A.D. MDLXII

UNIVERSITÀ DEGLI STUDI DI SASSARI

Scuola di Dottorato in Scienze Biomediche

Direttore: Prof. Andrea Piana

XXVII CICLO DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE
INDIRIZZO IN GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

Responsabile di indirizzo: Prof. Francesco Cucca

Sequence-based GWAS using thousands Sardinian genomes: an application to quantitative traits

Relatori:
Francesco Cucca
Serena Sanna

Dottoranda:
Eleonora Porcu

Anno Accademico 2013/2014

The results showed in this thesis are part of two manuscripts, one submitted to *Nature Genetics* and one in press at *Nature Communications*.

The complete citations of the publications are:
Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, et al., Genome sequencing elucidates Sardinian genetic architecture and augments GWAS findings: the examples of lipids and blood inflammatory markers, *Nature Genetics*, submitted.

Taylor P, Porcu E, et al., Whole-genome sequence-based analysis of thyroid function, *Nature Communications*, In press

Index

1. Introduction

With the completion of the Human Genome Project in 2003[1] and the International HapMap Project[2] in 2005, researchers began to pinpoint areas of the genome that varies between individuals. Shortly thereafter, they discovered that the most common type of DNA sequence variation found in the genome is the single nucleotide polymorphism (SNP).

The public data of the HapMap project, and the more recent 1000 Genomes project[3], which systematically and comprehensively catalog human variations of different populations, have facilitated a new type of research effort: the genome-wide association study (GWAS).

The basic approach in GWAS is to evaluate the association between each SNP and a quantitative, or qualitative, trait of interest that has been measured across a large number of individuals. The power of a GWAS study is proportional to the effect of the causative variant to be found, therefore strongly depends on the number of individuals and the number of genetic markers assessed.

The first successful GWAS was published in 2005 and investigated a few hundreds patients with age-related macular degeneration and controls[4]. Since 2005, the GWAS approach has been applied to hundreds of complex traits and diseases, constantly enlarging the number of individuals and improving genetic resolution – although with limitations dictated by the status of available technologies. All such efforts led to more than 2,000 published human GWAS (*http://www.genome.gov/gwastudies*).

Although these studies have revealed thousands of loci predisposing to hundreds of human diseases and traits, these variants have explained relatively little of the heritability

- the portion of phenotypic variance in a population attributable to additive genetic factors - of most complex traits. For example, the estimated heritability for the human height is 80% but although several studies of tens to hundreds of thousands individuals has been conducted, the loci associated with height explain only about 16% of phenotypic variance[5].

Several proposed explanations for this "missing heritability" include[6,7]:

1. effect sizes of associated variants may be underestimates due to incomplete linkage disequilibrium (LD) between causal variants and SNPs we tested;
2. the polygenic model of complex traits includes several low-frequency polymorphisms (minor allele frequency (MAF) < 5%) or rare variants (MAF < 1%) that are not sufficiently frequent to be captured by current genotyping arrays;
3. heritability may be overestimated[8], with epistasis, epigenetics, and genotype–environment interactions contributing to trait heritability;
4. many additional, currently undetected small effects (both at common and rare variants) may together comprise a significant contribution to heritability.

For the first two hypotheses, whole-genome sequencing represents a good investment as it allows an accurate identification of the lowest frequency variants and distinguishing causal variants among so many.

However efficient detection of rare and low frequency variants requires sequencing hundreds to thousands of individuals which costs are still prohibitively high.

An alternative cost-effective approach is to sequence a subset from a study sample that incorporates maximal number of variants (i.e. founders individuals), and use their haplotypes to impute the missing genotypes in the rest study sample. In parallel, meta-analyses of different GWAS cohorts is another cost-effective strategy to assess variants in hundreds of thousands of individuals at a minimum-cost.

This work is subdivided in two parts where I will present two different genetic studies that involved whole-genome sequencing data.

In the first section I will present how whole-genome low-pass sequencing of 2,120 individuals from the Sardinian founder population allowed assessment of ~13.6 million variants that, tested in 6,602 individuals, yielded novel associations with the levels of five inflammatory biomarkers: adiponectin, high-sensitivity C-reactive protein, erythrocyte sedimentation rate, monocyte chemotactic protein-1 and interleukin-6.

In the second part I will present a meta-analysis of 7 cohorts (totaling up to 16,335 individuals) for common and low-frequent variants (MAF >= 1%) associated with thyroid function using whole-genome sequence data from various sources and deeply imputed datasets. In particular, I will show how increasing coverage in whole-genome sequence association studies permits identification of novel variants associated with thyroid stimulating hormone and free thyroxine.

1.1 Genome-wide association studies

In the genome-wide association studies (GWAS) hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested for association with a phenotype in hundreds or thousands of individuals. They were made possible by the availability of chip-based microarray technology for assaying hundreds of thousand SNPs.

In GWAS no *a priori* biological knowledge is needed and they are therefore an agnostic method for localizing the genetic effects even in unsuspected genes.

The association analysis of genome-wide data is a series of single-locus statistical tests, examining each SNP independently for association to the phenotype. The specific statistical test chosen to assess association depends on a variety of factors but one major distinction is dictated by the type of phenotype of interest: binary or quantitative.

Quantitative traits are generally analyzed using generalized linear model (GLM) approaches, most commonly the Analysis of Variance (ANOVA), which is similar to linear regression with a categorical predictor variable, in this case genotype classes. The null hypothesis of an ANOVA using a single SNP is that there is no difference between the trait means of any genotype group.

Binary traits are generally analyzed using either logistic regression or contingency table methods. Researchers often prefer logistic regression as it allows for adjustment for clinical covariates. Indeed, the statistical test should be adjusted for all of the factors that are known to influence the trait (age, sex, …) in order to reduce spurious associations due to sampling artifacts.

For each statistical test, we have a p-value, i.e. the probability of seeing a test statistic equal to or greater than the observed test statistic if the null hypothesis is true. This effectively means that lower p-values indicate that if there is no association, the chance of seeing this result is extremely small.

Statistical tests are generally called significant when the p-value is lower than 0.05. This threshold is relative to a single test but in the case of GWAS, we have millions of tests and a multiple testing correction is needed.

A consensus has emerged that $5x10^{-08}$ is the genome-wide significance threshold in a non-African population-based GWAS. This is a conservative Bonferroni correction which adjusts the single test threshold 0.05 to $0.05/k$ where k is the number of tests conducted(9); in this case k is one million as it is the number of independent common SNPs throughout the genome(10).

With the inclusion of low-frequency and rare variants catalogued within the 1000 Genomes Project, in which ~50 million markers (SNPs, insertions and deletions) have been reported, the number of independent loci under study will be significantly larger than the 1 million markers estimated previously. An even more stringent threshold, compared to the typical $5x10^{-8}$, may be required to ensure robust findings(11). Additional studies are needed to indicate the optimal threshold but this threshold may depend on number of low-frequency variants present in the study and may therefore vary from study to study.

Reproducibility of the findings is a key part of the GWAS. Indeed, a repeated observation demonstrates that the association is not due to chance or uncontrolled bias affecting a single study. Furthermore, the replication allows a more precise estimate of the findings and a generalization of them to the wider population.

The best strategy for a replication study is to repeat the same analysis in an independent cohort, from the same population as the GWAS, using identical criteria for exclusions and adjustments of phenotypes, and only for the SNPs passing the genome-wide significant threshold in the GWAS. SNPs showing a significant association (0.05/N, where N is the number of SNPs tested for replication) and with the direction of effect consistent with the GWAS finding for the same allele are considered “replicated”. Replication can be searched in more than one independent cohort, strengthening the finding. Furthermore, the chance that a variant assessed is causative increases when replication is seen also in cohorts of different ethnicities and without heterogeneity in effect size.

1.2 Genotype imputation

Genotype imputation is a statistical technique that is often used to increase the power and resolution of genetic association studies.

Imputation methods infer untyped markers in a study sample by using the LD structure among markers assessed in an external reference panel for which a much denser genetic map is available(12).

Typically, the study sample is genotyped with a commercial genotyping platform for hundreds of thousands to millions of single nucleotide polymorphisms (SNPs) located across the entire genome(13,14).

The HapMap haplotypes have been used to carry out imputation for most of the GWAS published to date, but its use is now being replaced by the larger and more comprehensive set of individuals characterized within the 1000 Genomes Project (1000G). Indeed, while the HapMap set characterized 270 individuals with genotyping arrays for roughly 3 million markers, the 1000G reference set has been generated from whole-genome sequencing of 1,092 individuals (181 samples from Admixed American, 246 from African, 286 from East Asian, and 379 from European ancestry groups), leading to the discovery of about 39.7 million bi-allelic variants; of these, approximately 1.4 million markers are short indels and large deletions, and the rest are SNPs. Imputation performed with this much denser data set will yield a higher resolution of the genome for detection of association signals, thus increasing the power of the existing GWAS to identify novel variants beyond what was found after imputation with the HapMap data set and to pinpoint the causal variants at known associated loci(15).

1.3 Meta-analysis approach

Individual GWAS are generally too small to provide sufficient power to detect associated variants with small effect. To identify these variants, tens of thousands or even hundreds of thousands samples are needed, but such large cohorts are impractical to collect. To augment the sample size and increase the power in a 'virtual' manner, the genetics community has widely adopted the cost-effective approach of combining summary statistics from multiple independent GWASs into a single analysis called meta-analysis(16,17,18).

There are several approaches for GWAS meta-analysis and all of those have as fundamental principle that each study provides statistical results without transferring any genotype or clinical information to the other studies. Notably, it is not necessary that all studies genotype the same set of SNPs because data from different genotyping platforms are imputed to a common reference set and then combined in a joint analysis.

As only key condition is that all studies included adopt the same criteria to collect the phenotype and for modeling it (phenotype transformation method, if any, and adjustment for highly impacting covariates). Indeed, the power to find associations also depends on phenotype definition -- variability in definitions may cause heterogeneity in effect size or even spurious associations.

The most popular and the most powerful method in meta-analysis is the fixed-effect approach assuming that there is one true effect size which is shared by all the included studies. There are different models for fixed-effect meta-analysis, but the inverse variance weighting, in which each study is weighted according to the inverse of its squared standard error, is predominantly used(19).

