefore, the accuracy of haplotype reconstruction depends – among many parameters – on the frequency of the variant that is going to be predicted: the rarer is the variant, the harder is to predict it, because less similar haplotypes would be found. Consequently, the imputation accuracy for rare sites is expected to work well only with large reference panels (>1,000), which became available only recently. Accuracy will also be more precise if the population is highly homogeneous and the individuals in the reference set are genetically close to the population itself, because the haplotype where the rare variant lies can be selected from the panel with less ambiguity. The logical next step after early studies was therefore to create large population-specific reference panels to efficiently analyze rare sites in GWAS. Creating this set requires an investment in money, time and personnel, therefore it is important to evaluate the cost to benefit ratio before generating this massive amount of genomic data. In **Chapter 1**, we give a brief overview of the genotype imputation approach and how GWAS have used this tool to impute common and rare variations; we describe the limitations of the approach and provide a brief overview on current studies that have included a population-specific panel in their study design to efficiently assess rare variants. In **Chapter 2**, we described our work that aimed to evaluate the benefits of a population-specific reference panel in Sardinians. We carried out an experiment in a worst case scenario: we sequenced, by Sanger, a reduced number of individuals (256 samples) and focused on exons of a handful set of loci known to be associated with low-density lipoprotein cholesterol (LDL-C) levels. We assessed whether benefits in imputation using population-specific haplotypes could be translated in benefits in detecting traits-associated rare and low frequent variants even at known loci for a well-studied phenotype. The study revealed that genotype imputation performed with a moderate size reference panel could already be useful, in an isolate, to impute functionally interesting rare variants. Thereafter, we undertook large scale whole-genome

sequencing to create a Sardinian-specific reference panel for imputation that would maximize genetic information; in **Chapter 3** we evaluated its performance compared to other public resources. We showed that a population-specific panel confers a remarkably higher gain in accuracy in Sardinians, especially at low frequency and rare variants, compared to other studies with a similar design (number and coverage of sequenced samples, scaffold of baseline genotypes and number of samples to be imputed). We also showed that in the Sardinian isolate and with the specific reference panel, very accurate genome reconstruction can be made even from sparser genetic maps, such as MetaboChip and Human Exome arrays.

While the primary quantifiable outcome of genotype imputation is accuracy, the efficiency of the approach is concrete when the application results in enhanced gene discoveries. In the past years, there have been a few examples of genetic studies that were empowered by the use of a sequencing-based population-specific reference panel for imputation, which highlighted variants that have changed in frequency in the population being studied and that would not have been well imputed with generic reference panels. Chapters 4-7 include four of those studies, where GWAS carried out in the SardiNIA cohort were enhanced by the Sardinian-specific reference panel created with whole-genome sequencing of up to 2,120 Sardinians (Chapters 4-7). In **Chapter 4**, in a GWAS that used a combination of MetaboChip and ImmunoChip SNP arrays as a framework for imputation, we identified 23 independent variants of which 20 were imputed. This was striking because the two arrays led to very poor imputation quality when using other European reference panels. In **Chapter 5**, a parallel comparison of a series of GWAS performed using either the Sardinian reference panel or the latest release of 1000Genomes (phase 3) revealed four novel loci for lipid levels and blood inflammatory markers that would have not being identified without the Sardinian sequencing panel. Similarly, in **Chapter 6** and **7**, we identified novel loci associated with hemoglobin subtypes levels and height variation by means of this specific imputation reference set. Altogether, those studies indicate that isolates are an ideal setting to study rare variants in GWAS when an appropriate design is used. In fact, most of the novel signals we identified were more frequent in Sardinians than in other European populations, demonstrating that isolation and genetic drift confer an intrinsic enhanced power for association of rare variants. Still, this advantage can be lost if a population-specific reference panel is not used for imputation because many of the detected loci were not detectable with public reference sets.

Genetic drift may increase frequency of certain variants not only in isolates but also in small geographical area of open-populations. Therefore, population-specific reference panels could lead to better results compared to, or in combination with, the 1000 Genomes data in non-isolates. In **Chapter 8** we described current ongoing efforts in other isolates and in open populations that have set up their own reference panel by whole-genome sequencing a subset of individuals, including the Genome of the Netherlands Project, the UK10K Consortium and the deCode group. Novel associations at rare and less frequent variants were detected for complex traits and diseases by those studies; a few times the variant was specific of the population being studied, as observed in Sardinia. This final observation came with a consequence. Replication of association findings in independent cohorts, considered a standard step in the era of common variants GWAS, is not always possible when the frequency drops. Other factors have to be considered to evaluate if findings are genuine, for example examining consistency of association at other variants according to linkage disequilibrium patterns, use of more stringent significance levels, evaluation of robustness of estimates with statistical resampling methods, or integration of functional data.

The cost of whole-genome sequencing will probably drop in the next years until a point where sequencing will replace genotyping arrays. Meanwhile, genotype imputation will play a crucial role in genetic studies.

Sample size of both GWAS studies and reference panels are going to increase; changes in the algorithm underlying the genotype imputation are being proposed to allow the approach to scale with the complexity of the data. In **Chapter 9,** we discuss future prospects of GWAS and their evolution in the near future.

Samenvatting

(translation service provided by Elsevier's Webshop)

Genoom-breed associatieonderzoek (GWAS) initieerde een tijdperk van genontdekkingen, gekarakteriseerd door robuuste en reproduceerbare bevindingen in onafhankelijke onderzoeken. Deze werden rond tien jaar geleden voor het eerst geïmplementeerd met de destijds beschikbare methoden en technologieën. De technologieën en de statistische methoden ontwikkelden zich verder door modellering van de behoeften en het ondersteunen van de uitdagingen die de aanpak en de resultaten opleverde.

De eerste GWAS beoordeelden ~100.000-2.000.000 vaak voorkomende, enkele nucleotide polymorfismen (single nucleotide polymorphisms, SNP's), gekarakteriseerd door genotypische scala's of genotype-toerekening, een aanpak om statistisch varianten af te leiden die niet direct gegenotypeerd zijn. Met dit doel maakten GWAS'en intensief gebruik van gegevens uit het HapMap Project, dat bestond uit een referentieset van 270 personen die gekarakteriseerd werden voor ~2 miljoen vaak voorkomende polymorfismen (HapMap-fase II). Het gebruik van HapMap bij genotype-toerekening is nu veelal vervangen door de 1000 Genomes-referentieset. Het panel bestaat uit ~2.500 personen, waarvan het gehele variatiespectrum in een genoom (SNP's, indels en structurele veranderingen) gekarakteriseerd is met sequentiemachines van de volgende generatie. Op dit moment kunnen GWAS'en miljoenen vaak en zelden voorkomende varianten beoordelen door rechtstreekse sequentie van het gehele genoom of, opnieuw, door genotype-toerekening. Het is opmerkelijk dat het tot voor kort onduidelijk was of de aanpak van genotype-toerekening om zelden voorkomende genotypen af te leiden werkzaam zou zijn, en dus of nieuw ontdekte varianten met sequenties in referentiesets effectief te beoordelen zouden zijn. De aanpak van genotype-toerekening is daadwerkelijk gebaseerd op het idee dat gedeelde chromosoomreeksen zelfs in niet-gerelateerde personen gevonden kunnen worden, zodat het haplotypenpaar van elke persoon gereconstrueerd kan worden tot een mozaïek van kleine stukjes haplotypen uit een referentieset. Daarom is de nauwkeurigheid van haplotype-reconstructie (naast vele andere parameters) afhankelijk van de frequentie van de variant die zal worden voorspeld: hoe meer zelden de variant, hoe moeilijker de voorspelling, omdat er minder vergelijkbare haplotypen zouden worden gevonden. Als gevolg daarvan wordt verwacht dat de nauwkeurigheid van de toerekening aan zelden voorkomende varianten alleen goed zal werken met grote referentiepanels (>1.000), die pas kortgeleden beschikbaar werden. De nauwkeurigheid zal ook toenemen als de populatie grotendeels homogeen is en de personen in de referentieset genetisch dichtbij de populatie zelf liggen, omdat het haplotype met de zelden voorkomende variant met minder dubbelzinnigheid geselecteerd kan worden uit het panel. De logische vervolgstap na de vroege onderzoeken was dan ook de samenstelling van grote populatiespecifieke referentiepanels om zelden voorkomende varianten werkzaam te analyseren met GWAS. Het creëren van deze set vereist een investering in geld, tijd en personeel en het is dan ook belangrijk om een kosten-batenanalyse uit te voeren alvorens deze enorme hoeveelheid genomische gegevens te genereren. In **Hoofdstuk 1** geven we een kort overzicht van de aanpak van genotype-toerekening en de wijze waarop GWAS dit hulpmiddel gebruikte bij het toerekenen van vaak en zelden voorkomende variaties. We beschrijven de beperkingen van de aanpak en bieden een kort overzicht van huidige onderzoeken die in hun onderzoeksontwerp een populatiespecifiek panel gebruiken om zelden voorkomende varianten werkzaam te beoordelen. In **Hoofdstuk 2** beschrijven we ons werk dat gericht was op het evalueren van de voordelen van een populatiespecifiek referentiepanel bij Sardijnen. We voerden een experiment uit met een worst case scenario: we sequentieerden, naar Sanger, een gereduceerd aantal personen (256 monsters) en richtten

