A bird’s-eye view of Italian genomic variation through whole-genome sequencing

Title

A bird's eye view of Italian genomic variation and deleterious variants pattern

Cocca Massimiliano[1], Barbieri Caterina[2], Concas Maria Pina[1,3], Gandin Ilaria[1], Brumat Marco[1,3], Robino Antonietta[1], Trudu Matteo[7], Sala Cinzia[2], Vuckovic Dragana[4], Girotto Giorgia[1,3], Matullo Giuseppe[5], Polasek Ozren[6], Ivana Kolčić[6], Paolo Gasparini[1,3], Soranzo Nicole[4], Toniolo Daniela[2], Massimo Mezzavilla[1]

Affiliations

1) Institute for Maternal and Child Health IRCCS Burlo Garofolo, Trieste, Italy.

2) Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan, Italy.

3) Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste, Italy.

4) Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.

5) Department of Medical Sciences, University of Turin, Turin, Italy

6) Public Health, University of Split, Croatia

7) Molecular Genetics of Renal Disorders Unit, Division of Genetics and Cell Biology, San Raffaele Scientific Institute, Milan

Abstract

The genomic variation in the Italian peninsula populations is currently under represented: the only Italian whole genome reference are the Tuscans from the 1000 Genome Project. To address this issue, we sequenced a total of 947 Italian genomes from three different geographical areas that could be representative of a large portion of the whole country genomic pool. First, we defined a new Italian Genome Reference Panel (IGRP) for imputation, which showed high-performance, especially for rare variants imputation, and we subsequently validated it by GWAS analysis. Furthermore, we widened the catalogue of genetic variation and investigated population structure, pattern of natural selection, distribution of deleterious variants and human knockouts

(HKOs). All the results emphasise a high level of genomic differentiation between populations, diverse signatures of natural selection and a distinctive distribution of deleterious variants and HKO, confirming the necessity of multiple genome references for the Italian population.

Introduction

Large sequencing projects have identified the majority of common variants and millions of rare and low frequency variants (Gudbjartsson et al., 2015; The 1000 Genomes Project Consortium, 2015; The UK10K Consortium, 2015). Most of the rare variants were detected in protein coding genes and it was calculated that each individual may carry more than 20.000 variants per exome (Karczewski et al., 2017; The ENCODE Project Consortium, 2007), a finding that complicates our understanding of gene function since only few genes may underline a disorder or be associated with a given phenotype. The filtering of candidate variants by frequency in unselected individuals is a key step in any pipeline for the discovery of causal variants. The efficacy of such filtering depends on both the size and the ancestral diversity of the available reference data. From this point of view, the catalogue of rare and low frequency variants is still largely incomplete, and its completion will represent a major challenge.

In the available human genome reference sequence data sets (i.e 1000G PH3, ExAC databases, etc.), Southern European populations, which represent a significant proportion of the overall European populations, are highly underrepresented (i.e. only a small group of subjects from Tuscany, Italy, and Spain). To fill this gap, we obtained whole genomes from founder populations - for which the presence of stratification (Esko et al., 2013; Sazzini et al., 2016) and the different level of isolation were demonstrated (Xue et al., 2017) - localized in three different parts of Italy: North-West (Val Borbera), North-East (Friuli Venezia Giulia) and South-East (Carlantino). In founder populations, variants that are rare or absent elsewhere can occur at higher frequencies and overcome the difficulty of identifying rare and low frequency variants. In this respect, our Italian genomes could be also extremely useful for the genetic analysis of other Italian and South European populations. An Italian Reference Genome panel for imputation

was also developed, tested and validated with GWAS analysis for red blood cells parameters and results were compared with those previously obtained using the 1000G data imputation panel only. Our work aims to answer the following questions: *1) Are we able to increment the catalogue of genotypic variation, possibly in the low frequency spectrum, with new data? 2) Do we add useful information in terms of genetic variability, non-redundant with respect to the South European-Italian data already present in the commonly used reference panels? 3) Will we be able to identify new loci/variants, characteristic of a South-European subpopulation through GWAS using the new reference panel for imputation? 4) How much homogeneous are genomes coming from different regions of Italy in terms of population structure, natural selection signatures, deleterious variants distribution and human knockouts (HKO)? And, as a consequence, how reliable is to use only one reference population for Italians such as Tuscans?*