In meta-analysis, as we are combining results from multiple studies performed by different analysts using different software for single SNP association tests, in populations with slightly different ethnic background and sometimes dealing with phenotypes measured with different instruments, heterogeneity is inevitable. The most popular measure of heterogeneity is Cochran's Q[20], which is calculated as the weighted sum of squared differences between individual study effects and the overall meta-analysis estimate, weighting the contribution of each study in the same manner as in the meta-analysis. Q is distributed as a chi-square statistic with *k*-1 degrees of freedom where *k* is the number of studies included.

Like primary GWAS, meta-analysis usually define P-value threshold at which a finding can be considered genome-wide significant. Usually, the standard genome-wide threshold of $5x10^{-08}$ is used combined with less stringent level ($1x10^{-04}$ or $1x10^{-05}$)[21] to warrant further bioinformatics analysis or replication in independent cohorts.

1.4 Rare variants tests

With the recent technological advances in high-throughput sequencing platforms, the focus of genetic associations is shifting to rare variants[(22,23)].

While statistical methods for detecting associations of common variants have been extensively studied and thousands common variants were found to be associated with hundreds complex traits, methods for statistical analysis of rare variants are limited.

Although methods used with common variants are applicable to rare variants, their performance might not be optimal because they are underpowered unless sample sizes or effect sizes are very large.

In recent years, considerable efforts have been done for developing rare-variant analysis focusing on testing cumulative effects of rare variants in genetic regions, such as genes. These tests can be broadly classified as burden and non-burden tests[(24,25,26,27)].

The approach in burden tests is to fix an allele-frequency threshold (1% - 5%) and combine multiple variants from the same gene below that threshold in a single unit. In this way, each rare variant has the same weight and the genes, rather than individual alleles, are treated as a unit for the association test[(28)].

A more general approach uses a variable allele-frequency threshold (VT test)[(29)], instead of a fixed threshold: rare alleles are grouped together optimizing an allele-frequency threshold that maximizes the difference between distributions of trait values for individuals with and without rare alleles.

Examples of burden tests are the cohort allelic sum test (CAST)[(30)] and the combined multivariate and collapsing method (CMC). CAST collapses information on all rare variants within a gene into a single dichotomous variable for each subject by indicating whether or not the subject has any rare variants within the gene and then applies a univariate test. CMC collapses by counting the rare variants within a gene and then applies a multivariate test.

Kernel-based test methods, such as the sequence kernel association test (SKAT)(31), are non-burden tests. SKAT uses a multiple regression model to directly regress the phenotype on genetic variants in a region and on covariates, and so allows different variants to have different directions and magnitude of effects, including no effects.

Both burden and non-burden tests present limitations. A limitation for all burden tests is that they implicitly assume that all the rare variants in a gene are causal and affect the phenotype in the same direction with the same magnitude. When these assumptions are violated they are underpowered because collapsing all variants is likely to introduce noise into the collapsed value.

By contrast, for SKAT, as for multiple regression models, neither directionality nor magnitudes of the associations are assumed a priori but are instead estimated from the data. Hence, SKAT is more powerful when a large fraction of the variants in a region are noncausal or the effects of causal variants have different directions.

Although SKAT makes few assumptions about rare-variant effects, it can be less powerful than burden tests if a large proportion of the rare variants in a gene are truly causal and influence the phenotype in the same direction.

Taking in account of these limitations, an optimized test has been developed, known as SKAT-O(32). This approach maximizes statistical power by applying both burden-based and sequence kernel association tests: when the burden test is more powerful than SKAT,

SKAT-O behaves like the burden test and when the SKAT is more powerful than the burden test, it behaves like SKAT.

2. The SardiNIA project

2.1 Samples description

The SardiNIA project started in 2001 and recruited 6,921 Sardinians (age 14-102 older), from a cluster of four towns in the Lanusei Valley in the Ogliastra region: Arzana, Elini, Ilbono and Lanusei. This sample corresponded to approximately 62% of the population eligible in the area for recruitment(33,34).

The samples can be grouped in >1000 families, up to 5 generations deep; the largest family has more than 625 genotyped individuals.

While GWAS studies are designed to find common variants with low/moderate attributable risks, family-based studies may facilitate the detection of rare variants with high attributable risk because predisposing variants will be present at much higher frequency in affected relatives of an index case. Moreover, family-based designs can better control both genetic and environmental background and are robust to heterogeneity and population stratification.

All volunteers have been characterized for more than 800 quantitative traits. Traits include anthropomorphic measures, plasma and serum markers (including cholesterol and other markers of 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).

2.2 Genotyping

The entire SardiNIA cohort was genotyped using the HumanOmniExpress GWAS array, containing ~750K markers, and three different Illumina custom arrays: the Cardio-

MetaboChip, ImmunoChip and the HumanExome, each containing about 200,000 markers(35,36). Genotyping calling was performed using the Illumina GenCall algorithm, and an additional 2,968 rare variants were called for HumanExome using Zcall(37).

All samples had a genotyping call rate > 90% in OmniExpress and > 98% in the other arrays. SNP genotypes were carefully assessed though several quality control checks. In particular, the four arrays were analysed independently and removed markers with call rate < 98%, with strong deviation from HWE ($p < 10^{-6}$), that were monomorphic (or with MAF < 1% for OmniExpress) or leading to an excess of Mendelian errors (defined as > 1% of the families or > 1 for ExomeChip SNPs called with Zcall).

In addition, SNPs in common between the arrays that showed a high level of discordance or that generated > 1% discrepancies when comparing genotypes across 13 twins were removed. After performing quality control checks and merging genotypes from the four arrays, the quality checked 886,938 autosomal markers were used as baseline genotypes to impute variants detected through sequencing, as described below.

2.3 Sequencing

Samples to be sequenced were selected in trios, being those highly informative for haplotypes reconstruction. Trios were selected starting from the founders of all available families to assure the representation of all haplotypes that have been propagated within families. Of the 2,120 Sardinian samples sequenced at 4X coverage, 1,122 were part of the SardiNIA project, whereas the remaining 998 were individuals enrolled in case-control studies of Multiple Sclerosis and Type 1 Diabetes(38,39).

To avoid over-representation of rare variants, related samples were removed (mostly child of a trios) and we generated a reference panel containing phased haplotypes of 1,488 individuals and 17 million variants.

2.4 Genotype imputation

Before performing imputation using minimac[(40)], genotypes of all individuals were phased 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. Then, imputation was performed using the phased haplotypes as baseline and Sardinia sequences as reference panel.

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. This strategy lead to 13.6 million markers useful for analyses.

To better understand the benefits of a population based reference panel, an other run of imputation was performed in parallel using the same baseline but the 1000 Genomes data as reference panel (March 2012 release).

We used RSQR > 0.3 for all variants as a filter for imputation accuracy, as recommended (Ref. cookbook). This results in the analysis of 13.5 million markers.

3. Statistical methods

3.1 Genome-wide association analysis

Since the majority of the variants, both directly genotyped with the arrays and coming from the imputation, are low frequency (MAF between 1% and 5%) and rare (MAF<1%) variants, the effect of cryptic relatedness and population stratification could be a cause of spurious associations.

For each trait, each SNP was tested for association using EPACTS[41], a software that performs a linear mixed model adjusted with a genomic-based kinship matrix calculated using all quality checked genotyped SNPs with MAF > 1%.

The advantage of this model is that the kinship matrix encodes a wide range of sample structures, including both cryptic relatedness than population stratification.

3.2 Association analysis - Conditional analysis

There may be multiple causal variants at the same locus, each independently contributing to genetic association with the phenotype. A tool to detect secondary independent signal at a locus is the conditional analysis which consists in a GWAS performed for each trait by adding the leading SNPs found in the primary GWAS as covariates to the basic model. A SNP reaching the standard genome-wide significance threshold ($P < 5x10^{-08}$) was considered a significant independent signal.

3.3 Meta-analysis

We used the GWAMA (Genome-Wide Association Meta Analysis) software[42] to perform meta-analysis of the results of each GWAS from the 7 cohorts participant (see description in Appendix): TwinsUK WGS, TwinsUK GWAS, Avon Longitudinal Study of Parents and Children (ALSPAC) WGS, ALSPAC GWAS, SardiNIA, ValBorbera and Busselton Health Study (BHS).

We performed fixed effects meta-analyses using estimates of the allelic effect size and standard error. Two meta-analysis were performed for each phenotype: a meta-analysis of the two UK10K WGS cohorts (TwinsUK WGS and ALSPAC WGS), and a meta-analysis of all seven cohorts. The ValBorbera cohort does not have FT4 phenotype data so this cohort was not included in the meta-analysis for this phenotype.

In each GWAS cohort, genotyping was performed using different Illumina genome-wide chips and >9 million SNPs were imputed using three different panels as reference: UK10K (http://www.uk10k.org/studies/cohorts.html), 1000 Genomes Phase I and Sardinia. Cohort-specific quality control filters relating to call rate and Hardy-Weinberg equilibrium (HWE) were applied before imputation. Genotype imputation was performed using either the IMPUTE(43), MaCH(44) or Minimac software packages with poorly imputed variants excluded.

An inverse normal transformation was applied to each trait and each SNP was modeled using an additive genetic effect (allele dosage for imputed SNPs), including age and sex as covariates in the model as well as study-specific covariates.

Association analysis within each cohort was performed using either the SNPTEST v247(45), GEMMA (Genome-wide Efficient Mixed Model Association)(46), EPACTS (Efficient and Parallelizable Association Container Toolbox) or ProbABEL(47) software packages.

In the meta-analysis, any variants that were missing from > 2 cohorts or with a combined MAF < 1% were excluded. However, in the meta-analysis performed using whole-genome sequence data a MAF of 0.5% in either cohort was accepted to prevent marginal MAF drop-outs; the MAF < 1% exclusion was then applied during the meta-analysis.

To identify independent association signals, each study repeated the analysis using the top SNPs as covariates. In cohorts where the top SNP was not present, the best proxy (r^2>0.8) was included when available. A meta-analysis was then performed using these results and using the same filters and the same model as in the primary analysis.

Separate work in the UK10K project has identified that a combined analytic strategy of testing common variants (MAF 0.5% or above) using single-SNP tests combined with detailed rare variants analysis would have marginally lower significance threshold of around $1.5x10^{-08}$. To take account of this we have reset the significance threshold to $1.5x10^{-08}$.

3.4 Burden test on inflammatory markers

To improve power on the analysis of rare variants, we performed the Combined Multivariate and Collapsing (CMC) and Variable Threshold tests implemented in EPACTS. 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 $5x10^{-06}$ to declare significance.