ons op exonen van een handvol sets met loci die gekend geassocieerd waren aan low-density-lipoproteïnecholesterol (LDL-C). We beoordeelden of voordelen van toerekening door middel van populatiespecifieke haplotypen vertaald konden worden in voordelen voor het detecteren van factorgeassocieerde, zelden voorkomende en laagfrequente varianten, zelfs in gekende loci, voor een goed onderzocht fenotype. Het onderzoek toonde aan dat genotype-toerekening in een referentiepanel van gemiddelde omvang al nuttig kan zijn, in een isolaat, voor het toerekenen van functioneel interessante, zelden voorkomende varianten. Daarna ondernamen we grootschalige sequentie van complete genomen om een Sardijns-specifiek referentiepanel voor toerekening samen te stellen, dat genetische informatie zou maximaliseren. In **Hoofdstuk 3** beoordeelden we de prestaties in vergelijking met andere openbare bronnen. We toonden aan dat een populatiespecifiek panel een aanzienlijk grotere winst in nauwkeurigheid oplevert bij Sardijnen, vooral bij lage frequentie- en zelden voorkomende varianten, in vergelijking met andere onderzoeken met een vergelijkbaar ontwerp (aantal en dekking van gesequentieerde monsters, staffels van genotypen aan de baseline en aantal toe te rekenen monsters). We toonden ook aan dat in het Sardijnse isolaat en met het specifieke referentiepanel zeer nauwkeurige genoomreconstructie kan worden uitgevoerd, zelfs op basis van spaarzamere genetische kaarten als MetaboChip- en Human Exome-scala's.

Hoewel het primair kwantificeerbare resultaat van genotype-toerekening de nauwkeurigheid is, is de werkzaamheid van de aanpak concreet als de toepassing tot genetische ontdekkingen leidt. In de afgelopen jaren waren er een aantal voorbeelden van genetische onderzoeken die werden ondersteund door het gebruik van een sequentiegebaseerd, populatiespecifiek referentiepanel voor toerekening en varianten aantoonden die in frequentie wijzigden in de onderzochte populatie en niet goed toegerekend zouden zijn bij generische referentiepanels. De hoofdstukken 4 t/m 7 bevatten vier van deze onderzoeken, waarbij het in het Sardiniëcohort uitgevoerde GWAS versterkt werd door het Sardijns-specifieke referentiepanel, gecreëerd met sequentie van complete genomen van tot 2.120 Sardijnen (Hoofdstuk 4-7). In **Hoofdstuk 4** identificeerden we 23 onafhankelijke varianten in een GWAS die een combinatie van MetaboChip en ImmunoChip SNP-scala's gebruikte als toerekeningskader, waarvan 20 werden toegerekend. Dit was opvallend, omdat de twee scala's tot zeer matige toerekeningskwaliteit leidden in andere Europese referentiepanels. In **Hoofdstuk 5** toonde een parallelvergelijking van een reeks uitgevoerde GWAS in ofwel het Sardijnse referentiepanel, ofwel de laatste uitgave van 1000 Genomes (fase 3) vier nieuwe loci aan voor lipideniveaus en inflammatoire bloedmarkers, die niet geïdentificeerd zouden zijn zonder het Sardijnse sequentiepanel. Op vergelijkbare wijze identificeerden we in **Hoofdstuk 6** en **7** nieuwe loci geassocieerd aan hemaglobinesubtypeniveaus en lengtevariatie in deze specifiek toegerekende referentieset. Samen tonen deze onderzoeken aan dat isolaten een ideale setting vormen voor onderzoek naar zelden voorkomende varianten in GWAS, bij gebruik van een passend ontwerp. De meeste nieuwe signalen die we identificeerden, kwamen daadwerkelijk frequenter voor bij Sardijnen dan in andere Europese populaties, wat aantoont dat isolatie en genetische drift een intrinsiek versterkt onderscheidend vermogen oplevert voor associatie van zelden voorkomende varianten. Dit voordeel kan echter teniet worden gedaan als er geen populatiespecifiek referentiepanel wordt gebruikt voor toerekening, omdat vele van de gedetecteerde loci niet detecteerbaar waren met openbare referentiesets.

Door genetische drift kan de frequentie van bepaalde varianten niet alleen in isolaten toenemen, maar ook in kleine geografische gebieden met open populaties. Daarom kunnen populatiespecifieke referentiepanels bij non-isolaten in vergelijking of in combinatie met de 1000 Genomes-gegevens tot betere resultaten leiden. In **Hoofdstuk 8** beschreven we de inspanningen die op dit moment gaande zijn in andere isolaten en in open populaties die een eigen referentiepanel creëerden door sequentie van complete genomen in een

personensubgroep, waaronder het Genoom van Nederland, het UK10K Consortium en de deCode-groep. In die onderzoeken werden bij zelden voorkomende en laagfrequente varianten nieuwe associaties voor complexe factoren en aandoeningen gedetecteerd. Een aantal keren was de variant specifiek voor de onderzochte populatie, zoals geobserveerd in Sardinië. Deze laatste observatie leverde een consequentie op. Replicatie van geassocieerde bevindingen in onafhankelijke cohorten, beschouwd als een normale stap in het tijdperk van GWAS bij vaak voorkomende varianten, is niet altijd mogelijk als de frequentie afneemt. Er dienen andere factoren in ogenschouw te worden genomen om te evalueren of bevindingen echt zijn, bijvoorbeeld het onderzoeken van de consistentie van associatie met andere varianten met betrekking tot koppeling van disequilibriumpatronen, het gebruik van stringentere significantieniveaus, evaluatie van robuustheid van schattingen met statistische herbemonsteringsmethoden of integratie van functionele gegevens.

De kosten van sequentie van complete genomen zullen in de komende jaren wellicht afnemen tot een punt waarop sequentie de plek inneemt van genotypische scala's. In de tussentijd zal genotype-toerekening een cruciale rol spelen in genetisch onderzoek. De steekproefgrootte van zowel GWAS als referentiepanels zal toenemen; er worden veranderingen voorgesteld voor het onderliggende algoritme voor genotype-toerekening om de aanpak voor schaalbaarheid van de complexiteit van gegevens mogelijk te maken. In **Hoofdstuk 9** bespreken we toekomstmogelijkheden van GWAS en de evolutie ervan in de nabije toekomst.

Acknowledgements

Foremost, I would like to express my sincere gratitude to Prof. Cisca Wijmenga for helping me to achieve this goal, with guidance, continuous support and bright suggestions. I would like to thank also the rest of my thesis committee: Prof Lude Frank, Prof. C.M. van Duijn, Prof. P. van der Harst, and Prof. H. Snieder for their insightful comments.

I am also grateful to many individuals without whom this dissertation would have never been possible:

Prof. Giuseppe Pilia and David Schlessinger for their brilliant idea of initiating the SardiNIA project and later offering me the opportunity to work on it;

Prof. Goncalo Abeçasis, for his patience during my early years work in research, for being an inspiration for motivation, enthusiasm, creativity and humility;

Prof. Antonio Cao for his scientific support, suggestions and sincere criticisms – I miss you Prof.;

Manuela Uda and Prof. Francesco Cucca for their successful leadership of the SardiNIA study after prof. Pilia passed away;

Nicole Soranzo, a colleague and friend, for encouraging me to pursue my studies further with a PhD;

The many colleagues and collaborators, in Sardinia and around the world, who have contributed to my scientific progress, sharing days and nights of work; special thanks for the studies described in this thesis to Gonçalo Abecasis, Fabio Busonero, Charleston Chiang, Francesco Cucca, Fabrice Danjou, Edoardo Fiorillo, Bingshan Li, Andrea Maschio, Antonella Mulas, David Schlessinger, Silvia Naitza, John Novembre, Valeria Orrú, Eleonora Porcu, Giorgio Pistis, Carlo Sidore, Maristella Steri and Magdalena Zoledziewska;

My parents for supporting my wish to work in research after my bachelor;

Last but not the least, my husband, a loving, patient, supporting and always-present companion in life.

Serena

Short Biography

Serena Sanna

Serena Sanna was born on 1980 in Sardinia, Italy. After secondary school, she obtained her Master of Science (Laurea) *cum laude* in Mathematics at the University of Cagliari, Italy, on 2003. After one year fellowship at the Italian National Research Council (CNR), she was introduced to genetics and decided to join the Center for Statistical Genetics at the University of Michigan, where she learned innovative statistical genetics methods and developed a strong attitude to research. She then returned to Sardinia on 2007 and continued to use her statistical genetics skills for the SardiNIA study, a Genetic and Epidemiological study on aging-associated conditions, a project conceived by Prof. Giuseppe Pilia. Her main research focus was to study the genetics of quantitative traits and complex diseases in Sardinians, through genome-wide association scan approach. She was also involved in several international consortia where, in some, she played leadership roles in coordinating meta-analysis efforts. She has dedicated particular attention to use and evaluate tools to enhance discoveries as imputation of untyped markers in both families and unrelated samples, using both array-based and sequencing-based reference panels.