Results

WGS data generation: variant calling and quality control

A total of 947 DNAs from three cohorts were sequenced at 6 to 10X coverage; 381 individuals from Friuli Venezia Giulia (FVG), 433 from Val Borbera (VBI) and 133 from Carlantino (CAR) (**Figure 1a).** Genotype calls for autosomal chromosomes were produced separately for each population. After filtering, 926 samples were retained. Approximately 27M sites (i.e. >24M SNVs and >2M indels) were detected (**Table 1**) in the joint dataset. Overall, 7.1 M sites (26%) were common (MAF>5%), 3.1M (12%) were low frequency (MAF between 1% and 5%) and 16.6M (62%) were rare (MAF <1%) with a similar partition in all cohorts. Singletons variants (AC=1) were >6M (24%) (**Table 1 and Figure 1 b)**. For each individual, we identified on average ~3.5M variant sites including ~0.56M indels and ~7.000 singletons. Considering each cohort separately, we noticed an excess of singletons in Carlantino cohort (CAR): most of them were shared with the other INGI cohorts, confirming that this is an artefact due to the lower sample size (124 samples, after QC). The comparison with outbred references (EUR subset from 1000G Phase 3, the whole 1000G Phase 3 and UK10K) highlighted that 34% to 45% of the INGI variants are not represented (~12M with EUR, ~10M with 1000G and

~9M with UK10K respectively): 89% of those variants are private to each INGI cohort. Moreover 8% of the sites shared between two or all three INGI cohorts were not found either in the whole 1000G or in the EUR subpopulation from 1000G (which includes Italian samples from the Tuscany region - TSI), suggesting that they may be characteristic of the general Italian population. The majority of the private variants are within the range of the low and rare frequencies (MAF < 1%) (**Figure 1c)** while the proportion of low frequency and common variants are similar in the pool of shared sites **(figure supplement 1, table supplement 1**).

IGRP1.0: Reference panel and imputation

To increase the burden of good quality low frequency sites imputed in our isolated cohorts and possibly in the general Italian population, we generated a custom reference panel integrating our WGS data with already available resources from the 1000 Genomes project.

Variants with read depth (DP) lower than 5 and all singleton variants not overlapping between all INGI populations or the 1000 Genome project data were excluded. After filtering, 95.6%, 94.29% and 92.06% variants were retained for CAR, FVG and VBI, respectively (**table supplement 2**). Merging our data with the 1000G Phase 3 reference resulted in the addition of 6.9M Italian population specific variants, 7.8% of the merged INGI+1000G (IGRP1.0, from now on) panel (**table supplement 3**).

We tested our resource on the INGI populations and on an outbred Italian cohort of randomly selected samples. As shown in **Figure 2**, the panel including our data (red line) always outperforms the 1000G phase 3 reference panel for the INGI cohorts in terms of genotype concordance (r^2 - right y-axes), while there are not significant improvements for the outbred population (NW-ITA).

We compared our resource performances also in terms of the IMPUTE ‘info score’ metric. The proportion of well imputed sites (info score >= 0.4) in the IGRP1.0 reference panel was always higher compared to the 1000G phase 3 reference panel (red and blue bars respectively) with an increase from 20% to 36% of the rare sites (MAF<0.5%) with info score >=0.4 (**Figure 2, table supplement 4**). Imputation of an outbred Italian

population showed a similar outcome: the variants added by our resource spread evenly across the info score bins without jeopardizing the imputation results. In particular, for the lowest frequency bin we could impute 800.721 sites with IGRP1.0 versus 698.140 sites with 1000G phase 3 panel with info scores >=0.4 and a 13% increase of good imputation of the rare sites. We further validated our resource on three Croatian cohorts (VIS, KORCULA, SPLIT): the IGRP1.0 panel has higher proportion of well imputed sites with respect to other panels with a result similar to the outbred Italian population (**figure supplement 2 and table supplement 5**). A direct comparison with the recent HRC reference panel (McCarthy et al., 2016) was not performed since our populations (as well as the 1000G samples) are included in that reference. However, we checked the quality of sites belonging to the INGI cohorts that are excluded because of filtering from the HRC reference: among seven test cohorts, we identified 696.895 to 624.434 polymorphic sites with an average proportion of good quality sites (info score>=04) of 71% (63% - 81.5%). Focusing on rare variants for this subset, we can identify 256.222 to 326.076 polymorphic sites with a proportion of good quality sites between 15 and 63% **(table supplement 6)**.