3.5 SKAT analysis on thyroid related traits

We conducted GWAS candidate gene (*AADAT*, *ABO*, *B4GALT6*, *CAPNS2*, *CAPZB*, *DIO1*, *DIRC3*, *ELK3*, *FBXO15*, *FGF7*, *FOXA2*, *FOXE1*, *GLIS3*, *HACE1*, *IGFBP2*, *IGFBP5*, *INSR*, *ITPK1*, *LHX3*, *LOC440389/LOC102467146*, *LPCAT2*, *MAF*, *MBIP*, *MIR1179*, *NETO1*, *NFIA*, *NKX2-3*, *NR3C2*, *NRG1*, *PDE10A*, *PDE8B*, *PRDM11*, *RAPGEF5*, *SASH1*, *SIVA1*, *SLC25A52*, *SOX9*, *SYN2*, *TMEM196*, *TPO*, *TTR*, *VAV3*, *VEGFA*) based analyses to test for association of the combined effects of rare variants on TSH and FT4 using SKAT-O software. We used the TwinsUK WGS, ALSPAC WGS and SardiNIA data to examine loci with a known association with TSH and FT4. We examined all SNPs within the candidate gene regions, including variants within 50kb on either side of the gene with MAF less than 1% down to a MAF of 0.04% (in a cohort), or 0.02% (overall). These analyses used sequential nonoverlapping windows each containing 50 SNPs. Association at $P < 1.55 \times 10^{-05}$ (Bonferroni corrected) was considered significant. For the meta-analysis of rare variant data from the WGS cohorts we used SkatMeta(48).

3.6 Calculation of variance explained on thyroid related traits

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 full model contains all the independent SNPs associated to the specific trait in addition to the covariates age, age^2, sex in the basic model for TSH and FT4 and age, age^2, sex, smoke and BMI for inflammatory traits.

Variance for all available SNPs was calculated using GCTA software(49) taking account of both closely and distantly related pairs of individuals. For each trait, we quantified the variance explained by all quality checked SNPs after removing those which were monomorphic in the subset of individuals phenotyped (also known as “accessible genome”).

4. Whole-genome sequence-based GWAS on inflammatory markers

Inflammation is a process by which our organism protects itself from harmful stimuli - such as germs, damaged cells, or irritants - and begins the healing process. It has also been implicated, with both protective and predisposing effects, in several diseases(50,51); but many important details of this complex phenomenon are still unknown. Identifying the genes that influence levels of pro-inflammatory molecules can help to elucidate the factors and mechanisms underlying inflammation and their consequence on health.

We conducted a population sequencing-based GWAS on the levels of five key inflammatory biomarkers: adiponectin (ADPN), high-sensitivity C-reactive protein (hsCRP), erytrocyte sedimentation rate (ESR), monocyte chemotactic protein-1 (MCP-1) and interlukin-6 (IL-6).

4.1 Results

Using the Sardinian reference panel we assessed up to 13.6 million variants and found several SNPs above the standard genome-wide significant threshold ($5x10^{-08}$). In particular, we found 5 hits at 4 novel loci, along with 2 new independent variants at previously reported loci.

Table1. Association results at the genome-wide significant loci. The table shows the association results at the genome-wide significant loci. For each lead SNP, we reported the nearest gene, the rs ID when available, the effect allele and its frequency, the regression coefficients, the imputation accuracy (RSQR) for those that were imputed, the biological type of the corresponding nucleotide change. Novel loci are shown in bold; independent signals are shown in italics.

Nearest Gene	Chr:position	rs name	Effect Allele / Other	Freq	Effect (StdErr)	pvalue	RSQR	Type
ADPN								
ADIPOQ	3:186559460	rs17300539	A/G	0.156	0.247 (0.025)	1.35x10^{-22}	genotyped	intergenic
ABDH13	**13:108884835**	**N/A**	**A/G**	**0.001**	**-1.519 (0.275)**	**3.35x10^{-08}**	**0.921**	**UTR5**
hsCRP								
CRP	1:159684665	rs3091244	A/G	0.428	0.207 (0.019)	5.28x10^{-27}	genotyped	intergenic
PDGFRL	**8:17450500**	**rs73198138**	**A/G**	**0.004**	**-0.894 (0.151)**	**3.31x10^{-09}**	**0.977**	**intronic**
HNF1A	*12:121415293*	*rs7139079*	*G/A*	*0.377*	*-0.123 (0.020)*	*7.70x10^{-10}*	*0.998*	*intergenic*
AACS	**12:125533106**	**rs183233091**	**A/G**	**0.01**	**1.054 (0.094)**	**1.09x10^{-28}**	**0.941**	**intergenic**
APOE/APOC1	19:45411941	rs429358	C/T	0.073	-0.237 (0.036)	3.78x10^{-11}	1	nonsyn
ESR								
TMEM57	*1:25724005*	*rs71721472*	*T/C*	*0.297*	*-0.109 (0.020)*	*4.26x10^{-08}*	*0.957*	*intronic*
CR1	1:207684359	rs11117956	T/G	0.4	-0.153 (0.018)	9.43x10^{-18}	genotyped	intronic
HBB	11:5248004	rs76728603	A/G	0.048	-0.437 (0.042)	1.02x10^{-25}	0.918	stop
AACS	**12:125406340**	**N/A**	**G/A**	**0.007**	**1.034 (0.104)**	**4.40x10^{-23}**	**0.952**	**intergenic**
MCP-1								
DARC	1:159175354	rs12075	G/A	0.446	-0.405 (0.019)	1.08x10^{-96}	-	nonsyn
CADM3	*1:159164454**	*rs2852718*	*C/T*	*0.022*	*-0.515 (0.063)*	3.34x10^{-16}	*0.999*	*intronic*
DARC	***1:159175494****	***rs34599082***	***T/C***	***0.037***	***-0.338 (0.049)***	***8.23x10^{-12}***	**-**	***nonsyn***
CCR2	3:46383906	rs113403743	T/G	0.099	0.273 (0.034)	1.47x10^{-15}	0.997	intergenic
CCR2	***3:46399764****	***rs200491743***	***A/T***	***0.005***	***0.799 (0.130)***	***9.94x10^{-10}***	**-**	***nonsyn***
CBLN1	**16:49072490****	**rs76135610**	**T/C**	**0.005**	**0.969 (0.172)**	**1.76x10^{-08}**	**0.915**	**intergenic**
IL-6								
IL6R	1:154428283	rs12133641	G/A	0.255	0.118 (0.020)	6.87x10^{-09}	1	intronic
ABO	9:136142355	rs643434	A/G	0.263	-0.221 (0.020)	5.80x10^{-27}	-	intronic

Specifically, we detected one novel association at the 3'UTR of the *ABHD13* gene on chromosome 13 (chr13:108884835; p=3.35x10^{-08}) for ADPN, two new signals for hsCRP near the *PDGFRL* (rs73198138; p=3.31x10^{-09}) and *AACS* (rs125533106; p=1.09x10^{-28}) genes, and one for ESR near *AACS* (chr12:125406340; p=4.40x10^{-23}). It is interesting to observe that the top SNPs at the *AACS* locus were only partially correlated (r^2=0.20), but the association with hsCRP disappeared when conditioning for the lead variant for ESR and viceversa. Thus, the two markers are likely representing the same causal variant, consistent with the biological correlation between hsCRP and ESR.

Performing conditional analysis, we also detected two novel independent signals for MCP-1 in the *DARC* (rs34599082; p=8.23x10^{-12} after conditioning on top SNPs rs12075 and rs2852718) and *CCR2* (rs200491743; p=9.94x10^{-10} after conditioning on top SNP rs113403743) genes. Both variants are non-synonymous and cause non-conservative amino acid changes in the corresponding protein.

Furthermore, at 3 loci we were able to detect a more strongly associated variant than previously reported, likely representing the causative one. For example, at the *HBB* gene the top variant associated with ESR levels is now the Q40X stop codon mutation, also known as β° 39, responsible for β-thalassemia when carried in homozygosity. Other refinements were seen on chromosome 1 for ESR, where a previously reported signal(52,53) in an intron of *TMEM57*, encoding a protein with unknown function, has been now mapped to intron 3 of the nearby *RHCE* gene, encoding for the Rh blood group C and E antigens. Finally, we could fine map the previously reported association signal for hsCRP at the *APOC1* locus, with a non-synonymous variant in the *APOE* gene, C130R, which has been associated with Alzheimer's disease but not yet directly with CRP levels(54).