From 2011, she is a Permanent Researcher at the CNR Institute. She has a strong record of publications and also received several awards, including the 2008 ESHG Young Scientists award for outstanding science. She continues to be a co-investigator of the SardiNIA project and involved in a large scale sequencing effort (>3,500 Sardinians samples at 4x coverage), with the aim to further enhance genetic discoveries for complex traits and diseases in this population.

Full List of Publications

A list of all publications is given, following chronological order. Please note that * indicates equal contributions.

2015

1. Winkler TW, Justice AE, Graff M, Barata L, et al. *The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study.* **PLoS Genet.** 2015 Oct 1;11(10):e1005378. doi: 10.1371/journal.pgen.1005378. eCollection 2015 Oct. PMID: 26426971

2. Day FR, Ruth KS, Thompson DJ, Lunetta KL, et al. *Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair.* **Nat Genet.** 2015 Sep 28. doi: 10.1038/ng.3412

3. Zoledziewska M*, Sidore C*, Chiang CWK*, ***Sanna S****, Mulas A, Steri M, Busonero F, Marcus JH, Marongiu M, Maschio A, Ortega Del Vecchyo D, Floris M, Meloni A, Delitala A, Concas MP, et al. *Major height reducing variants and selection for short stature on the island of Sardinia.* ***Nature*** **Genetics**. 47, 1352–1356 (2015) doi:10.1038/ng.3403 Epub 15 Sept 2015

4. Danjou F*, Zoledziewska M*, Sidore C , Steri M, Busonero F , Maschio A, Mulas A, Perseu L, Barella S, Porcu E, Pistis G, Pitzalis M, Mauro Pala M, Menzel M, Metrustry S, Spector TD, Leoni L, Angius A, Uda M, Moi P, Thein SL, Galanello R, Abecasis GR, Schlessinger D, ***Sanna S****, Cucca F*. *Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels.* **Nature Genetics**. 47, 1264–1271 (2015) doi:10.1038/ng.3307 Epub 15 Sept 2015

5. Sidore C*, Busonero F*, Maschio A*, Porcu E*, Naitza S*, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F, Kwong A, Ortega del Vecchyo VD, Chiang CWK, Bragg-Gresham J, Pitzalis M, Nagaraja R, Tarrier B, Brennan C, Uzzau S, Fuchsberger C, Atzeni R, Reinier F, Berutti R, Huang J, Timpson NJ, Toniolo D, Gasparini P, Malerba G, Dedoussis G, Zeggini E, Soranzo N, Jones C, Lyons R, Angius A, Kang HM, Novembre J, ***Sanna S****, Schlessinger D*, Cucca F*, Abecasis GR*. *Genome sequencing elucidates Sardinian genetic architecture and augments GWAS findings: the examples of lipids and blood inflammatory markers*. **Nature Genetics**. 47, 1272–1281 (2015) doi:10.1038/ng.3368 Epub 15 Sept 2015

6. Barizzone N, Zara I, Sorosina M, Lupoli S, Porcu E, Pitzalis M, Zoledziewska M, Esposito F, Leone M, Mulas A, Cocco E, Ferrigno P, Guerini FR, Brambilla P, Farina G, Murru R, Deidda F, Sanna Sonia, Loi A, Barlassina C, Vecchio D, Zauli A, Clarelli F, Braga D, Poddie F, Cantello R, Martinelli V, Comi G, Frau J, Lorefice L, Pugliatti M, Rosati G, the PROGEMUS Group, the PROGRESSO Group, Melis M, Marrosu MG, Cusi D, Cucca F*, Martinelli Boneschi F*, ***Sanna S****,* D'Alfonso S*. *The burden of Multiple Sclerosis variants in continental Italians and Sardinians.* **Mult Scler** *Epub 5 October 2015*

7. Reinier F, Zoledziewska M, Hanna D, Smith JD, Valentini M, Zara I, Berutti R, ***Sanna S***, Oppo M, Cusano R, Satta R, Montesu MA, Jones C, Cerimele D, Nickerson DA, Angius A, Cucca F, Cottoni F, Crisponi L. *Mandibular hypoplasia, deafness, progeroid features and lipodystrophy (MDPL) syndrome in the context of inherited lipodystrophies.* Metabolism. 2015 Nov;64(11):1530-40. doi: 10.1016/j.metabol.2015.07.022. Epub 2015 Aug 1. Review.

8. Predazzi IM*, Sobota RS*, ***Sanna S****, Bush WS, Bartlett J, Lilley JS, Linton MF, Schlessinger D, Cucca F, Fazio S, Williams SM. *Gender-specific Parental Effects on Offspring Lipid Levels.* **J Am Heart Assoc.** 2015 Jun 30;4(7). pii: e001951. doi: 10.1161/JAHA.115.001951.

9. Palomba G, Loi A, Porcu E, Cossu A, Zara I, Budroni M, Dei M, Lai S, Mulas A, Olmeo N, Ionta MT, Atzori F, Cuccuru G, Pitzalis M, Zoledziewska M, Olla N, Lovicu M, Pisano M, Abecasis GR, Uda M, Tanda F, Michailidou K, et al. *Genome-wide association study of susceptibility loci for breast cancer in Sardinian population.* **BMC Cancer**. 2015 May 10;15(1):383. doi: 10.1186/s12885-015-1392-9.PMID:25956309 | PMCID:PMC4434540

10. Taylor PN, Porcu E, Chew S, Campbell PJ, Traglia M, Brown SJ, Mullin BH, Shihab HA, Min J, Walter K, Memari Y, Huang J, Barnes MR, Beilby JP, Charoen P, Danecek P, Dudbridge F, Forgetta V, Greenwood C, Grundberg E, Johnson AD, Hui J, et al. *Whole-genome sequence-based analysis of thyroid function.* **Nat Commun.** 2015 Mar 6;6:5681. doi: 10.1038/ncomms6681.PMID:25743335 | PMCID:PMC4366514

11. Feng S, Pistis G, Zhang H, Zawistowski M, Mulas A, Zoledziewska M, Holmen OL, Busonero F, ***Sanna S***, Hveem K, Willer C, Cucca F, Liu DJ, Abecasis GR. *Methods for association analysis and meta-analysis of rare variants in families.* **Genet Epidemiol**. 2015 May;39(4):227-38. doi: 10.1002/gepi.21892. Epub 2015 Mar 4.PMID:25740221 | PMCID:PMC4459524

12. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, Croteau-Chonka DC, Esko T, Fall T, Ferreira T, Gustafsson S, Kutalik Z, Luan J, MÃngi R, Randall JC, Winkler TW, Wood AR, Workalemahu T, et al. *Genetic studies of body mass index yield new insights for obesity biology*. **Nature**. 2015 Feb 12;518(7538):197-206. doi: 10.1038/nature14177.PMID:25673413 | PMCID:PMC4382211

13. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, MÃngi R, Strawbridge RJ, Pers TH, Fischer K, Justice AE, Workalemahu T, Wu JM, Buchkovich ML, Heard-Costa NL, Roman TS, Drong AW, Song C, Gustafsson S, Day FR, Esko T, Fall T, Kutalik Z, et al. *New genetic loci link adipose and insulin biology to body fat distribution.* **Nature**. 2015 Feb 12;518(7538):187-96. doi: 10.1038/nature14132.PMID:25673412 | PMCID:PMC4338562

2014

14. Baumert J, Huang J, McKnight B, Sabater-Lleal M, Steri M, Chu AY, Trompet S, Lopez LM, Fornage M, Teumer A, Tang W, Rudnicka AR, MÃ¤rz W, Hottenga JJ, Kavousi M, Lahti J, Tanaka T, Hayward C, Huffman JE, Morange PE, Rose LM, Basu S, et al. *No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects.* **PLoS One.** 2014;9(12):e111156. doi: 10.1371/journal.pone.0111156.PMID:25551457 | PMCID:PMC4281156

15. Benyamin B, Esko T, Ried JS, Radhakrishnan A, Vermeulen SH, Traglia M, GÃgele M, Anderson D, Broer L, Podmore C, Luan J, Kutalik Z, ***Sanna S***, van der Meer P, Tanaka T, Wang F, Westra HJ, Franke L, Mihailov E, Milani L, HÃ¤ldin J, Winkelmann J, et al. *Novel loci affecting iron homeostasis and their effects in individuals at risk for hemochromatosis.* **Nat Commun**. 2014 Oct 29;5:4926. doi: 10.1038/ncomms5926.PMID:25352340 | PMCID:PMC4215164

16. Kus A, Szymanski K, Peeters RP, Miskiewicz P, Porcu E, Pistis G, ***Sanna S***, Naitza S, Ploski R, Medici M, Bednarczuk T. *The association of thyroid peroxidase antibody risk loci with susceptibility to and phenotype of Graves' disease.* **Clin Endocrinol** (Oxf). 2014 Oct 24. doi: 10.1111/cen.12640. [Epub ahead of print]PMID:25345847

17. Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A, Brennan C, Lai S, Miller MB, Marcelli M, Urru MF, Pitzalis M, Lyons RH, Kang HM, Jones CM, Angius A, Iacono WG, Schlessinger D, McGue M, Cucca F*, Abecasis GR*, ***Sanna S*.*** *Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs.* **Eur J Hum Genet.** 2015 Jul;23(7):975-83. doi: 10.1038/ejhg.2014.216. Epub 2014 Oct 8.PMID:25293720

18. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, Amin N, Buchkovich ML, Croteau-Chonka DC, Day FR, Duan Y, Fall T, Fehrmann R, Ferreira T, Jackson AU, Karjalainen J, Lo KS, Locke AE, et al. *Defining the role of common variation in the genomic and biological architecture of adult human height.* **Nat Genet**. 2014 Nov;46(11):1173-86. doi: 10.1038/ng.3097. Epub 2014 Oct 5.PMID:25282103 | PMCID:PMC4250049

19. Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, He C, Chasman DI, Esko T, Thorleifsson G, Albrecht E, Ang WQ, Corre T, Cousminer DL, Feenstra B, Franceschini N, Ganna A, Johnson AD, Kjellqvist S, Lunetta KL, McMahon G, Nolte IM, et al. *Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche*. **Nature**. 2014 Oct 2;514(7520):92-7. doi: 10.1038/nature13545. Epub 2014 Jul 23.PMID:25231870 | PMCID:PMC4185210

20. Simino J, Shi G, Bis JC, Chasman DI, Ehret GB, Gu X, Guo X, Hwang SJ, Sijbrands E, Smith AV, Verwoert GC, Bragg-Gresham JL, Cadby G, Chen P, Cheng CY, Corre T, de Boer RA, Goel A, Johnson T, Khor CC; LifeLines Cohort Study, Lluis-Ganella C, et al. *Gene-age interactions in blood pressure regulation: a large-scale investigation with the CHARGE, Global BPgen, and ICBP Consortia.* **Am J Hum Genet.** 2014 Jul 3;95(1):24-38. doi: 10.1016/j.ajhg.2014.05.010. Epub 2014 Jun 19.PMID:24954895 | PMCID:PMC4085636

21. Arking DE, Pulit SL, Crotti L, van der Harst P, Munroe PB, Koopmann TT, Sotoodehnia N, Rossin EJ, Morley M, Wang X, Johnson AD, Lundby A, Gudbjartsson DF, Noseworthy PA, Eijgelsheim M, Bradford Y, Tarasov KV, Dorr M, Muller-Nurasyid M, Lahtinen AM, Nolte IM, Smith AV, et al. *Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization*. **Nat Genet**. 2014 Aug;46(8):826-36. doi: 10.1038/ng.3014. Epub 2014 Jun 22.PMID:24952745 | PMCID:PMC4124521

22. Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sa¡nchez-Quinto F, Zara I, Pitzalis M, Sidore C, Busonero F, Maschio A, Angius A, Jones C, Mendoza-Revilla J, Nekhrizov G, Dimitrova D, Theodossiev N, Harkins TT, Keller A, Maixner F, Zink A, Abecasis G, et al. *Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe.* **PLoS Genet**. 2014 May;10(5):e1004353. doi: 10.1371/journal.pgen.1004353.PMID:24809476 | PMCID:PMC4014435

23. Medici M*, Porcu E*, Pistis G*, Teumer A, Brown SJ, Jensen RA, Rawal R, Roef GL, Plantinga TS, Vermeulen SH, Lahti J, Simmonds MJ, Husemoen LL, Freathy RM, Shields BM, Pietzner D, Nagy R, Broer L, Chaker L, Korevaar TI, Plia MG, Sala C, ……. Cappola A, Toniolo D, ***Sanna S****, Naitza S*, Peeters RP*. *Identification of novel genetic Loci associated with thyroid peroxidase antibodies and clinical thyroid disease.* **PLoS Genet**. 2014 Feb;10(2):e1004123. doi: 10.1371/journal.pgen.1004123.PMID:24586183 | PMCID:PMC3937134

2013

24. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang HY, Demirkan A, Den Hertog HM, Do R, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, et al. *Discovery and refinement of loci associated with lipid levels.* **Nat Genet.** 2013 Nov;45(11):1274-83. doi: 10.1038/ng.2797. Epub 2013 Oct 6.PMID:24097068 | PMCID:PMC3838666

25. Do R, Willer CJ, Schmidt EM, Sengupta S, Gao C, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, Beckmann JS, Bragg-Gresham JL, Chang HY, Demirkan A, Den Hertog HM, Donnelly LA, Ehret GB, Esko T, Feitosa MF, Ferreira T, et al. *Common variants associated with plasma triglycerides and risk for coronary artery disease*. **Nat Genet**. 2013 Nov;45(11):1345-52. doi: 10.1038/ng.2795. Epub 2013 Oct 6.PMID:24097064 | PMCID:PMC3904346

26. Orru V, Steri M, Sole G, Sidore C, Virdis F, Dei M, Lai S, Zoledziewska M, Busonero F, Mulas A, Floris M, Mentzen WI, Urru SA, Olla S, Marongiu M, Piras MG, Lobina M, Maschio A, Pitzalis M, Urru MF, Marcelli M, Cusano R, Deidda F, Serra V, Oppo M, Pilu R, Reinier F, Berutti R, Pireddu L, Zara I, Porcu E, Kwong A, Brennan C, Tarrier B, Lyons R, Kang HM, Uzzau S, Atzeni R, Valentini M, Firinu D, Leoni L, Rotta G, Naitza S, Angius A, Congia M, Whalen MB, Jones CM, Schlessinger D, Abecasis GR, Fiorillo E*, ***Sanna S****, Cucca F.* *Genetic variants regulating immune cell levels in health and disease*. **Cell**. 2013 Sep 26;155(1):242-56. doi: 10.1016/j.cell.2013.08.041.PMID:24074872

27. Sabater-Lleal M, Huang J, Chasman D, Naitza S, Dehghan A, Johnson AD, Teumer A, Reiner AP, Folkersen L, Basu S, Rudnicka AR, Trompet S, Milarstig A, Baumert J, Bis JC, Guo X, Hottenga JJ, Shin SY, Lopez LM, Lahti J, Tanaka T, Yanek LR, et al. *Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease.* **Circulation**. 2013 Sep 17;128(12):1310-24. doi: 10.1161/CIRCULATIONAHA.113.002251. Epub 2013 Aug 22.PMID:23969696 | PMCID:PMC3842025

28. Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, Sanna D, Useli A, Urru MF, Marcelli M, Cusano R, Oppo M, Zoledziewska M, Pitzalis M, Deidda F, Porcu E, Poddie F, Kang HM, et al. *Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny*. **Science**. 2013 Aug 2;341(6145):565-9. doi: 10.1126/science.1237947.PMID:23908240

29. Porcu E, ***Sanna S***, Fuchsberger C, Fritsche LG. *Genotype imputation in genome-wide association studies.* **Curr Protoc Hum Genet.** 2013 Jul;Chapter 1:Unit 1.25. doi: 10.1002/0471142905.hg0125s78. Review.PMID:23853078

30. Graff M, Ngwa JS, Workalemahu T, Homuth G, Schipf S, Teumer A, Volzke H, Wallaschofski H, Abecasis GR, Edward L, Francesco C, ***Sanna S***, Scheet P, Schlessinger D, Sidore C, Xiao X, Wang Z, Chanock SJ, Jacobs KB, Hayes RB, Hu F, Van Dam RM; et al. *Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course.* **Hum Mol Genet.** 2013 Sep 1;22(17):3597-607. doi: 10.1093/hmg/ddt205. Epub 2013 May 12.PMID:23669352 | PMCID:PMC3736869

31. den Hoed M, Eijgelsheim M, Esko T, Brundel BJ, Peal DS, Evans DM, Nolte IM, Segri AV, Holm H, Handsaker RE, Westra HJ, Johnson T, Isaacs A, Yang J, Lundby A, Zhao JH, Kim YJ, Go MJ, Almgren P, Bochud M, Boucher G, Cornelis MC, et al. *Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders.* **Nat Genet**. 2013 Jun;45(6):621-31. doi: 10.1038/ng.2610. Epub 2013 Apr 14.PMID:23583979 | PMCID:PMC3696959

32. Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR, Esko T, Fall T, Ferreira T, Gentilini D, Jackson AU, Luan J, Randall JC, Vedantam S, Willer CJ, Winkler TW, Wood AR, Workalemahu T, et al. *Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture.* **Nat Genet.** 2013 May;45(5):501-12. doi: 10.1038/ng.2606. Epub 2013 Apr 7.PMID:23563607 | PMCID:PMC3973018