IGRP1.0: GWAS studies

To assess the reliability of our new reference panel, we conducted a GWAS study with the newly imputed data on a series of red blood cells (RBC) traits (MCH, HGB, MCHC, RBC, HCT and MCV) for each INGI cohort followed by a meta-analysis. A total of 3292 individuals (age>=18 years) were included in the analysis. The characteristics of the samples are summarised in **table supplement 7**. Results from this analysis were compared with GWAS results for our cohorts with data imputed on the 1000G reference. Manhattan plots of all the meta-analysis results are given in **figure supplement 3**.

Lambda values of GWAS with 1000G showed no stratification (**figure supplement 4)**. Meta-analysis of GWAS with 1000G showed significant results (P<6.23E-9) only for MCH and MCV (**table supplement 8**). MCH analysis identified rs4820268 (P= 4.54E-10) in TMPRSS3, a gene already associated to MCH, MCV and MCHC (Ferreira et al., 2009; Kullo et al., 2010)**.** A locus on chromosome 11 at 3.8 Mb was identified for MCH and MCV both (rs117802349, MCH P= 2.67E-10, MCV P= 2.33E-11) and two loci on

chromosome 11 at 5.2 and 5.4 Mb were identified for MCV, near beta-globin cluster (rs113853911, p-value = 4.01E-09 and rs80297185, p-value = 1.84E-14 - D' = 1, r2=0.328). Other significant results were obtained for low frequency variants on chromosome 2 (MCH, P= 9.44E-10), 5 (MCH P= 1.98E-10 and MCV P= 1.16E-09), and 8 (MCV P=3.86E-10). Finally, a significant association was found for rs112483810 which lies 3Mb close to rs7844723, already found in association with HGB (Yang et al., 2007).

As shown in **figure supplement 5**, lambda of meta-analysis of GWAS with IGRP1.0 imputation were higher than lambda of 1000G imputation, due the high number of rare variants included in the new panel. However, the values ranged from 1.032 (MCHC) to 1.0505 (RBC) indicating adequate control of population stratification. The meta-analysis of results of IGRP1.0 imputed data showed several GWAS significant results, mainly in low frequency and rare variants (**table supplement 9**). The best hits for HGB, MCH, MCV and RBC were found in HBB cluster (chr11p15.4). The IGRP1.0 meta-analysis for HGB, MCH, MCV and RBC identified the pathogenic SNP rs11549407 located in HBB gene and responsible of beta-thalassemia (MCV P=1.86E-59, MCH P=4.88E-52, RBC P=8.30E-14, HGB P=3.67E-10) (Danjou et al., 2015). This SNP was also replicated with higher p-values for MCHC (P=0.0001) and HCT (P=5.38E-07). This locus was found only in CAR and VBI and this rare variant (CAR MAF= 0.48% and VBI MAF= 0.28%) was present neither in FVG nor in 1000G EUR. Furthermore, this variant is at very low frequency in Exac European (AF=0.07%) and has a r2 value of 0.328 (D'=1) with the rs113853911 variant identified in the previous analysis (replicated with IGRP1.0 only in HGB and HCT with p-values of 1.68E-4 and 2.49E-4 respectively).

IGRP1.0 meta-analysis confirmed TMPRSS6 gene for MCH and MCV already found in 1000G analysis (i.e. significant only in MCH).

Overall, IGRP1.0 imputation panel allowed us to replicate known loci and loci identified through the 1000G imputation, increasing also the number of significant variants, as shown in **Figure 3 a-b**.

Population structure

Using only European populations for PCA analysis, each INGI population separates from each other in the first four principal components (**Figure 4 a and Figure 4 b).** Regarding the FVG cohort, we can appreciate the separation of the six villages included in the isolate: Erto (ERT), Illegio (ILG), Resia (RSI), Sauris (SAU), San Martino del Carso (SMC) and Clauzetto (CLZ), underlining the evidence of population structure and of a high degree of isolation, as shown previously (Xue et al., 2017). Analyses and clustering using genomic pairwise Fst (**Figure 4 c**) highlight how INGI populations cluster with Europeans. However, the six villages from FVG show high levels of Fst in respect to other Italians. A closer look was taken with Treemix (Pickrell and Pritchard, 2012) analyses (**Figure 4 d**). Different lengths in the tree due to both inbreeding and genetic drift confirmed the peculiar structure of the FVG cohorts but, most importantly, it showed gene flow between North European population and North Eastern Italians. This adds more complexity to the Italian genomic pool.