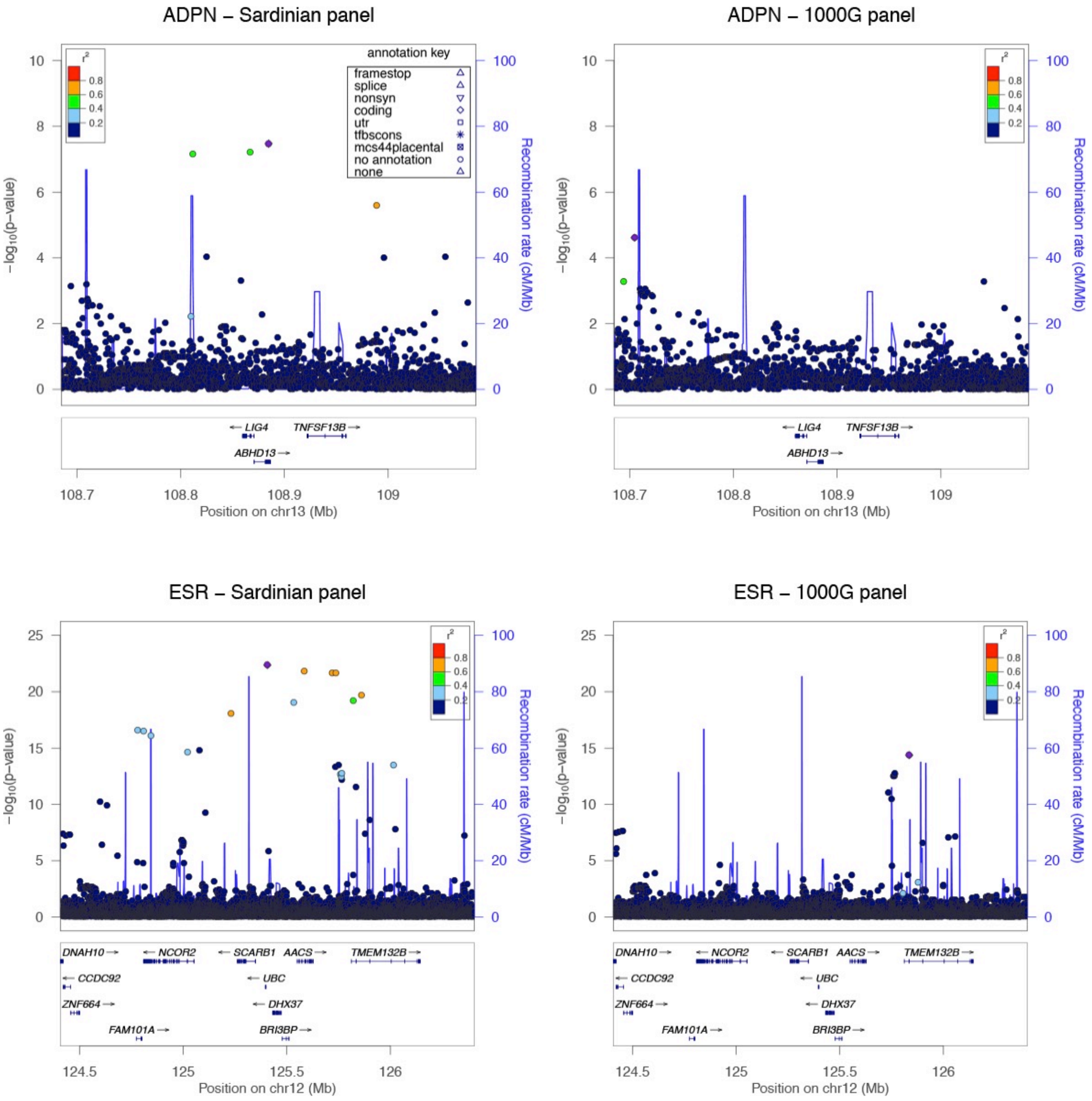
To assess the potential sex-specific impact of the associated variants, we conducted two different GWAS for males and females separately. [Supplementary Table 1 – Sex specific results]

We found a novel signal for MCP-1 at *CBLN1* gene which was significant only in females (rs76135610, $p=1.76x10^{-08}$ and p=0.13 in females and males respectively, heterogeneity $p=8.06x10^{-03}$). Furthermore, of all top associated variants in main analysis, significant evidence for heterogeneity of effects in males and females ($p< 0.002$, corresponding to a Bonferroni threshold of 0.05/18) was observed at *AACS* (heterogeneity $p=6.36x10^{-03}$, the G allele lead to a stronger decrease of both hsCRP and ESR in males than in females), *CR1* and *HBB* (heterogeneity $p=2.88x10^{-03}$ and heterogeneity $p=7.75x10^{-03}$ respectively, the effect sizes for top SNPs for increased ESR levels were doubled in females than in males).

4.2 Comparison with 1000 Genomes data

To assess the effectiveness of our Sardinian imputation panel, we carried out the same GWAS analyses using the 1000G phase I reference panel for imputation and made a direct comparison of the results obtained using the two panels. Interestingly, not only we were unable to find additional loci, but we missed signals detected using the Sardinian reference panel. [Supplementary Table2 – 1000G GWAS results] At the *ABHD13* locus on chromosome 13 for ADPN and at *AACS* for ESR we missed the signal because the associated variants are absent in 1000G reference panel. For ESR, a significant signal was seen at variants in the downstream region.

Fig1. Regional association plots for *ABHD13* and *AACS*. Regional association plots at the *ABDH13* locus for ADPN, and at *AACS* for ESR for imputation performed using the Sardinian and 1000 Genomes reference panels, respectively. At each locus, we plotted the association strength (Y axis shows the –log 10 pvalue) versus the genomic positions (on 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 r2 values calculated on Sardinian and 1000G haplotypes, for left and right panels, respectively). Symbols reflect genomic functional annotation, as indicated in the inner box of the first plot. 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 LocusZoom package[60].



Marker chr13:108884835 at *ABHD13* is present in the recent release of the UK10K project, where its frequency is extremely low (AC=1), and the variant chr12:125406340 near *AACS* is still missing, so it is unclear whether it is specific to Sardinians or just rarer elsewhere.

Variants in *PDGFRL*, *AACS* (rs183233091) and *CBLN1* are present in 1000G panels but are poorly imputed in our cohort, thus association did not passed the genome-wide significant threshold.

Fig2. Regional association plots for *PDGFRL* and *AACS*. Regional association plots at the *PDGFRL* and *AACS* loci for hsCRP for imputation performed using the Sardinian and 1000 Genomes reference panels, respectively. For a description of the plot style, see Fig1 legend.

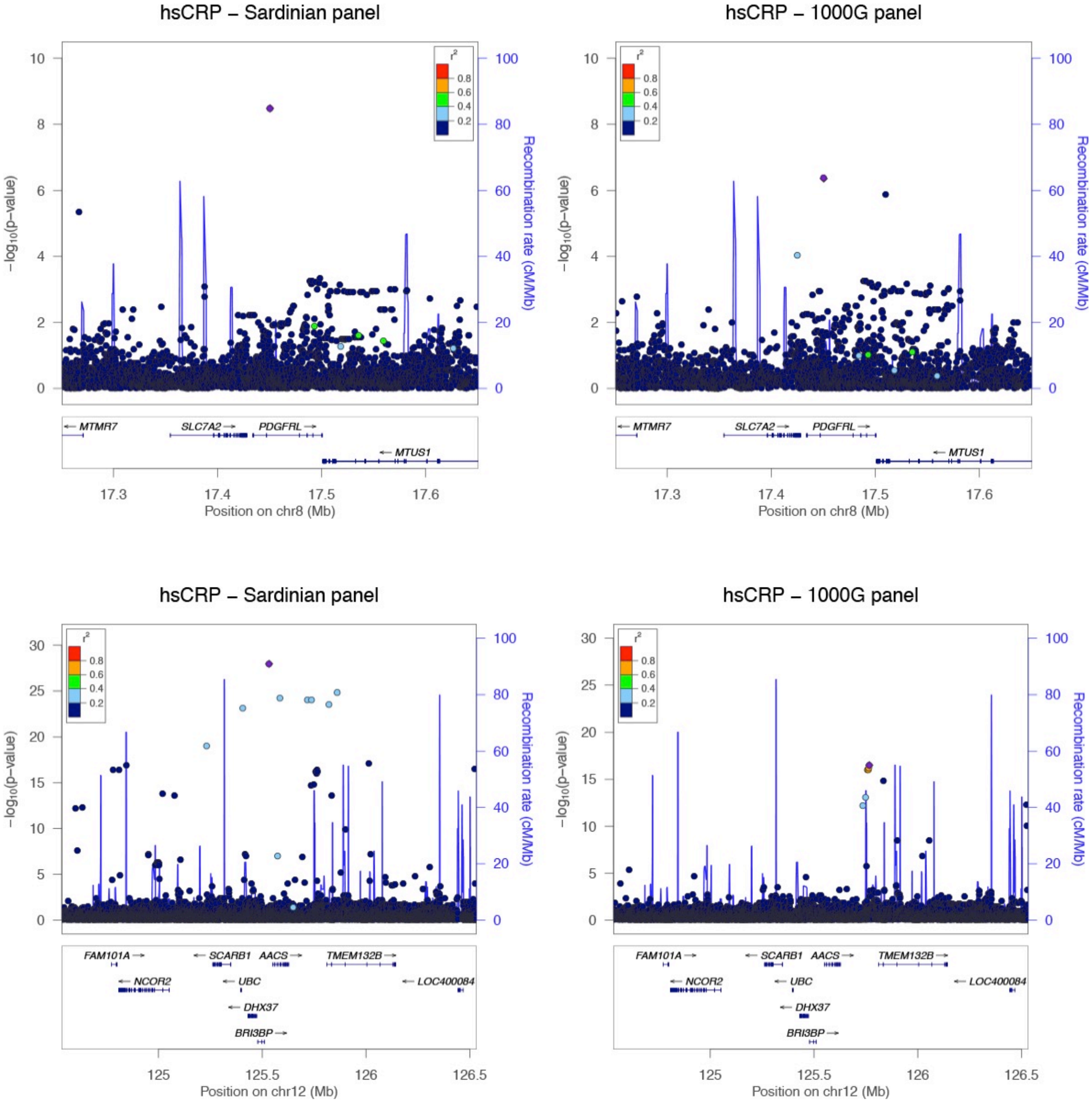
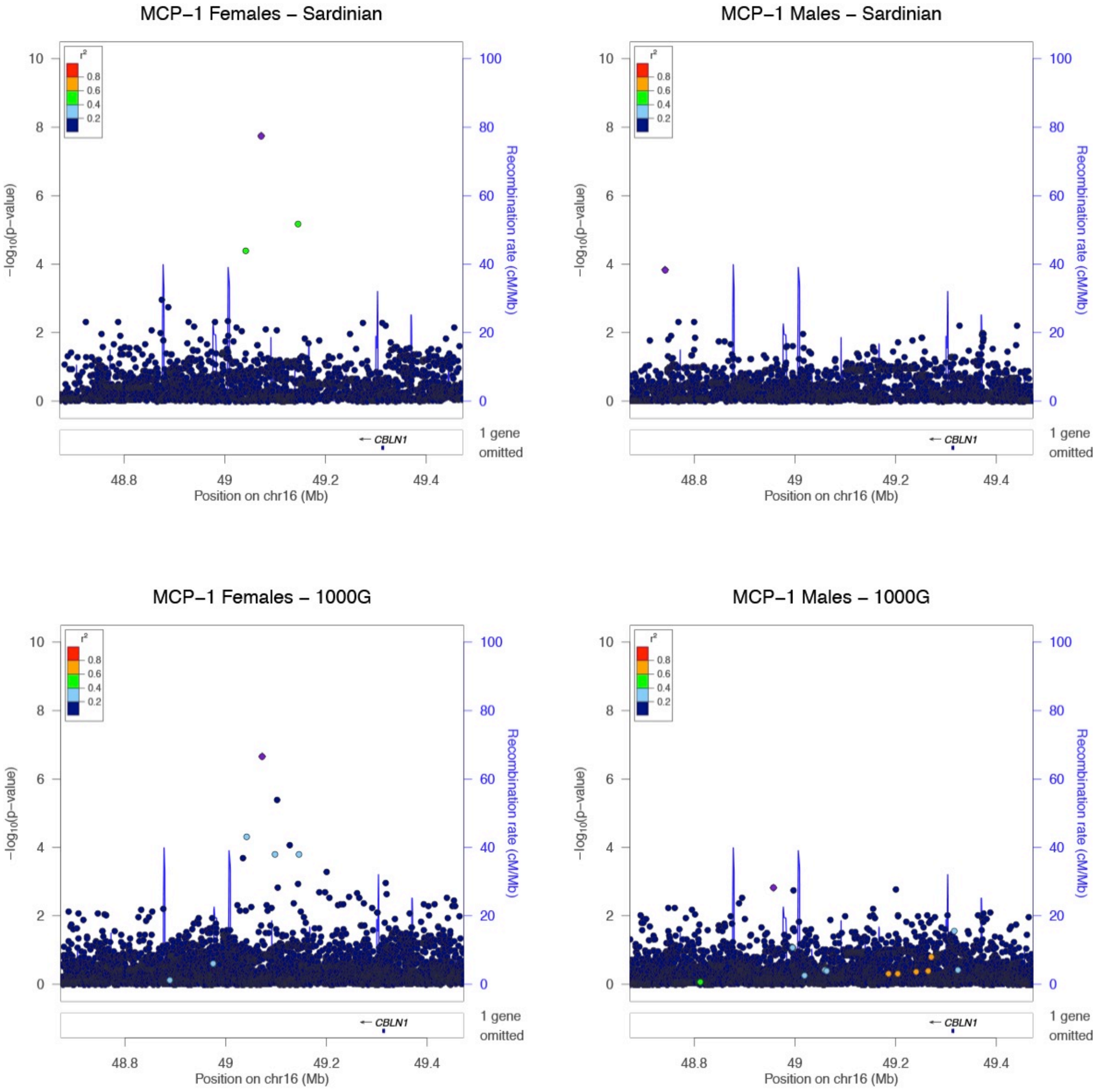
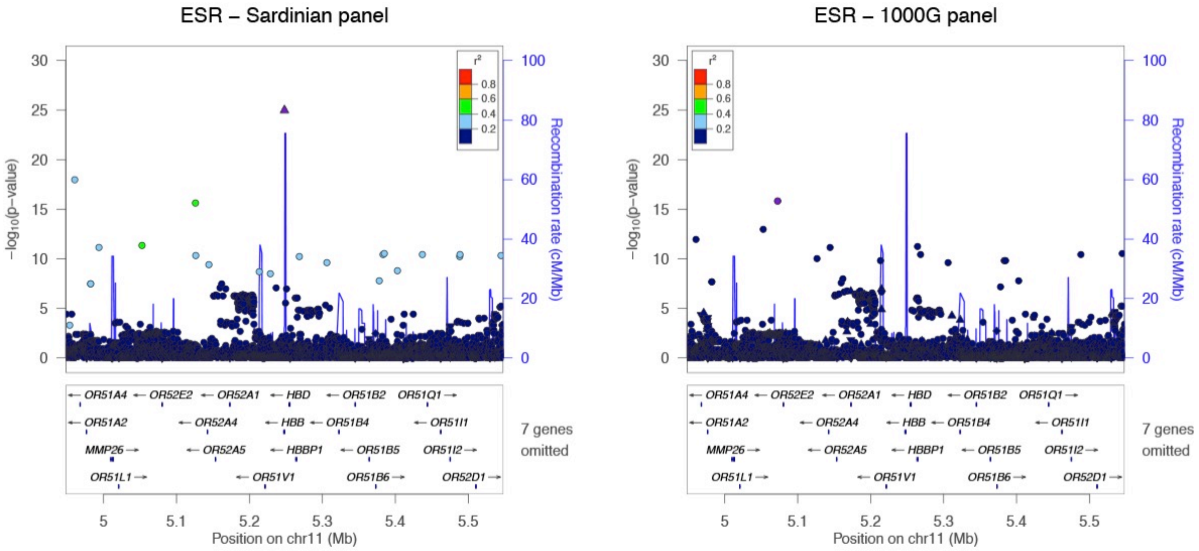