33. Porcu E, Medici M, Pistis G, Volpato CB, Wilson SG, Cappola AR, Bos SD, Deelen J, den Heijer M, Freathy RM, Lahti J, Liu C, Lopez LM, Nolte IM, O'Connell JR, Tanaka T, Trompet S, Arnold A, Bandinelli S, Beekman M, Böhringer S, Brown SJ, Buckley BM, Camaschella C, de Craen AJ, Davies G, de Visser MC, Ford I, Forsen T, Frayling TM, Fugazzola L, Gögele M, Hattersley AT, Hermus AR, Hofman A, Houwing-Duistermaat JJ, Jensen RA, Kajantie E, Kloppenburg M, Lim EM, Masciullo C, Mariotti S, Minelli C, Mitchell BD, Nagaraja R, Netea-Maier RT, Palotie A, Persani L, Piras MG, Psaty BM, Räikkönen K, Richards JB, Rivadeneira F, Sala C, Sabra MM, Sattar N, Shields BM, Soranzo N, Starr JM, Stott DJ, Sweep FC, Usala G, van der Klauw MM, van Heemst D, van Mullem A, Vermeulen SH, Visser WE, Walsh JP, Westendorp RG, Widen E, Zhai G, Cucca F, Deary IJ, Eriksson JG, Ferrucci L, Fox CS, Jukema JW, Kiemeney LA, Pramstaller PP, Schlessinger D, Shuldiner AR, Slagboom EP, Uitterlinden AG, Vaidya B, Visser TJ, Wolffenbuttel BH, Meulenbelt I, Rotter JI, Spector TD, Hicks AA, Toniolo D, ***Sanna S****,* Naitza S*, Peeters RP*. *A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function*. **PLoS Genet.** 2013;9(2):e1003266. doi: 10.1371/journal.pgen.1003266. Epub 2013 Feb 7.PMID:23408906 | PMCID:PMC3567175

34. Kottgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, Pistis G, Ruggiero D, O'Seaghdha CM, Haller T, Yang Q, Tanaka T, Johnson AD, Kutalik Z, Smith AV, Shi J, Struchalin M, Middelberg RP, Brown MJ, Gaffo AL, Pirastu N, Li G, et al. *Genome-wide association analyses identify 18 new loci associated with serum urate concentrations.* **Nat Genet.** 2013 Feb;45(2):145-54. doi: 10.1038/ng.2500. Epub 2012 Dec 23.PMID:23263486 | PMCID:PMC3663712

2012

35. van der Harst P, Zhang W, Mateo Leach I, Rendon A, Verweij N, Sehmi J, Paul DS, Elling U, Allayee H, Li X, Radhakrishnan A, Tan ST, Voss K, Weichenberger CX, Albers CA, Al-Hussani A, Asselbergs FW, Ciullo M, Danjou F, Dina C, Esko T, Evans DM, et al. *Seventy-five genetic loci influencing the human red blood cell.* **Nature**. 2012 Dec 20;492(7429):369-75. doi: 10.1038/nature11677. Epub 2012 Dec 5.PMID:23222517 | PMCID:PMC3623669

36. Meirelles OD, Ding J, Tanaka T, ***Sanna S,*** Yang HT, Dudekula DB, Cucca F, Ferrucci L, Abecasis G, Schlessinger D.SHAVE: shrinkage estimator measured for multiple visits increases power in GWAS of quantitative traits. **Eur J Hum Genet.** 2013 Jun;21(6):673-9. doi: 10.1038/ejhg.2012.215. Epub 2012 Oct 24. Erratum in: Eur J Hum Genet. 2014 Jan;22(1):154. PMID:23092954 | PMCID:PMC3658185

37. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. *An integrated map of genetic variation from 1,092 human genomes*. **Nature**. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632. PMID: 23128226

38. Chen W, Li B, Zeng Z, ***Sanna S***, Sidore C, Busonero F, Kang HM, Li Y, Abecasis GR. *Genotype calling and haplotyping in parent-offspring trios.* **Genome Res.** 2013 Jan;23(1):142-51. doi: 10.1101/gr.142455.112. Epub 2012 Oct 11.PMID:23064751 | PMCID:PMC3530674

39. Li B, Chen W, Zhan X, Busonero F, ***Sanna S***, Sidore C, Cucca F, Kang HM, Abecasis GR. *A likelihood-based framework for variant calling and de novo mutation detection in families.* **PLoS Genet.** 2012;8(10):e1002944. doi: 10.1371/journal.pgen.1002944. Epub 2012 Oct 4.PMID:23055937 | PMCID:PMC3464213

40. Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, Chasman DI, Rose LM, Thorleifsson G, Steinthorsdottir V, Magi R, Waite L, Smith AV, Yerges-Armstrong LM, Monda KL, Hadley D, Mahajan A, Li G, Kapur K, Vitart V, Huffman JE, Wang SR, Palmer C, et al. *FTO genotype is associated with phenotypic variability of body mass index*. **Nature**. 2012 Oct 11;490(7419):267-72. doi: 10.1038/nature11401. Epub 2012 Sep 16.PMID:22982992 | PMCID:PMC3564953

41. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, Luan J, Maagi R, Strawbridge RJ, Rehnberg E, Gustafsson S, Kanoni S, Rasmussen-Torvik LJ, Yengo L, Lecoeur C, Shungin D, ***Sanna S***, Sidore C, Johnson PC, Jukema JW, Johnson T, Mahajan A, Verweij N, et al. *Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways.* **Nat Genet**. 2012 Sep;44(9):991-1005. doi: 10.1038/ng.2385. Epub 2012 Aug 12.PMID:22885924 | PMCID:PMC3433394

42. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J, Frayling TM, Heid IM, Jackson AU, Johnson T, KilpelÃ¤inen TO, Lindgren CM, Morris AP, Prokopenko I, Randall JC, Saxena R, Soranzo N, Speliotes EK, Teslovich TM, Wheeler E, Maguire J, Parkin M, Potter S, Rayner NW, Robertson N, Stirrups K, Winckler W, ***Sanna S***, Mulas A, Nagaraja R, Cucca F, Barroso I, Deloukas P, Loos RJ, Kathiresan S, Munroe PB, Newton-Cheh C, Pfeufer A, Samani NJ, Schunkert H, Hirschhorn JN, Altshuler D, McCarthy MI, Abecasis GR, Boehnke M. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. **PLoS Genet.** 2012;8(8):e1002793. doi: 10.1371/journal.pgen.1002793. Epub 2012 Aug 2. Erratum in: PLoS Genet. 2013 Apr;9(4). doi: 10.1371/annotation/0b4e9c8b-35c5-4dbd-b95b-0640250fbc87. PMID:22876189 | PMCID:PMC3410907

43. Matesanz F, Gonzalez-Perez A, Lucas M, ***Sanna S***, Gayan J, Urcelay E, Zara I, Pitzalis M, Cavanillas ML, Arroyo R, Zoledziewska M, Marrosu M, FernÃ¡ndez O, Leyva L, Alcina A, Fedetz M, Moreno-Rey C, Velasco J, Real LM, Ruiz-Pena JL, Cucca F, Ruiz A, et al. *Genome-wide association study of multiple sclerosis confirms a novel locus at 5p13.1.***PLoS One.** 2012;7(5):e36140. doi: 10.1371/journal.pone.0036140. Epub 2012 May 3.PMID:22570697 | PMCID:PMC3343041

44. Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikainen LP, Fuchsberger C, Tanaka T, Morris AP, Small K, Isaacs A, Beekman M, Coassin S, Lohman K, Qi L, Kanoni S, Pankow JS, Uh HW, et al. *Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals*. **PLoS Genet**. 2012;8(3):e1002607. doi: 10.1371/journal.pgen.1002607. Epub 2012 Mar 29.PMID:22479202 | PMCID:PMC3315470

45. He C, Chasman DI, Dreyfus J, Hwang SJ, Ruiter R, ***Sanna S***, Buring JE, FernÃ¡ndez-Rhodes L, Franceschini N, Hankinson SE, Hofman A, Lunetta KL, Palmieri G, Porcu E, Rivadeneira F, Rose LM, Splansky GL, Stolk L, Uitterlinden AG, Chanock SJ, Crisponi L, Demerath EW, et al. *Reproductive aging-associated common genetic variants and the risk of breast cancer.* **Breast Cancer Res.** 2012 Mar 20;14(2):R54.PMID:22433456 | PMCID:PMC3446388

46. Naitza S, Porcu E, Steri M, Taub DD, Mulas A, Xiao X, Strait J, Dei M, Lai S, Busonero F, Maschio A, Usala G, Zoledziewska M, Sidore C, Zara I, Pitzalis M, Loi A, Virdis F, Piras R, Deidda F, Whalen MB, Crisponi L, Concas A, Podda C, Uzzau S, Scheet P, Longo DL, Lakatta E, Abecasis GR, Cao A, Schlessinger D, Uda M, ***Sanna S****, Cucca F*. *A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation*. **PLoS Genet.** 2012 Jan;8(1):e1002480. doi: 10.1371/journal.pgen.1002480. Epub 2012 Jan 26.PMID:22291609 | PMCID:PMC3266885

47. Stolk L, Perry JR, Chasman DI, He C, Mangino M, Sulem P, Barbalic M, Broer L, Byrne EM, Ernst F, Esko T, Franceschini N, Gudbjartsson DF, Hottenga JJ, Kraft P, McArdle PF, Porcu E, Shin SY, Smith AV, van Wingerden S, Zhai G, Zhuang WV, et al. *Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways.* **Nat Genet**. 2012 Jan 22;44(3):260-8. doi: 10.1038/ng.1051.PMID:22267201 | PMCID:PMC3288642