Admixture (Alexander et al., 2009) analyses at different cluster solutions from K=2 to K=14 were also performed using worldwide reference populations from 1000G. The cluster solution with lowest cross validation error was for K=9 (**figure supplement 6**). VBI showed an admixture pattern similar to the one of Tuscany (TSI) from 1000G. The more isolated FVG populations showed their own ancestral component, that was however present at different fractions in all European and Italian populations suggesting a strong differential isolation of Italian subpopulations.

Finally, inbreeding coefficients and total homozygosity (due to ROH) showed high levels of variance among different Italian subpopulations as shown by the shape of the beanplots (**Figure 4 e-f)**. In particular, VBI shows the lowest mean coefficient (mean=0.008) CAR, CLZ and SMC had similar distributions (0.0149, 0.0134, 0.0151, respectively). Inbreeding was particularly high for ERT, SAU, ILG, and RSI (0.0191, 0.0325, 0.0304, 0.0311, respectively), the same pattern is followed by the total homozygosity due to ROH, which is quite different from the reference Italian population.

Natural Selection

We tested natural selection using the statistics iHS (Voight et al., 2006), we grouped the markers accordingly: markers with |iHS|>=2 in only one population and markers with

|iHS|>=2 in all Italian populations. We applied the following stringent criteria to select genes with signature of positive selection: at least 20 markers with |iHS|>=2.

A total of 37 other genes was found under putative selection in all Italian populations (**Supplementary table 10**). Interestingly, six genes (FHIT, CSMD1, CNTNAP2, MACROD2, RBFOX1 and PTPRD) were found under putative selection in all Italian populations but with different markers. Some of them had been previously associated with complex traits such as FHIT associated with BMI (Hoffmann et al., 2018), CSMD1 associated with 79 different phenotypes, including age of menarche (Perry et al., 2014), schizophrenia (Bergen et al., 2012) and educational attainment (Lee et al., 2018) (data from GWAS catalogue), CNTNAP2 associated with mathematical ability (Lee et al., 2018) and DNA methylation variation (Zhang et al., 2018), MACROD2 associated with several phenotypes, including educational attainment (Lee et al., 2018) and blood protein levels (Sun et al., 2018), RBFOX1 associated with eyes (Pickrell et al., 2016), other neurological traits and also educational attainment (Lee et al., 2018) and PTPRD associated with restless leg syndrome (Schormair et al., 2017) and blood pressure (Evangelou et al., 2018)**.**

As shown in **Figure 5,** the majority of genes with signatures of selection are not found in the TSI (the available Italian reference population from 1000G project). More in detail the fraction of private genes under selection ranges from 74% in VBI to 86% in RSI.

As a further example of the complex puzzle of signature of selection present in the Italian peninsula we selected the highest-ranking genes in terms of |iHS| and number of SNPs with |iHS|>=2 that are found only in one population. We then provided some examples reporting the genes with the average highest |iHS| and highest number of SNPs with |iHS|>=2 that are found only in one population.

Starting from the current Italian reference TSI, we found a strong signal for TYW1B, associated with triglycerides (Teslovich et al., 2010) and educational attainment. We found signature in CYP2C19 in CAR (associated with diastolic blood pressure (Liu et al., 2016), ABCG8 in VBI associated with lipid traits (Chasman et al., 2009), SLC25A12 in SMC (Educational attainment (Lee et al., 2018)), ERI3 in CLZ (associated with Educational attainment (Lee et al., 2018)), in ERT we found strong signatures for ANKRD30A, which was associated with paediatric autoimmune diseases,metabolite

levels and vestibular neuritis (Li et al., 2015), in SAU evidences were found for CLOCK gene associated with height (Wood et al., 2014), we found SSPN in ILG (associated with atrial fibrillation,(Nielsen et al., 2018)) and finally PBRM1 in RSI which was linked to blood protein levels (Sun et al., 2018),schizophrenia,general cognitive ability (Davies et al., 2018, p. 4). These are few examples of the different genomic patterns that can be found in the various subpopulations of the Italian peninsula.

Deleterious variants enrichment

The profound differentiation in all our Italian samples, and subsequent different level of isolation and selection signature led to the question whether there was any difference in deleterious or neutral variant distribution among different populations compared to the Italian reference population.