Fig3. Regional association plots for *CBLN1*. Regional association plots at the *CBLN1* for MCP-1 in females and males for imputation performed using the Sardinian and 1000 Genomes reference panels, respectively. For a description of the plot style, see Fig1 legend.



Finally, the association at the *HBB* gene was seen with weaker evidence at marker rs186042619 (p=1.28x10^{-16}) versus p=1.02x10^{-25} at the putative causative variant Q40X with our imputation panel. Indeed, the causative variant Q40X was imputed with poor quality (RSQR=0.31) and incorrect frequency (0.000089), being present in only one haplotype in the reference set and thus difficult to impute.

Fig4. Regional association plots for *HBB*. Regional association plots at the *HBB* locus with ESR levels, using the Sardinian and 1000Genomes reference panels for imputation, respectively. For a description of the plot style, see Fig1 legend.



4.3 Validation of findings

Replication is a key step in GWAS but it becomes unfeasible when the variant to be replicated is extremely rare or population-specific because large cohorts are needed to have enough power to detect association.

Our findings belong to this case, so we validated rather than replicated them. Validation allows to ensure the associations are due to true variation and not to imputation artefacts. Using Sanger sequencing, we validated imputed genotypes, and therefore association, for all of the 5 imputed signals.

Table2. Validation. For each SNP, we show the number of heterozygotes and homozygotes for the reference and alternative alleles that were imputed using the Sardinian panel, the number of these that were validated by Sanger sequencing along with the genotype mismatch rate, the pvalue observed in our primary analysis (as reported in Table 1) and the pvalue obtained replacing imputed genotypes with those derived by Sanger sequencing.

SNP	N hom ref/het/hom alt	N Sanger sequencing			Original Pvalue	Pvalue after validation
		Hom Ref (mismatch %)	Het (mismatch %)	Hom alt (mismatch %)		
13:108884835	5824/12/0	12 (0%)	12 (0%)	0	3.35x10^{-08}	2.84x10^{-08}
8:17450500	5588/42/0	20 (0%)	42 (7%)	0	3.31x10^{-09}	2.65x10^{-09}
12:125533106	5524/105/1	20 (0%)	63 (0%)	1 (100%)	1.09x10^{-28}	1.80x10^{-28}
12:125406340	5864/77/0	20 (0%)	16 (0%)	0	4.40x10^{-23}	4.41x10^{-23}
16:49072490	3312/35/0	21 (0%)	33 (0%)	0	1.76x10^{-08}	1.76x10^{-08}

4.4 Rare variants association

We also assessed global gene effects using two burden tests: the Combined and Multivariate Collapsing (CMC) and the variable thresholds method (VT), both adapted in EPACTS to account for familiar relationship. In each test, we assessed 10,000 regions and thus considered a Bonferroni threshold of 5x10^{-06} to declare significance. Interestingly, four loci were significant for both tests, thus likely representing true signals. In particular, two overlapped with results for single-variant association tests: the *CCR2* gene for MCP-1 and *HBB* gene for ESR.

Table3. Burden test. The table shows results for the rare variants association tests at gene passing the significant threshold for at least on the two statistical tests (CMC and VT). Of note, no significant results were observed for 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 is reported. For the VT we reported the pvalue and the pvalue observed after adjusting for the lead variant at the same or the nearby gene. Specifically, *STAB1* was adjusted for rs7639267; *CCR2* was adjusted for rs113403743 and rs200491743; *IFI16* was adjusted for rs12075, rs2852718 and rs34599082; *HBB* and *OR52H1* were adjusted for rs76728603, and *PTPRH* was adjusted for the best lead in the region (rs7253814). Genes that remain significant after adjustment are marked in bold.

Gene	Chr:Start-end	#SNPs	#Pass	Burden Count	Fraction with rare	CMC test			VT test	
						Effect (StdErr)	Pvalue	Adjusted pvalue	Pvalue	Adjusted pvalue
ADPN										
STAB1	3: 52535766-52558237	25	23	752	0.12886	0.245 (0.039)	4.71x10^{-10}	**1.92x10^{-09}**	1.00x10^{-07}	**1.00x10^{-07}**
MCP1										
CCR2	3: 46399158-46401290	4	3	105	0.01797	0.541 (0.104)	1.84x10^{-07}	0.7092	1.00x10^{-06}	0.92
IFI16	1: 158979950-159024668	10	8	567	0.09702	0.218 (0.046)	2.50x10^{-06}	0.1564	1.40x10^{-05}	0.115
ESR										
HBB	11: 5247914-5248004	2	2	613	0.10318	-0.345 (0.039)	9.77x10^{-19}	0.015	1.00x10^{-07}	0.025
OR52H1	11: 5565906-5566751	5	3	529	0.08904	-0.205 (0.042)	1.23x10^{-06}	0.345	3.40x10^{-06}	0.69
PTPRH	19: 55693244-55716713	22	15	1152	0.19391	-0.146 (0.029)	8.31x10^{-07}	**4.22x10^{-06}**	1.18x10^{-05}	1.90x10^{-05}

Of note, 4 of 6 signals were not a result of cumulative effects of rare variants but they were driven by a single top SNP in the region. Indeed, the signal at the genes *CCR2* and *IFI16* for MCP-1, at *HBB* and *OR52H1* for ESR disappeared when we repeated the test adding the top SNP in the region as covariate.

The other two significant loci include the gene *STAB1* for ADPN, and the *PTPRH* gene for ESR. Of note, *PTPRH* was significant only with the CMC test (p=8.31x10^{-07} and p=1.18x10^{-05} with CMC and VT respectively). *STAB1* acts as a scavenger receptor for acetylated low density lipoprotein, and variants in this gene has been associated to waist-hip ratio[(55)]. The protein encoded by *PTPRH* is a member of the protein tyrosine phosphatase (PTP) family, known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Its relation to ESR is at the moment unclear.

4.5 Variance explained

The variance explained calculated by using the top variants is higher for all traits when using Sardinian rather than 1000G reference panel, with the exception of IL-6 where the same variants were detected using both panels and thus the variance explained did not change. Thus, the Sardinian reference panel provided more precise information than a general, freely available panel from multiple populations. We confirmed the higher efficiency of Sardinian-imputed results when we estimated the variance explained by all the 13.6 SNPs successfully genotyped or imputed.