48. Palmer ND, McDonough CW, Hicks PJ, Roh BH, Wing MR, An SS, Hester JM, Cooke JN, Bostrom MA, Rudock ME, Talbert ME, Lewis JP; DIAGRAM Consortium; MAGIC Investigators, Ferrara A, Lu L, Ziegler JT, Sale MM, Divers J, Shriner D, Adeyemo A, Rotimi CN, et al. *A genome-wide association search for type 2 diabetes genes in African Americans.* **PLoS One.** 2012;7(1):e29202. doi: 10.1371/journal.pone.0029202. Epub 2012 Jan 4.PMID:22238593 | PMCID:PMC3251563

49. Cagliani R, Guerini FR, Fumagalli M, Riva S, Agliardi C, Galimberti D, Pozzoli U, Goris A, Dubois B, Fenoglio C, Forni D, ***Sanna S***, Zara I, Pitzalis M, Zoledziewska M, Cucca F, Marini F, Comi GP, Scarpini E, Bresolin N, Clerici M, Sironi M. *A trans-specific polymorphism in ZC3HAV1 is maintained by long-standing balancing selection and may confer susceptibility to multiple sclerosis.* **Mol Biol Evol.** 2012 Jun;29(6):1599-613. doi: 10.1093/molbev/mss002. Epub 2012 Jan 6. PMID: 22319148

2011

50. Luciano M, Lopez LM, de Moor MH, Harris SE, Davies G, Nutile T, Krueger RF, Esko T, Schlessinger D, Toshiko T, Derringer JL, Realo A, Hansell NK, Pergadia ML, Pesonen AK, ***Sanna S***, Terracciano A, Madden PA, Penninx B, Spinhoven P, Hartman CA, Oostra BA, et al. *Longevity candidate genes and their association with personality traits in the elderly*. **Am J Med Genet B Neuropsychiatr Genet**. 2012 Mar;159B(2):192-200. doi: 10.1002/ajmg.b.32013. Epub 2011 Dec 27.PMID:22213687 | PMCID:PMC3583011

51. Gieger C*, Radhakrishnan A*, Cvejic A*, Tang W*, Porcu E*, Pistis G, Serbanovic-Canic J, Elling U, Goodall AH, Labrune Y, Lopez LM, MÃ¤gi R, Meacham S, Okada Y, Pirastu N, Sorice R, Teumer A, Voss K, Zhang W, Ramirez-Solis R, Bis JC, Ellinghaus D, ……. Stemple D, Toniolo D, Wernisch L, ***Sanna S*,*** Hicks AA*, Rendon A*, Ferreira MA*, Ouwehand WH*, Soranzo N*. *New gene functions in megakaryopoiesis and platelet formation.* **Nature**. 2011 Nov 30;480(7376):201-8. doi: 10.1038/nature10659.PMID:22139419 | PMCID:PMC3335296

52. Erriu M, ***Sanna S***, Nucaro A, Orru' G, Garau V, Montaldo C. *HLA-DQB1 Haplotypes and their Relation to Oral Signs Linked to Celiac Disease Diagnosis*. **Open Dent J.** 2011;5:174-8. doi: 10.2174/1874210601105010174. Epub 2011 Nov 4.PMID:22135701 | PMCID:PMC3227877

53. Mitchell GF, Verwoert GC, Tarasov KV, Isaacs A, Smith AV, Yasmin, Rietzschel ER, Tanaka T, Liu Y, Parsa A, Najjar SS, O'Shaughnessy KM, Sigurdsson S, De Buyzere ML, Larson MG, Sie MP, Andrews JS, Post WS, Mattace-Raso FU, McEniery CM, Eiriksdottir G, Segers P, et al. *Common genetic variation in the 3'-BCL11B gene desert is associated with carotid-femoral pulse wave velocity and excess cardiovascular disease risk: the AortaGen Consortium*. **Circ Cardiovasc Genet.** 2012 Feb 1;5(1):81-90. doi: 10.1161/CIRCGENETICS.111.959817. Epub 2011 Nov 8.PMID:22068335 | PMCID:PMC3288392

54. Terracciano A, Piras MG, Lobina M, Mulas A, Meirelles O, Sutin AR, Chan W, ***Sanna S***, Uda M, Crisponi L, Schlessinger D. *Genetics of serum BDNF: meta-analysis of the Val66Met and genome-wide association study.* **World J Biol Psychiatry**. 2013 Dec;14(8):583-9. doi: 10.3109/15622975.2011.616533. Epub 2011 Nov 2.PMID:22047184 | PMCID:PMC3288597

55. Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, Holm H***, Sanna S***, Kavousi M, Baumeister SE, Coin LJ, Deng G, Gieger C, Heard-Costa NL, Hottenga JJ, KÃhnel B, Kumar V, Lagou V, Liang L, Luan J, Vidal PM, Mateo Leach I, et al. *Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma.* **Nat Genet.** 2011 Oct 16;43(11):1131-8. doi: 10.1038/ng.970.PMID:22001757 | PMCID:PMC3482372

56. Bis JC, Kavousi M, Franceschini N, Isaacs A, Abecasis GR, Schminke U, Post WS, Smith AV, Cupples LA, Markus HS, Schmidt R, Huffman JE, LehtimÃ¤ki T, Baumert J, Manzel T, Heckbert SR, Dehghan A, North K, Oostra B, Bevan S, Stoegerer EM, Hayward C, et al. *Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque.* **Nat Genet.** 2011 Sep 11;43(10):940-7. doi: 10.1038/ng.920.PMID:21909108 | PMCID:PMC3257519

57. ***Sanna S****, Li B*, Mulas A*, Sidore C, Kang HM, Jackson AU, Piras MG, Usala G, Maninchedda G, Sassu A, Serra F, Palmas MA, Wood WH 3rd, Nja¸lstad I, Laakso M, Hveem K, Tuomilehto J, Lakka TA, Rauramaa R, Boehnke M, Cucca F, Uda M, et al. *Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability*. **PLoS Genet.** 2011 Jul;7(7):e1002198. doi: 10.1371/journal.pgen.1002198. Epub 2011 Jul 28.PMID:21829380 | PMCID:PMC3145627

58. Pichler I*, Minelli C*, ***Sanna S****, Tanaka T, Schwienbacher C, Naitza S, Porcu E, Pattaro C, Busonero F, Zanon A, Maschio A, Melville SA, Grazia Piras M, Longo DL, Guralnik J, Hernandez D, Bandinelli S, Aigner E, Murphy AT, Wroblewski V, Marroni F, Theurl I, et al. *Identification of a common variant in the TFR2 gene implicated in the physiological regulation of serum iron levels.* **Hum Mol Genet**. 2011 Mar 15;20(6):1232-40. doi: 10.1093/hmg/ddq552. Epub 2010 Dec 28.PMID:21208937 | PMCID:PMC3043660

2010

59. Elks CE, Perry JR, Sulem P, Chasman DI, Franceschini N, He C, Lunetta KL, Visser JA, Byrne EM, Cousminer DL, Gudbjartsson DF, Esko T, Feenstra B, Hottenga JJ, Koller DL, Kutalik Z, Lin P, Mangino M, Marongiu M, McArdle PF, Smith AV, Stolk L, et al. *Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies.* **Nat Genet.** 2010 Dec;42(12):1077-85. doi: 10.1038/ng.714.PMID:21102462 | PMCID:PMC3140055

60. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, MÃ¤gi R, Randall JC, Vedantam S, Winkler TW, Qi L, Workalemahu T, Heid IM, Steinthorsdottir V, Stringham HM, Weedon MN, Wheeler E, Wood AR, Ferreira T, et al. *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* **Nat Genet**. 2010 Nov;42(11):937-48. doi: 10.1038/ng.686. Epub 2010 Oct 10.PMID:20935630 | PMCID:PMC3014648

61. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Magi R, Workalemahu T, White CC, Bouatia-Naji N, Harris TB, Berndt SI, Ingelsson E, Willer CJ, Weedon MN, Luan J, Vedantam S, Esko T, Kilpelainen TO, et al. *Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.* **Nat Genet.** 2010 Nov;42(11):949-60. doi: 10.1038/ng.685. Epub 2010 Oct 10. Erratum in: Nat Genet. 2011 Nov;43(11):1164. PMID:20935629 | PMCID:PMC3000924

62. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segr AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, et al. *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* **Nature**. 2010 Oct 14;467(7317):832-8. doi: 10.1038/nature09410. Epub 2010 Sep 29.PMID:20881960 | PMCID:PMC2955183

63. Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, Ripatti S, Aulchenko YS, Zhang W, Yuan X, Lim N, Luan J, Ashford S, Wheeler E, Young EH, Hadley D, Thompson JR, Braund PS, Johnson T, Struchalin M, Surakka I, Luben R, Khaw KT, et al. *Genetic variants influencing circulating lipid levels and risk of coronary artery disease.* **Arterioscler Thromb Vasc Biol**. 2010 Nov;30(11):2264-76. doi: 10.1161/ATVBAHA.109.201020. Epub 2010 Sep 23.PMID:20864672 | PMCID:PMC3891568