To answer this question, we applied the DVxy statistic (Xue et al., 2017) for DV variants (Drifted Variants respect to a reference) between 1-2 allele count (AC) and 3-5 AC in each population using as actual Italian reference (TSI). Variants were grouped according to CADD score. In our analysis we discovered a significant relative enrichment in deleterious variants with CADD>20 in the more isolated/inbred populations compared to the TSI (DVxy-sd>1), this is true when we are considering populations such as ILG, RSI, SAU and also SMC, whereas no differences were found when considering neutral or low deleterious variants (CADD 0-5, DVxy+/-sd=1) (**Figure 6**).

This level of enrichment could be explained with lower effectiveness of purifying selection due to isolation and small effective population size (Xue et al., 2017), however the interesting point lies somewhere else: these Italian populations show enrichment for high deleterious variants (CADD>20) at low frequency (3-5 AC).

In order to show the complexity of the Italian catalogue of deleterious variants, we estimated the ratio of DV variants (3-5 AC and CADD>20) that are different between pairs of sub-populations with DV variants shared between pairs (**Figure 7**), with the lowest value being 12 for the pair CLZ/VBI and the highest 31 for RSI/SAU. All values are highly positive indicating that the majority of DV variants are private of each subpopulations.

Human Knockout

Homozygote loss of Function (LoF) variants represent a category of deleterious variants and examples of human KO (HKO). Considering that the highest enrichment was found for variants with CADD score>20, we used this value to select LoF variation.

In our total cohort, 509 LoF presenting with a CADD score >20 were found at homozygous state in at least one individual per population (**table supplement 11**).Gene ontology analysis revealed an excess of transmembrane signalling receptor genes including olfactory receptors, as already described (MacArthur et al., 2012).

In order to have an high reliable dataset, we used a stringent filtering criteria analysing only variants that affected all transcripts (considered as TOTAL LoF in opposition to PARTIAL LoF), resulting in 205 variants affecting 195 different genes (**table supplement 12**)**.** Among these 205 variants, the majority (150, ~73%) was shared among all 3 populations and more than a half (~60%) had frequency >=0.05**.** A large number of HKOs was located in genes involved in hair/skin/epithelium or eye phenotypes, and many were members of gene families. As a matter of facts, 5 HKO were found in keratin genes: (KRT37, KRT24, KRT31 and KRT83) and 5 in keratin associated protein (KRTAP1-5, KRTAP1-1, KRTAP19-6 and KRTAP13-2, KRTAP29-1). Two different rare stop gain mutations were found in KRT83 (AF in Europeans rs146753414=0.0265, rs2857667 =0.0063). Two missense mutations in this gene were associated to a mild form of monilethrix (MNLIX; OMIM #158000), a rare autosomal dominant hair disease that results in fragile, brittle hair that tends to fracture and produce some degree of alopecia (Steensel et al., 2005). The hair of three carriers of the KO of KRT83 in the VB population and of nine heterozygotes in three families was investigated and resulted normal. Lack of KRT83 does not seem to affect hair structure as much as substitutions of amino acids that are highly conserved and affect the helix termination motifs, known hotspots for monilethrix mutations. Finally, we found a very rare stop gain (rs11355796) in COL6A5 gene (collagen type VI, alpha 5) identified in homozygous state in one individual from VBI. The variant is enriched in all three Italian populations (AF~0.015 compared with reference Europeans of 0.0013. Mutations in the COL6A5 gene were shown to cause familial neuropathic chronic itch (Martinelli-

Boneschi et al., n.d., p. 5). Unfortunately, we do not have so far clinical data on our homozygous HKO, but for sure a related heterozygous carrier reported to complain of itching all his life.

We then analysed only variants reported in gnomAD, resulting in 133 different genes which are distributed among the populations as shown in **Figure 8**: we found that the majority of genes are private of FVG, VBI and CAR (61, 36 and 10 respectively) whereas only 13 genes are shared among all populations. Among these HKO genes only few of them show evidence of selection (11 genes out of 133) (see t**able supplement 13**), in particular signatures of selection are found in populations where the LOF variants are not present with the exception of CDH23 (associated with LDL cholesterol) and KLHL23 (associated with obesity-related traits); however, these HKO are considered PARTIAL.

The novel aspect that we report is that we add new information into the variability and genomic signatures in HKO genes present in the Italian genomic pool.