Table4. Variance explained. For each of the five inflammatory markers, the table shows the heritability, the amount of phenotypic variance explained by the top signals identified using the Sardinian and the 1000 Genomes reference panel, as well as the variance explained by all variants in the accessible genome when the Sardinian or 1000 Genomes imputed data are considered.

Trait	**h^2**	**top hits**			**All SNPs**	
		Sardinian	**1000G**		**Sardinian**	**1000G**
ADPN	39.2	2.37	1.97		21.71	20.39
hsCRP	25.1	6.35	5.02		24.73	21.96
ESR	43.1	4.58	3.39		24.92	24.36
MCP-1	31	11.80	11.25		13.26	11.00
IL-6	15.3	2.49	2.49		5.46	3.39

5. Whole-genome sequence-based analysis of thyroid function

Levels of thyroid hormones are tightly regulated by TSH produced in the pituitary, and even mild alterations in their concentrations are strong indicators of thyroid pathologies, which are very common worldwide. In the last years, genome-wide association studies (GWAS) have identified several susceptibility loci for thyroid function markers(56,57,58). However, to gain insights into increasingly more modest signals of association, samples of many thousands of individuals are required. One approach to overcome this problem is to combine the results of GWAS from closely related populations via meta-analysis. The most recent meta-analysis(56) conducted in 26,420 individuals, identified 19 loci associated with TSH and 4 with fT4 explaining only 5.6% and 2.3% of the variance for TSH and fT4 respectively. To identify additional common (MAF>=1%) variants associated with TSH and FT4 levels, we carried out a two-stage meta-analysis of genome-wide association results from whole-genome sequence and deeply imputed datasets.

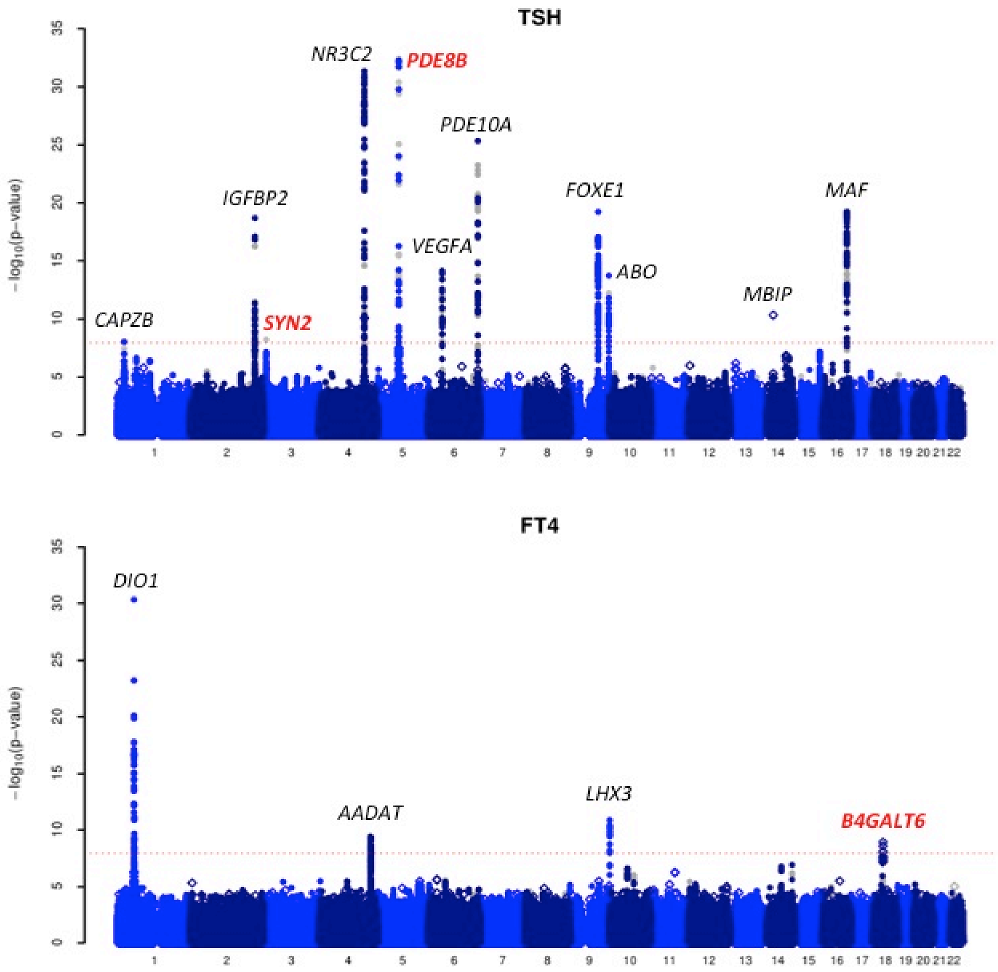
5.1 Results

In the stage1, using a meta-analysis of WGS data from the ALSPAC and TwinsUK cohorts (N=2,287) we analyzed up to 8,816,734 markers and we found associations at 2 previously described loci for TSH. These are *NR3C2* (rs11728154; p=8.21x10^{-09}; r^2=0.99 with the previously reported rs10028213) and *FOXE1* (rs1877431; p=2.29x10^{-10}; r2=0.99 with the previously reported rs965513). We found one borderline signal at a novel locus *FAM222A* (rs11067829; p=3.73x10^{-08}).

No variants showed genome-wide significant association for FT4.

In the stage2, we conducted a meta-analysis of the stage1 cohorts and 5 additional cohorts (TwinsUK GWAS, ALSPAC GWAS, SardiNIA, ValBorbera and Busselton Health Study (BHS)) and we found associations for 13 SNPs at 11 loci for TSH (N=16,335) and 4 SNPs at 4 loci for FT4 (N=13,651).

Fig5. Manhattan plot. Annotated Manhattan plot from the overall analysis for TSH and FT4. SNPs (MAF>1%) are plotted on the X axis according to their position on each chromosome against association with TSH on the Y axis (shown as–log10 P value). The loci are regarded as genome-wide significant at P<5x10^{-8}. Variants with 1%<MAF <5% are show as open diamond symbols. Common SNPs (MAF>5%) are shown as sold circles with those present in Hapmap II reference panels in grey and those derived from WGS or deeply imputed using WGS and 1000 genomes reference panels in blue. Genes labeled in red represent novel genome-wide significant (P<5x10^{-8}) findings.



For TSH, 11 top SNPs represent known signals: our top variants are in strong LD (r^2>0.6) with those described in previous studies.

Two SNPs associated are novel: one at *SYN2* (rs310763; p=6.15x10^{-09}). *SYN2* is a member of a family of neuron-specific phosphoproteins involved in the regulation of neurotransmitter release with expression in the pituitary and hypothalamus.

With a conditional analysis on *PDE8B* we identified a novel variant (rs2928167; p=5.94x10^{-14}) in linkage equilibrium (r^2=0.002) with the previously described variant rs688509910 and independent from our top SNP rs2046045 (p=1.93x10^{-11} after conditional analysis).

In the overall meta-analysis we were unable to replicate the association between *FAM222A* and TSH detected in the stage1 (p=0.378); however, we observed evidence of heterogeneity between cohorts (p heterogeneity=4.70x10^{-06}), so potentially this locus may find support in future WGS studies.

For FT4, we confirmed 3 known associated loci (*DIO1*, *LHX3* and *AADAT*) and we found one novel uncommon variant (MAF=3.2%) at *B4GALT6/SLC25A52* (rs113107469; p=1.27x10^{-09}).

B4GALT6 is in the ceramide metabolic pathway, which inhibits cAMP production in TSH-stimulated cells. However rs113107469 is in weak LD (r^2<0.1) with the Thr139Met substitution (rs28933981; MAF=0.4%) and it may therefore be a marker for this functional change in TTR. The Thr139Met substitution was associated with FT4 levels in our single-point meta-analysis (p=2.14x10^{-11}), however was not originally observed as the MAF was lower than our 1% threshold. Conditional analysis of the TTR region using rs28933981 as the conditioning marker in the ALSPAC WGS cohort reveals no evidence of association between rs113107469 in *B4GALT6* and FT4 (p=0.124). Analysis using direct genotyping in the ALSPAC WGS and GWAS cohorts confirms the effect of the Thr139Met substitution on FT4 levels. Here, 0.79% of children were heterozygous for the Thr139Met substitution, which is positively associated with FT4 (p=3.89x10^{-24}). In the ALSPAC GWAS dataset, rs113107469 in *B4GALT6* was also positively associated with FT4 (p=0.0002); however, when conditioned on the Thr139Met substitution there was no longer any evidence of association (p=0.20). The Thr139Met substitution also appears to be functional: this mutation has increased protein stability compared with wild-type TTR and tighter binding of thyroxine, resulting in a two-fold increase in thyroxine binding affinity.

Of all 17 independent markers, significant evidence for heterogeneity (p<0.003, corresponding to a Bonferroni threshold of 0.05/17) was observed at *FOXE1* (p=2.02x10^{-06}) and *ABO* (p=4.11x10^{-04}).