64. Soranzo N*, ***Sanna S****, Wheeler E, Gieger C, Radke D, Dupuis J, Bouatia-Naji N, Langenberg C, Prokopenko I, Stolerman E, Sandhu MS, Heeney MM, Devaney JM, Reilly MP, Ricketts SL, Stewart AF, Voight BF, Willenborg C, Wright B, Altshuler D, Arking D, Balkau B, et al. *Common variants at 10 genomic loci influence hemoglobin A1(C) levels via glycemic and nonglycemic pathways.* **Diabetes**. 2010 Dec;59(12):3229-39. doi: 10.2337/db10-0502. Epub 2010 Sep 21. Erratum in: Diabetes. 2011 Mar;60(3):1050-1. multiple author names added. PMID:20858683 | PMCID:PMC2992787

65. Terracciano A, Tanaka T, Sutin AR, ***Sanna S***, Deiana B, Lai S, Uda M, Schlessinger D, Abecasis GR, Ferrucci L, Costa PT Jr. *Genome-wide association scan of trait depression.* **Biol Psychiatry.** 2010 Nov 1;68(9):811-7. doi: 10.1016/j.biopsych.2010.06.030. Epub 2010 Aug 25.PMID:20800221 | PMCID:PMC2955852

66. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, Pirruccello JP, Ripatti S, Chasman DI, Willer CJ, Johansen CT, Fouchier SW, Isaacs A, Peloso GM, Barbalic M, Ricketts SL, Bis JC, Aulchenko YS, Thorleifsson G, Feitosa MF, Chambers J, Orho-Melander M, et al. *Biological, clinical and population relevance of 95 loci for blood lipids.* **Nature**. 2010 Aug 5;466(7307):707-13. doi: 10.1038/nature09270.PMID:20686565 | PMCID:PMC3039276

67. Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, de Bakker PI, MÃ¼ller M, Morrison AC, Smith AV, Isaacs A, ***Sanna S***, Orri M, Navarro P, Fuchsberger C, Nolte IM, de Geus EJ, Estrada K, Hwang SJ, Bis JC, Rackert IM, Alonso A, Launer LJ, Hottenga JJ, Rivadeneira F, et al. *Genome-wide association analysis identifies multiple loci related to resting heart rate.* **Hum Mol Genet**. 2010 Oct 1;19(19):3885-94. doi: 10.1093/hmg/ddq303. Epub 2010 Jul 16.PMID:20639392 | PMCID:PMC3657480

68. Pruim RJ, Welch RP, ***Sanna S***, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. *LocusZoom: regional visualization of genome-wide association scan results*. **Bioinformatics.** 2010 Sep 15;26(18):2336-7. doi: 10.1093/bioinformatics/btq419. Epub 2010 Jul 15.PMID:20634204 | PMCID:PMC2935401

69. ***Sanna S****, Pitzalis M*, Zoledziewska M*, Zara I, Sidore C, Murru R, Whalen MB, Busonero F, Maschio A, Costa G, Melis MC, Deidda F, Poddie F, Morelli L, Farina G, Li Y, Dei M, Lai S, Mulas A, Cuccuru G, Porcu E, Liang L, et al. *Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis.* **Nat Genet.** 2010 Jun;42(6):495-7. doi: 10.1038/ng.584. Epub 2010 May 9.PMID:20453840 | PMCID:PMC3786343

70. Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, Sehmi JS, Gale DP, Wass MN, Ahmadi KR, Bakker SJ, Beckmann J, Bilo HJ, Bochud M, Brown MJ, Caulfield MJ, Connell JM, Cook HT, Cotlarciuc I, Davey Smith G, de Silva R, Deng G, Devuyst O, et al. *Genetic loci influencing kidney function and chronic kidney disease.* **Nat Genet.** 2010 May;42(5):373-5. doi: 10.1038/ng.566. Epub 2010 Apr 11.PMID:20383145 | PMCID:PMC3748585

71. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren CM, Magi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, et al. *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* **Nat Genet.** 2010 Feb;42(2):105-16. doi: 10.1038/ng.520. Epub 2010 Jan 17. Erratum in: Nat Genet.2010 May;42(5):464. PMID:20081858 | PMCID:PMC3018764

72. Pfeufer A*, van Noord C*, Marciante KD*, Arking DE*, Larson MG*, Smith AV*, Tarasov KV*, Muller M, Sotoodehnia N, Sinner MF, Verwoert GC, Li M, Kao WH, Kottgen A, Coresh J, Bis JC, Psaty BM, Rice K, Rotter JI, Rivadeneira F, Hofman A, Kors JA, …….. Ellinor PT*, ***Sanna S****, Kääb S*, Witteman JCM*, Alonso A*, Benjamin EJ*, and Susan R. Heckbert18,20,*. *Genome-wide association study of PR interval*. **Nat Genet.** 2010 Feb;42(2):153-9. doi: 10.1038/ng.517. Epub 2010 Jan 10.PMID:20062060 | PMCID:PMC2850197

73. Terracciano A, Tanaka T, Sutin AR, Deiana B, Balaci L, ***Sanna S***, Olla N, Maschio A, Uda M, Ferrucci L, Schlessinger D, Costa PT Jr. *BDNF Val66Met is associated with introversion and interacts with 5-HTTLPR to influence neuroticism.* **Neuropsychopharmacology.** 2010 Apr;35(5):1083-9. doi: 10.1038/npp.2009.213. Epub 2009 Dec 30.PMID:20042999 | PMCID:PMC2840212

2009

74. Tarasov KV*, ***Sanna S****, Scuteri A, Strait JB, Orru M, Parsa A, Lin PI, Maschio A, Lai S, Piras MG, Masala M, Tanaka T, Post W, O'Connell JR, Schlessinger D, Cao A, Nagaraja R, Mitchell BD, Abecasis GR, Shuldiner AR, Uda M, Lakatta EG, et al. *COL4A1 is associated with arterial stiffness by genome-wide association scan*. **Circ Cardiovasc Genet.** 2009 Apr;2(2):151-8. doi: 10.1161/CIRCGENETICS.108.823245. Epub 2009 Feb 18.PMID:20031579 | PMCID:PMC2801872

75. Li Y, Willer C, ***Sanna S***, Abecasis G. *Genotype imputation.* **Annu Rev Genomics Hum Genet.** 2009;10:387-406. doi: 10.1146/annurev.genom.9.081307.164242. Review.PMID:19715440 | PMCID:PMC2925172

76. Galanello R, ***Sanna S***, Perseu L, Sollaino MC, Satta S, Lai ME, Barella S, Uda M, Usala G, Abecasis GR, Cao A. *Amelioration of Sardinian beta0 thalassemia by genetic modifiers*. **Blood.** 2009 Oct 29;114(18):3935-7. doi: 10.1182/blood-2009-04-217901. Epub 2009 Aug 20.PMID:19696200 | PMCID:PMC2925722

77. Milan DJ, Kim AM, Winterfield JR, Jones IL, Pfeufer A,***Sanna S***, Arking DE, Amsterdam AH, Sabeh KM, Mably JD, Rosenbaum DS, Peterson RT, Chakravarti A, Kaab S, Roden DM, MacRae CA. *Drug-sensitized zebrafish screen identifies multiple genes, including GINS3, as regulators of myocardial repolarization.* **Circulation.** 2009 Aug 18;120(7):553-9. doi: 10.1161/CIRCULATIONAHA.108.821082. Epub 2009 Aug 3.PMID:19652097 | PMCID:PMC2771327

78. Nolte IM, Wallace C, Newhouse SJ, Waggott D, Fu J, Soranzo N, Gwilliam R, Deloukas P, Savelieva I, Zheng D, Dalageorgou C, Farrall M, Samani NJ, Connell J, Brown M, Dominiczak A, Lathrop M, Zeggini E, Wain LV; Wellcome Trust Case Control Consortium; DCCT/EDIC Research Group, Newton-Cheh C, et al. *Common genetic variation near the phospholamban gene is associated with cardiac repolarisation: meta-analysis of three genome-wide association studies.* **PLoS One**. 2009 Jul 9;4(7):e6138. doi: 10.1371/journal.pone.0006138.PMID:19587794 | PMCID:PMC2704957

79. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, Jackson AU, Lim N, Scheet P, Soranzo N, Amin N, Aulchenko YS, Chambers JC, Drong A, Luan J, Lyon HN, Rivadeneira F,***Sanna S***, et al. *Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution*. **PLoS Genet.** 2009 Jun;5(6):e1000508. doi: 10.1371/journal.pgen.1000508. Epub 2009 Jun 26. Erratum in: PLoS Genet. 2009 Jul;5(7). doi: 10.1371/annotation/b6e8f9f6-2496-4a40-b0e3-e1d1390c1928. PMID:19557161 | PMCID:PMC2695778

80. Kolz M*, Johnson T*, ***Sanna S****, Teumer A*, Vitart V, Perola M, Mangino M, Albrecht E, Wallace C, Farrall M, Johansson A, Nyholt DR, Aulchenko Y, Beckmann JS, Bergmann S, Bochud M, Brown M, Campbell H; EUROSPAN Consortium, Connell J, Dominiczak A, Homuth G, et al. *Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations.* **PLoS Genet.** 2009 Jun;5(6):e1000504. doi: 10.1371/journal.pgen.1000504. Epub 2009 Jun 5.PMID:19503597 | PMCID:PMC2683940

81. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, Papadakis K, Voight BF, Scott LJ, Zhang F, Farrall M, Tanaka T, Wallace C, Chambers JC, Khaw KT, Nilsson P, van der Harst P, Polidoro S, et al. *Genome-wide association study identifies eight loci associated with blood pressure.* **Nat Genet.** 2009 Jun;41(6):666-76. doi: 10.1038/ng.361. Epub 2009 May 10.PMID:19430483 | PMCID:PMC2891673

82. ***Sanna S****, Busonero F*, Maschio A, McArdle PF, Usala G, Dei M, Lai S, Mulas A, Piras MG, Perseu L, Masala M, Marongiu M, Crisponi L, Naitza S, Galanello R, Abecasis GR, Shuldiner AR, Schlessinger D, Cao A, Uda M. *Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia.* **Hum Mol Genet**. 2009 Jul 15;18(14):2711-8. doi: 10.1093/hmg/ddp203. Epub 2009 May 6.PMID:19419973 | PMCID:PMC2701337

83. Pfeufer A*, ***Sanna S****,* Arking DE*, Maller M, Gateva V, Fuchsberger C, Ehret GB, Orru M, Pattaro C, Kottgen A, Perz S, Usala G, Barbalic M, Li M, Putz B, Scuteri A, Prineas RJ, Sinner MF, Gieger C, Najjar SS, Kao WH, Muhleisen TW, et al. *Common variants at ten loci modulate the QT interval duration in the QTSCD Study.* **Nat Genet.** 2009 Apr;41(4):407-14. doi: 10.1038/ng.362. Epub 2009 Mar 22.PMID:19305409 | PMCID:PMC2976045

84. Terracciano A, Balaci L, Thayer J, Scally M, Kokinos S, Ferrucci L, Tanaka T, Zonderman AB,***Sanna S***, Olla N, Zuncheddu MA, Naitza S, Busonero F, Uda M, Schlessinger D, Abecasis GR, Costa PT Jr. *Variants of the serotonin transporter gene and NEO-PI-R Neuroticism: No association in the BLSA and SardiNIA samples*. **Am J Med Genet B Neuropsychiatr Genet.** 2009 Dec 5;150B(8):1070-7. doi: 10.1002/ajmg.b.30932.PMID:19199283 | PMCID:PMC2788669

2008

85. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C, Lettre G, Lim N, Lyon HN, McCarroll SA, Papadakis K, Qi L, Randall JC, Roccasecca RM, ***Sanna S***, Scheet P, Weedon MN, Wheeler E, et al. *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation*. **Nat Genet**. 2009 Jan;41(1):25-34. doi: 10.1038/ng.287. Epub 2008 Dec 14.PMID:19079261 | PMCID:PMC2695662

86. Prokopenko I, Langenberg C, Florez JC, Saxena R, Soranzo N, Thorleifsson G, Loos RJ, Manning AK, Jackson AU, Aulchenko Y, Potter SC, Erdos MR,***Sanna S***, Hottenga JJ, Wheeler E, Kaakinen M, Lyssenko V, Chen WM, Ahmadi K, Beckmann JS, Bergman RN, Bochud M, et al. *Variants in MTNR1B influence fasting glucose levels.* **Nat Genet.** 2009 Jan;41(1):77-81. doi: 10.1038/ng.290. Epub 2008 Dec 7.PMID:19060907 | PMCID:PMC2682768

87. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, Voight BF, Bonnycastle LL, Jackson AU, Crawford G, Surti A, Guiducci C, Burtt NP, Parish S, Clarke R, Zelenika D, Kubalanza KA, Morken MA, et al. *Common variants at 30 loci contribute to polygenic dyslipidemia.* **Nat Genet**. 2009 Jan;41(1):56-65. doi: 10.1038/ng.291. Epub 2008 Dec 7.PMID:19060906 | PMCID:PMC2881676

88. Lettre G, Sankaran VG, Bezerra MA, Arajo AS, Uda M, ***Sanna S,*** Cao A, Schlessinger D, Costa FF, Hirschhorn JN, Orkin SH. *DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease.* **Proc Natl Acad Sci U S A**. 2008 Aug 19;105(33):11869-74. doi: 10.1073/pnas.0804799105. Epub 2008 Jul 30.PMID:18667698 | PMCID:PMC2491485

89. Chen WM, Erdos MR, Jackson AU, Saxena R, ***Sanna S***, Silver KD, Timpson NJ, Hansen T, OrrÃ¹ M, Grazia Piras M, Bonnycastle LL, Willer CJ, Lyssenko V, Shen H, Kuusisto J, Ebrahim S, Sestu N, Duren WL, Spada MC, Stringham HM, Scott LJ, Olla N, et al. *Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels.* **J Clin Invest.** 2008 Jul;118(7):2620-8. doi: 10.1172/JCI34566.PMID:18521185 | PMCID:PMC2398737

90. Arnaud-Lopez L, Usala G, Ceresini G, Mitchell BD, Pilia MG, Piras MG, Sestu N, Maschio A, Busonero F, Albai G, Dei M, Lai S, Mulas A, Crisponi L, Tanaka T, Bandinelli S, Guralnik JM, Loi A, Balaci L, Sole G, Prinzis A, Mariotti S, et al. *Phosphodiesterase 8B gene variants are associated with serum TSH levels and thyroid function.* **Am J Hum Genet.** 2008 Jun;82(6):1270-80. doi: 10.1016/j.ajhg.2008.04.019.PMID:18514160 | PMCID:PMC2427267

91. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, Berndt SI; Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Jacobs KB, Chanock SJ, Hayes RB, Bergmann S, Bennett AJ, Bingham SA, Bochud M, Brown M, Cauchi S, Connell JM, et al. *Common variants near MC4R are associated with fat mass, weight and risk of obesity.* **Nat Genet.** 2008 Jun;40(6):768-75. doi: 10.1038/ng.140. Epub 2008 May 4.PMID:18454148 | PMCID:PMC2669167

92. Lettre G*, Jackson AU*, Gieger C*, Schumacher FR*, Berndt SI*, ***Sanna S****, Eyheramendy S, Voight BF, Butler JL, Guiducci C, Illig T, Hackett R, Heid IM, Jacobs KB, Lyssenko V, Uda M; Diabetes Genetics Initiative; FUSION; KORA; Prostate, Lung Colorectal and Ovarian Cancer Screening Trial; Nurses' Health Study; SardiNIA, et al. *Identification of ten loci associated with height highlights new biological pathways in human growth*. **Nat Genet.** 2008 May;40(5):584-91. doi: 10.1038/ng.125. Epub 2008 Apr 6.PMID:18391950 | PMCID:PMC2687076

93. Uda M*, Galanello R*, ***Sanna S****, Lettre G, Sankaran VG, Chen W, Usala G, Busonero F, Maschio A, Albai G, Piras MG, Sestu N, Lai S, Dei M, Mulas A, Crisponi L, Naitza S, Asunis I, Deiana M, Nagaraja R, Perseu L, Satta S, et al. *Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia.* **Proc Natl Acad Sci U S A.** 2008 Feb 5;105(5):1620-5. doi: 10.1073/pnas.0711566105. Epub 2008 Feb 1.PMID:18245381 | PMCID:PMC2234194

94. ***Sanna S****, Jackson AU*, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G, Chines PS, Stringham HM, Scott LJ, Dei M, Lai S, Albai G, Crisponi L, Naitza S, Doheny KF, Pugh EW, Ben-Shlomo Y, Ebrahim S, et al. *Common variants in the GDF5-UQCC region are associated with variation in human height.* **Nat Genet**. 2008 Feb;40(2):198-203. doi: 10.1038/ng.74. Epub 2008 Jan 13.PMID:18193045 | PMCID:PMC2914680

95. Willer CJ*,**Sanna S***, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, et al. *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* **Nat Genet.** 2008 Feb;40(2):161-9. doi: 10.1038/ng.76. Epub 2008 Jan 13.PMID:18193043

2007

96. Li S*, ***Sanna S****, Maschio A, Busonero F, Usala G, Mulas A, Lai S, Dei M, Orru M, Albai G, Bandinelli S, Schlessinger D, Lakatta E, Scuteri A, Najjar SS, Guralnik J, Naitza S, Crisponi L, Cao A, Abecasis G, Ferrucci L, Uda M, et al. *The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts*. **PLoS Genet.** 2007 Nov;3(11):e194.PMID:17997608 | PMCID:PMC2065883

97. Scuteri A*, ***Sanna S****, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, OrrÃº M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, et al. *Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.* **PLoS Genet.** 2007 Jul;3(7):e115.PMID:17658951 | PMCID:PMC1934391

98. Balaci L, Spada MC, Olla N, Sole G, Loddo L, Anedda F, Naitza S, Zuncheddu MA, Maschio A, Altea D, Uda M, Pilia S,***Sanna S***, Masala M, Crisponi L, Fattori M, Devoto M, Doratiotto S, Rassu S, Mereu S, Giua E, Cadeddu NG, et al. *IRAK-M is involved in the pathogenesis of early-onset persistent asthma*. **Am J Hum Genet**. 2007 Jun;80(6):1103-14. Epub 2007 Apr 27.PMID:17503328 | PMCID:PMC1867098