Discussion

The ability to interrogate all kind of genetic variations is critical for the classification of genetic determinants of complex and monogenic disorders: the whole genome sequencing of peculiar populations such as isolates has given a significant contribution, providing denser data and allowing a better mapping of the genomic features under study (Hatzikotoulas et al., 2014).

Here, we report the results of a series of analyses obtained through the investigation of WGS from 947 subjects coming from different Italian geographic areas (i.e. South, North West and North East) and their contribution to the identification and description of a relevant proportion of the Italian population pool of genetic variation.

The number of new variants described and discovered, especially in comparison with 1000G data, which include the Italian reference TSI (~1.86 M variants shared between all INGI cohorts but not TSI) confirms that these genomes are able to increment the catalogue of Italian genotypic variation, in particular in the low frequency spectrum.

This leads us to the inevitable next step: the creation of a reference panel for imputation using the Italian whole genome data. The “Italian core” of the reference panel was

assembled with INGI data only and it proved to be an extremely useful resource when merged with other larger reference panels: the IGRP1.0 outperformed the 1000G Phase 3 reference panel for imputation of inbred and outbred Italian and other European populations such as the Croatians cohorts.

More in detail, our new reference panel could facilitate the imputation of rare variants for GWAS studies and help the identification of population specific variants of different Italian and possibly Southern European populations: a notable point is that we are incrementing the total number of variants that are valuable for GWAS studies without adding “noise” neither INGI populations, as expected, nor in other outbred populations in terms of imputation quality.

With this resource at our disposal, another question arises: will we be able to increment the power to detect genome wide significant loci/variants using this new reference panel for imputation?

In this case, the reliability of IGRP1.0 panel was proven running a series of GWAS tests on some selected RBC traits, demonstrating that it performs better than the 1000G panel alone. As a matter of facts, GWAS studies carried out with IGRP1.0 panel imputed data, not only replicated previous findings with higher statistical significance, but also demonstrated that several previously found suggestive signals ($p<1E-5$) became genome wide significant ($p<1E-8$).

One interesting example is the RBFOX1 gene: we showed that it harbours signals of selection in all Italian populations and, moreover, it carries two variants significantly associated to MCV and MCH traits, that we were able to pinpoint only through our custom reference panel (**table supplement 11**). These results need further dissection, but are, again, a clue of the fact that the genetic features of our cohorts represent an important resource in the understanding of gene function and association to different traits.

On the other hand, one of the major point of our work consisted in addressing the issue of the underrepresentation of South European population in whole genome databases. Recent works based on array data pinpointed the genetic diversity in the Italian peninsula (Sazzini et al., 2016) along with the presence of isolates (Esko et al., 2013). This foregoing information prompted the inquiry about the homogeneity of genomes

coming from different regions of Italy in terms of diverse genomic aspects (population structure, natural selection signatures, deleterious variants distribution and HKO) and, as a consequence, how reliable is to use only one reference population for Italians such as the Tuscan (TSI).

Population structure analysis revealed that our populations fall in the European pole of variation, but their separation from the North European cohorts is clear.

Principal component analysis, tree graph analyses, ancestry coefficient distribution confirm the non-homogeneous genetic background of the Italian populations from North to South and highlight the fact that the use of the only TSI as genome reference leads to an underestimation of the Italian genomic variability, and if that was not enough, ROH pattern and inbreeding coefficient showed a wide array of values not comparable using the only South European reference of 1000 Genomes.

Discussing Natural Selection, it was demonstrated that environmental differences along the peninsula might have shaped the genome through mechanisms such as evolution and selective pressure (Sazzini et al., 2016). Our analyses pinpointed the presence of shared selective pressure on specific genes in all Italian populations such as HIT, CSMD1, CNTNAP2, MACROD2, RBFOX1 and PTPRD, however the striking point was the level of selection signatures that are private of single populations (when substructure is taken into account) ranging up to 86% of the total genes found for RSI. In addition, considering the relationship of some populations (RSI, SAU, SMC) with North European populations (as shown in Treemix analyses), we can suppose that a number of haplotypes passed in some North East Italian populations but not others: this peculiar gene flow could be responsible for some unique signals of selection.

For what concerns the distribution of deleterious variants it was already demonstrated how the relative relaxation of purifying selection in presence of isolation (Chheda et al., 2017; Xue et al., 2017) leads to an increased frequency of specific deleterious variants. This aspect reinforces our thesis about the need of a more broadened reference for the Italian genomic variation, as we demonstrated that not only have we an enrichment of low frequency deleterious variants (CADD>=20) in our genomes, but also most of this enrichment is population-specific.