Table1. Table shows the association results for SNPs that reached genome-wide level significance in the final meta-analysis. For each SNP, the best candidate gene is showed, as well as its genomic position, the effect allele (A1), the other allele (A2), the combined frequency of A1 across studies (Freq A1) the effect size (Beta - change in standardized thyroid measure by allele) and its standard error (Std Err), the p-value for association (P), the number of samples analyzed (N) and the p-values for heterogeneity of effects across the cohorts used in the meta-analysis (Het P). Entries in bold reflect novel identified SNPs

Gene	SNP	Chr	Position	A1/A2	Freq A1	Effect	Std Err	N	P	Het P
TSH										
CAPZB	rs12410532	1	19845279	T/C	0.164	-0.090	0.016	16,332	9.41 x10^{-09}	0.003
IGFBP2	rs7568039	2	217612321	A/C	0.250	-0.122	0.014	16,335	2.11 x10^{-19}	0.370
SYN2	**rs310763**	**3**	**12230704**	**T/C**	**0.235**	**0.083**	**0.014**	**16,334**	**6.15 x10^{-09}**	**0.252**
NR3C2	rs28435578	4	149646538	C/T	0.227	-0.166	0.014	16,333	4.59 x10^{-32}	0.109
PDE8B	rs2046045	5	76535811	G/T	0.414	0.142	0.012	16,334	4.05 x10^{-33}	0.653
PDE8B	**rs2928167**	**5**	**76477820**	**G/A**	**0.104**	**-0.145**	**0.019**	**16334**	**5.94 x10^{-14}**	**0.994**
VEGFA	rs6923866	6	43901184	C/T	0.280	-0.102	0.013	16,333	7.55 x10^{-15}	0.646
VEGFA	rs2396084	6	43804825	A/G	0.287	-0.096	0.013	16,333	4.33 x10^{-13}	0.422
PDE10A	rs3008034	6	166043862	C/T	0.312	-0.131	0.012	16,335	4.68 x10^{-26}	0.084
FOXE1	rs112817873	9	100548934	T/A	0.323	-0.14	0.015	11,544	6.15 x10^{-20}	2.02x10^{-6}
ABO	rs116552240	9	136149098	A/T	0.239	0.121	0.016	14,047	1.92 x10^{-14}	4.11x10^{-4}
MBIP	rs116909374	14	36738361	T/C	0.043	-0.208	0.032	15,037	4.69 x10^{-11}	0.179
MAF	rs17767742	16	79740541	G/C	0.354	-0.113	0.012	16,335	5.64 x10^{-20}	0.447
FT4										
DIO1	rs2235544	1	54375570	A/C	0.499	0.154	0.013	13,650	5.23 x10^{-34}	0.084
AADAT	rs7694879	4	170969799	T/C	0.095	0.137	0.022	13,650	4.15 x10^{-10}	0.168
LHX3	rs11103377	9	139097135	G/A	0.496	0.087	0.013	13,651	1.44 x10^{-11}	0.735
B4GALT6	**rs113107469**	**18**	**29306737**	**T/C**	**0.032**	**0.223**	**0.037**	**13,649**	**1.27 x10^{-11}**	**0.574**

5.2 Rare variants association

In the meta-analysis we analyzed only low frequency and common SNPs, excluding those with MAF<1%. To analyse rare variants we performed sequence kernel based association testing (SKAT) analysis using only individuals with WGS data.

We found no evidence of association with TSH, however for FT4 we identified one SKAT bin with multiple-testing corrected evidence for association in *NRG1* (p=2.53x10^{-06}). NRG1 is a glycoprotein that interacts with the NEU/ERBB2 receptor tyrosine kinase, and is critical in organ development.

5.3 Variance explained

To evaluate the improvement in variance explained given by our study, we compared the variance explained calculated by using all known SNPs from previous studies and then adding the novel top hits. We performed the analysis in ALSPAC WGS, TwinsUK WGS, BHS and SardiNIA and then combined the results with a fixed-effects meta-analysis.

We observed a small improvement: our estimates improved from 7.2% to 8% for TSH and from 1.8% to 2% for FT4.

Furthermore, we calculated that low frequency and common variants (MAF>=1%) collectively account for over 20% of the variance in TSH and FT4; a substantial advance on using only the top hits from GWAS meta-analysis.

6. Concluding Remarks

Advances in sequencing technologies revolutionized genetic studies of complex traits allowing the analysis of variants across the entire allele frequency spectrum.

Whole-genome sequencing provides many advantages over array-based genotyping for GWAS. Most importantly, the possibility to assess extremely large number of markers and even to discover new variants not yet classified in public databases augments the chance to directly assess the causal allele, with consequently higher statistical power and precision of results. Furthermore, it allows to compare association of genetic variations in different populations avoiding heterogeneity and lack of replication due to differences in linkage disequilibrium, important issue to consider when analyzing a fixed subset of markers from the genome.

In this work, I showed how single analysis and meta-analysis of sequencing-based GWAS improve the current knowledge of genetic variation associated to important human traits. In two different study designs with different statistical approaches, whole-genome sequencing integrated with genotyping arrays by statistical inference, led to the identification of novel common, low frequency and rare variants associated with levels of five inflammatory biomarkers and with two parameters related to thyroid function.

These two studies highlighted not only advantages but also current pitfalls of the sequencing based GWAS approach. One is related to the statistical methods utilized. In fact, although sequencing provides the opportunity to investigate the roles of low-frequency and rare variants in complex traits, a debate still exists on the optimal statistical method to be used for such variants. Given that the relative performance of these methods depends on the underlying genetic architectures of complex traits, which is unknown, it is difficult to have a test that is optimal for all scenarios.

Another limitation of the sequencing-based GWAS approach is the difficulty in replication of the association results. When the number of sequenced individuals increases, the proportion of low-frequency SNPs dramatically increases. A large proportion of rare variants are private to specific populations, absent in any commercial SNP array and even in large, public repositories, as dbSNP, so the associations hardly can be replicated. Furthermore, as the power of association tests is a function of the allele frequency, replication of rare-variant associations requires very large sample, with numbers that become unachievable for very rare or population-specific alleles.

Finally and contrary to expectations, the missing heritability will be not easily accounted for. In fact, because power to detect low-frequency and rare variant associations is lower than the power to detect common-variant associations, the observed proportion of heritability due to low frequency and rare variants in finite samples might be substantially underestimated. Indeed, even if these variants account for a large proportion of heritability, identifying them might require extremely large samples. For example, it has been demonstrated that when rare-variant association studies are carried out in a sample of 10,000 individuals, most rare causal variants will show no significant association(59). In this case, the apparent proportion of variance due to rare variants might be <0.1%, even when rare variants actually explain most of the heritability. The number of rare causal variants significantly associated increases with the sample sizes, and the variance explained by rare alleles become closer to the true value. Still, even after 1,000,000 individuals are studied, the estimated proportion of variance due to rare variants remains underestimated. Currently, there is no clear evidence as to which scenario represents the true genetic architecture of common complex diseases, and it is likely to vary across diseases and traits.

Nevertheless, only a few sequencing based GWAS studies have been published so far, and we expect this number to increase substantially as it has been for the HapMap GWAS studies. Therefore, despite the discussed limitations of the approach, their results will be valuable. They will enlarge our current knowledge of genes associated to traits and diseases, highlight novel biological pathways and elucidate underlying mechanisms, and suggest critical points and issues to be considered in further developments and improvements of necessary statistical methods.

Appendix

Appendix Table 1. Gender specific effects at variants associated with inflammatory markers. The table shows the association parameters of the SNPs listed in **Table 1** when analysed in males and females separately. Columns are defined as in **Table 1**. SNPs showing significant heterogeneity (HetPval column) between genders are marked in bold.

Chr:position	rs name	Candidate Gene	Effect Allele / Other	Males			Females			HetPVal
				Freq	Effect (StdErr)	pvalue	Freq	Effect (StdErr)	pvalue	
	ADPN (2486 males/3350 females)									
3:186559460	rs17300539	*ADIPOQ*	A/G	0.165	0.277 (0.038)	5.15x10^{-13}	0.15	0.256 (0.034)	1.28x10^{-13}	0.687
13:108884835	N/A	*ABDH13*	A/G	0.001	-1.982 (0.521)	1.44x10^{-04}	0.001	-1.426 (0.350)	4.68x10^{-05}	0.377
	hsCRP (2411 males/3219 females)									
1:159684665	rs3091244	*CRP*	A/G	0.414	0.196 (0.028)	2.46x10^{-12}	0.439	0.229 (0.025)	3.11x10^{-20}	0.393
8:17450500	rs73198138	*PDGFRL*	A/G	0.005	-0.899 (0.203)	9.55x10^{-06}	0.003	-0.884 (0.213)	3.38x10^{-05}	0.96
12:125533106	**rs183233091**	***BRI3BP, AACS***	**A/G**	**0.01**	**1.308 (0.134)**	**3.57x10^{-22}**	**0.01**	**0.807 (0.125)**	**1.25x10^{-10}**	**6.36x10^{-03}**
12:121415293	*rs7139079*	*HNF1A*	G/A	0.375	-0.127 (0.029)	9.60x10^{-06}	0.379	-0.134 (0.025)	1.19x10^{-07}	0.849
19:45411941	rs429358	*APOE, APOC1, APOC1P1*	C/T	0.076	-0.248 (0.051)	1.53x10^{-06}	0.071	-0.228 (0.047)	1.02x10^{-06}	0.774
	ESR (2531 males/3410 females)									
1:25724005	rs71721472	*RHCE, TMEM57*	T/C	0.305	-0.095 (0.032)	2.65x10^{-03}	0.291	-0.161 (0.028)	1.45x10^{-08}	0.123
1:207684359	**rs11117956**	***CR1***	**T/G**	**0.397**	**-0.102 (0.029)**	**3.65x10^{-04}**	**0.403**	**-0.215 (0.025)**	**8.51x10^{-18}**	**2.88x10^{-03}**
11:5248004	**rs76728603**	***HBB***	**A/G**	**0.05**	**-0.348 (0.067)**	**2.52x10^{-07}**	**0.046**	**-0.589 (0.060)**	**1.31x10^{-22}**	**7.75x10^{-03}**
12:125406340	N/A	*AACS, MIR5188*	G/A	0.008	1.260 (0.152)	2.00x10^{-16}	0.006	0.806 (0.164)	9.45x10^{-07}	0.044
	MCP-1 (2497 males/3347 females)									
1:159175354	rs12075	*DARC*	G/A	0.447	-0.415 (0.029)	4.91x10^{-46}	0.445	-0.408 (0.024)	3.15x10^{-60}	0.859
1:159164454	*rs2852718*	*CADM3*	C/T	0.021	-0.223 (0.103)	0.03	0.023	-0.450 (0.084)	8.33x10^{-08}	0.087
1:159175494	*rs34599082*	*DARC*	T/C	0.037	-0.068 (0.079)	0.39	0.037	-0.162 (0.066)	0.014	0.359
3:46383906	rs113403743	*CCR2, CCR3*	T/G	0.1	0.270 (0.050)	7.00x10^{-08}	0.098	0.273 (0.043)	3.65x10^{-10}	0.959
3:46399764	*rs200491743*	*CCR2*	A/T	0.005	1.115 (0.199)	2.50x10^{-08}	0.006	0.516 (0.170)	2.40x10^{-03}	0.022
16:49072490	**rs76135610**	***N4BP1, CBLN1***	**T/C**	**0.005**	**0.969 (0.172)**	**1.76x10^{-08}**	**0.006**	**0.286 (0.192)**	**0.1378**	**8.06x10^{-03}**
	IL-6 (2492 males/3346 females)									
1:154428283	rs12133641	*IL6R*	G/A	0.258	0.123 (0.030)	3.98x10^{-05}	0.253	0.117 (0.026)	9.58x10^{-06}	0.885
9:136142355	rs643434	*ABO*	A/G	0.267	-0.223 (0.030)	8.29x10^{-14}	0.26	-0.218 (0.026)	2.18x10^{-16}	0.9

Appendix Table 2. Association signals based on 1000G imputation for the inflammatory markers. The table reports top association signals identified with 1000G imputation. Columns are the same as defined in Table 2. Signals at novel loci are in bold. Independent signals, indicated in italics, are reported along with the regression coefficients from the conditional analysis.