In our analyses of human knockout we showed that the majority of the loss of function

(>70%) was shared among all the three populations and, as expected, they belong to the category of transmembrane signalling receptor genes, including olfactory receptors. Nevertheless, while analysing only the genes with at least 1 homozygous individual in each cohort, we discovered an inverse pattern: the majority of genes harbouring HKO are private of each cohort. In addition the majority of them (91%) was not found in any selection scan, suggesting the lack of evolutionary constraints for these genes, even though more accurate analyses are needed to confirm this consideration. Still, this gives us another hint of the necessity of multiple genomes to describe the catalogue of HKO present in Italy: with our data we are starting to scratch the surface, providing some of them.

Furthermore, HKO and pattern of deleterious variants are useful examples to show how clinical-relevant polymorphisms could be found enriched in frequency in specific populations within the same country and provide extremely relevant information that could be used for developing personalized medicine strategies: another great added value of our cohorts is that a large series of instrumental and clinical phenotypes is already available. Thus, future efforts should be pointed towards the functional characterization of putatively enriched variants and deep phenotyping of carriers of such polymorphisms as well as deep phenotyping of HKO.

In conclusion, we showed how our unique dataset of populations and WGS data fill a gap in publicly available human genome sequence data sets (i.e. 1000G, gnomAD databases, etc.), in which Southern European populations - a significant proportion of the overall European populations - are highly underrepresented and it will able to produce regionally appropriate reference panels.

Furthermore,considering the fact that in Italy, a National Genomic BioBank is not yet existing the availability of a catalogue of rare and low frequency variants for Italians populations will facilitate the understanding of these genetic loci improving the accuracy and efficacy of a series of genetics/genomics studies, and subsequently opening new perspectives for precise medicine and drug targets identification.

Materials and Methods

WGS data generation: variant calling and quality control

All samples selected for sequencing already had genotype data from other platforms (SNP array and Exome chip): this data allowed us to assess genotype concordance against a “trusted” set of variants. For all cohorts, samples were selected randomly. The sequencing was carried out at different sequencing centres: the Wellcome Trust Sanger Institute in Hinxton (UK), the BGI, Shenzhen (PRC) and the HSR in Milan. **Table supplement 14** summarises the total number of samples sequenced. All the data were post processed at the Sanger Institute. A written informed consent for participation was obtained from all subjects. Regarding the FVG cohort the project was approved by the Ethical committee of the IRCCS Burlo-Garofolo. Regarding the CAR cohort the project was approved by the local administration of Carlantino, the Health Service of Foggia Province, Italy, and ethical committee of the IRCCS Burlo-Garofolo of Trieste. For VBI cohort, data collection and genotyping were approved by the institutional ethical committee of the San Raffaele Hospital in Milan and by the Regione Piemonte.

The raw data were checked first at lane level to remove any sample with bad quality data. 54 samples were realigned to the hs37d5 reference sequence because they were aligned to a previous version of the GRCh37 build: this process has been carried out using the ‘Bridgebuilder system’ developed by the Human Genetics Informatic group at the Wellcome Trust Sanger Institute (*BridgeBuilder*, 2015) .

After the alignment, performed with bwa software (Li and Durbin, 2010), each bam file was improved through the implementation of the following steps: 1) Realignment around known and discovered INDELs using GATK (McKenna et al., 2010) RealignerTargetCreator and IndelRealigner; 2) Base Quality Recalibration by GATK BQSR using the BaseRecalibrator and PrintReads tools; 3) Recalculation of the MD tag by samtools (Li, 2011) calmd; 4) Bam indexing.

For the Carlantino cohort, sequencing was carried out using Illumina technology (Genome Analyzer and HiSeq 2000) at the Wellcome Trust Sanger Institute for 115 samples with an average coverage of 4x, an additional batch of 40 samples was sequenced at Beijing Genomics Institute (BGI) with an average coverage of 10x. Among the 115 samples sequenced at the Sanger Institute 27 failed the quality check at the