SNP	rs name	*Nearest Gene*	Effect allele / Other	Freq	pvalue	Effect (StdErr)	RSQR	Type
ADPN								
3:186562865	rs73185702	*ADIPOQ*	A/G	0.159	7.31x10^{-23}	0.249 (0.025)	0.98	Intronic
hsCRP								
1:159684665	rs3091244	*CRP*	A/G	0.428	5.28x10^{-27}	0.207 (0.019)	Genotyped	intergenic
12:121423659	rs9738226	*HNF1A*	A/G	0.355	2.65x10^{-09}	-0.120 (0.020)	0.998	Intronic
12:125766568	**rs142361132**	***TMEM132B***	**T/C**	**0.011**	**3.05x10^{-17}**	**0.762 (0.090)**	**0.92**	**intergenic**
19:45411941	rs429358	*APOE*	C/T	0.073	2.14x10^{-11}	-0.240 (0.036)	0.99	nonsyn
ESR								
*1:25769212**	*rs36055238*	*TMEM57*	*I/R*	*0.316*	*3.70x10^{-08}*	*-0.110 (0.020)*	*0.917*	*intergenic*
1:207690871	rs10863358	*CR1*	C/G	0.401	8.53x10^{-18}	0.153 (0.018)	0.997	intronic
11:5072356	rs186042619	*OR52J3*	G/A	0.048	1.28x10^{-16}	-0.349 (0.042)	0.894	Intergenic
12:125835147	**rs75220528**	***TMEM132B***	**C/G**	**0.01**	**4.14x10^{-15}**	**0.703 (0.089)**	**0.847**	**intronic**
MCP-1								
1:159175354	rs12075	*DARC*	A/G	0.446	1.19x10^{-96}	-0.405 (0.019)	0.999	nonsyn
1:159162174	*rs2814767*	*CADM3*	*G/T*	*0.023*	*7.82x10^{-17}*	*-0.524 (0.063)*	*0.973*	*intronic*
1:159175494	***rs34599082***	**DARC**	***C/T***	***0.037***	***6.62x10^{-12}***	***-0.340 (0.049)***	***0.998***	***nonsyn***
3:46391788	rs17141006	*CCR2*	T/G	0.099	2.67x10^{-15}	0.269 (0.034)	1	intergenic
IL6								
1:154428283	rs12133641	*IL6R*	G/A	0.25	3.92x10^{-09}	0.123 (0.021)	0.974	intronic
9:136143120	rs613534	*ABO*	G/A	0.262	5.50x10^{-27}	-0.221 (0.020)	0.998	intronic

Description of meta-analysis participating cohorts

Cohorts:

Seven populations were used in this study. They are known as the TwinsUK WGS cohort, the TwinsUK GWAS cohort, the Avon Longitudinal Study of Parents and Children (ALSPAC) WGS cohort, the ALSPAC GWAS cohort, the SardiNIA cohort, the ValBorbera cohort and the Busselton Health Study (BHS) cohort.

All human research was approved by the relevant institutional ethics committees.

Cohorts description:

Twins UK: The Twins UK cohort consists of 12,000 twins of northern European/UK ancestry, aged 16–82 yr, from St Thomas' UK Adult Twin Registry (TwinsUK), a volunteer sample recruited in the United Kingdom without selection for particular traits (www.twinsuk.ac.uk/). It has previously been shown to be representative of singleton populations and the UK population in general.

ALSPAC: ALSPAC is a geographically based UK cohort that recruited pregnant women residing in Avon (Southwest England) with an expected date of delivery between April 1, 1991, and December 31, 1992. A total of 15,247 pregnancies were enrolled, with 14,775 children born (see www.alspac.bris.ac.uk.). The ALSPAC cohort was the only child cohort used in this study however both the hypothalamic-thyroid axis and the impact of thyroid status on metabolism is considered to be generally comparable between children and adults.

SardiNIA: see *The SardiNIA Project* section.

ValBorbera (INGI): The Val Borbera (INGI) population is a collection of 1,785 genotyped samples collected in the Val Borbera Valley, a geographically isolated valley located within the Appennine Mountains in NorthWest Italy. The valley is inhabited by about 3000 descendants from the original population, living in 7 villages along the valley and in the mountains. The valley was inhabited by about 10,000 people in the 19th century when endogamy was >80%. Participants were healthy people between 18 and 102 years of age that had at least one grandfather living in the valley.

Busselton: The Busselton Health Study (http://bsn.uwa.edu.au) includes a series of cross-sectional health surveys carried out since 1966 of residents of Busselton, a rural town with a predominantly Caucasian population, located in the southwest of Western Australia 36. In 1994-5, there was a follow-up study of people who had participated in previous studies. Participants completed a health questionnaire, underwent physical examination, and gave a venous blood sample in the morning after an overnight fast.

Bibliography

1. Collins FS, Morgan M, Patrinos A The Human Genome Project: Lessons from large-scale biology. *Science* 2003, 300: 286–90.

2. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449:851–861.

3. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012, 491, 56–65.

4. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., and Hoh, J. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005, 308:385-389.

5. Wood AR, Esko T, Yang J, Vedantam S et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2004, 46(11):1173-86.

6. 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, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. *Nature* 2009, 461:747-753

7. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010, 11: 446–450.

8. Slatkin M. Epigenetic inheritance and the missing heritability problem. Genetics 2009, 182: 845–850

9. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ* 1995;310:170.

10. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 2011;187(2):367–383. doi: 10.1534/genetics.110.120907.

11. Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., Greenwood, C. M. T. Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies. *Genet. Epidemiol.* 2014, 38: 281–290

12. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10:387–406.

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

14. Hara K, Fujita H, Johnson TA et al. Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet* 2014;23:239–46.

15. Huang X., Feng Q., Qian Q., Zhao Q., Wang L., Wang A., et al. High-throughput genotyping by whole-genome resequencing. *Genome Res* 2009, 19, 1068–1076.

16. Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, et al. (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature* 2011, 480: 201–208

17. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, Zhai G, Zhao JH, Smith AV, Huffman JE, Albrecht E, Jackson CM, Evans DM, Cadby G, Fornage M, Manichaikul A, Lopez LM, Johnson T, Aldrich MC, Aspelund T, Barroso I, Campbell H, Cassano PA, Couper DJ, Eiriksdottir G, Franceschini N, Garcia M, Gieger C, Gislason GK, Grkovic I, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011;43:1082–1090.

18. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Consortium BIRAC, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJ, Costenbader KH, Criswell LA, Crusius JB, Cui J, de Bakker PI, De Jager PL, Ding B, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet.* 2010;42(6):508–514.

19. Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10, 191–201 (2009).

20. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;10: 101-29.

21. International Multiple Sclerosis Genetics C, Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* (2013) 45(11):1353–60.10.1038/ng.2770

22. M.A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C.K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, 43 (2011), pp. 1066–1073

23. Sanna S, Li B, Mulas A, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* 2011; 7:e1002198.

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

25. A.P. Morris, E. Zeggini - An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, 34 (2010), pp. 188–193

26. M. Zawistowski, S. Gopalakrishnan, J. Ding, Y. Li, S. Grimm, S. Zöllner - Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.*, 87 (2010), pp. 604–617

27. B.E. Madsen, S.R. Browning - A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5 (2009), p. e1000384

28. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *J. Hum. Genet.* 2008;83:311–321.

29. Price AL, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 2010;86:832–838.

30. S. Morgenthaler, W.G. Thilly - A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, 615 (2007), pp. 28–56

31. M.C. Wu, S. Lee, T. Cai, Y. Li, M.C. Boehnke, X. Lin - Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89 (2011), pp. 82–93

32. S. Lee, M.C. Wu, X. Lin - Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13 (2012), pp. 762–775

33. 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

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

35. 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.

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

37. 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.

38. 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.

39. 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.

40. Howie B, Fuchsberger C, Stephens M, Marchini J, and Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 2012

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

42. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC *Bioinformatics*. 2010;11:288.

43. 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.

44. 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.

45. J. Marchini, B. Howie, S. Myers, G. McVean and P. Donnelly (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* 39 : 906-913

46. Xiang Zhou and Matthew Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. 44: 821–824

47. Aulchenko YS, Struchalin MV, van Duijn CM (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* 11: 134

48. Voorman, A., Brody, J. and Lumley, T. . SkatMeta: an R Package for meta analyzing region-based tests of rare DNA variants. (http://cran.r-project.org/web/packages/skatMeta (2013)).

49. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88, 76-82 (2011).

50. Shanker J, Kakkar VV. Implications of genetic polymorphisms in inflammation-induced atherosclerosis. Open Cardiovasc Med J. 2010;4:30–37.

51. Raman D, Sobolik-Delmaire T, Richmond A. Chemokines in health and disease. *Exp Cell Res*. 2011;317:575–589.

52. 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

53. Kullo, I. J. et al. Complement receptor 1 gene variants are associated with erythrocyte sedimentation rate. *Am. J. Hum. Genet*. 89, 131–138 (2011).

54. 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

55. 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, Kilpeläinen TO, Kutalik Z, Li S, Monda KL, Dixon AL, Holmes CC, Kaplan LM, Liang L, Min JL. 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;42:949–960. doi: 10.1038/ng.685

56. Porcu E, Medici M, Pistis G, Volpato CB, Wilson SG, Cappola AR, et al. 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.

57. Arnaud-Lopez, L. *et al.* Phosphodiesterase 8B gene variants are associated with serum TSH levels and thyroid function. *Am J Hum Genet* 82, 1270-80 (2008).

58. Gudmundsson, J. *et al.* Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nat Genet* 44, 319-22 (2012).

59. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014 Jul 3;95(1):5-23. doi: 10.1016/j.ajhg.2014.06.009.

60. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–7 (2010).