lane level: 5 were re-processed while 22 were excluded from further analyses.
The most common cause of failure was the high percentage of adapter contamination and a bimodal insert size distribution.
For the Friuli Venezia Giulia cohort, 200 samples were sequenced at the Wellcome Trust Sanger Institute with a mean coverage of 4x and 192 samples at BGI with a mean coverage of 10x. Among the 200 Sanger samples, only 4 failed the quality check at the lane level, thus were excluded from further analyses. Among the BGI set 6 samples were duplicated from the Sanger pool: we merged the two sets of data to increase the coverage of each sample. We removed 1 additional sample from this set because of data corruption.
The data for the Val Borbera Cohort were generated at a mean coverage of 6x for all selected samples: 210 were sequenced at the Wellcome Trust Sanger Institute, 209 were sequenced at BGI and a small batch of 29 was processed at the San Raffaele Hospital. After the first step of quality check we removed 2 samples from the Sanger Institute set for contamination and bad quality DNA respectively, 12 samples from the OSR dataset for bad quality and 1 sample from the BGI set for data corruption.
Finally a total set of 947 samples was sent forward for the Variant Calling step.
We produced genotype calls for autosomal chromosomes separately for each population using the following pipeline.
Samtools mpileup (v.1.2) (Li, 2011) was used for multisample genotype calling (parameter set: -E -t DP,DV,SP -C50 -pm3 -F0.2 -d 10000). The generated BCF files were converted to VCF format with bcftools call (v.1.2) (parameter set: -Nvm) and filtered with bcftools filter (v.1.2) (parameter set: -m+ -sLowQual -e"%QUAL<=0"-g3 -G10 -Ov -). Variant Quality Score Recalibrator (VQSR) filtering was applied to the raw call data with GATK v.3.3 (DePristo et al., 2011). Raw calls from samtools were used with the UnifiedGenotyper module in "Given allele mode" to generate all the annotation needed to calculate the VQSLOD scores through the VariantRecalibrator module, separately for SNVs and INDELs. For SNVs we selected the following parameters: i) Annotations: QD, DP, FS, HaplotypeScore, MQRankSum, ReadPosRankSum, InbreedingCoeff; ii) Training set: HapMap 3.3, Omni 2.5M chip, 1000 Genomes Phase I; iii) Truth set: HapMap 3.3, Omni 2.5M chip; iv) Known set: dbSNP build 138. For

INDELs we selected: i) Annotations: DP, FS, ReadPosRankSum, MQRankSum; ii) Training set: Mills-Devine, 1000 Genomes Phase I, dbSNP v138; iii) Truth set: Mills-Devine; iv) Known set: Mills-Devine, dbSNP build 138. For each population the lowest VQSLOD threshold was chosen according to the output produced by VariantRecalibrator to select the best cut-off in terms of specificity and sensitivity of the trained model. The Transition/Transversion (Ti/Tv) ratio was used as a parameter to select the best threshold, taking as a reference the empirical value of ~ 2 calculated by (1000 Genomes Project Consortium et al., 2012) . For SNPs the minimum VQSLOD values selected were -59.1994 (99.94% truth sensitivity threshold), -15.0283 (99.80% truth sensitivity threshold), -22.6034 (99.9% truth sensitivity threshold) for VBI, FVG and CAR cohort respectively. For INDELs we used a more conservative approach, selecting a sensitivity threshold of 95% for each population. The filter was applied to each call set with GATK ApplyRecalibration module.

We performed several genotype refinement steps on the filtered data: 1. BEAGLEv4.r1230 (Browning and Browning, 2007) was used to assign posterior probabilities to all remaining genotypes. 2. SHAPEITv2 (Delaneau et al., 2013) to phase all genotypes calls and 3. IMPUTEv2 (Howie et al., 2009) to perform internal imputation in order to correct genotyping errors.

Finally, bcftools annotate (v.1.2) was used to add information about Ancestral Allele and allele frequencies from 1000G phase 3 (Sudmant et al., 2015) populations and rsIDs from dbSNP v.141 (Sherry et al., 2001). The Variant Effect Predictor v.90 (McLaren et al., 2010) provided all consequence annotation as well as Polyphen and Sift information. CADD score (Kircher et al., 2014) information was also added.

Samples and sites were again investigated for outliers or artefacts after the variant calling.

First, we looked for batch effects due to the different sequencing centres: we conducted an MDS analysis on each cohort testing the first PCA component for correlation with the sequencing centre variable with a Pearson’s correlation test and obtaining a significant outcome only for the FVG cohort ($p=0.001728$). We compared the analysis for the FVG cohort with data available from a previous work (Esko et al., 2013) showing that the pattern is consistent with the underlying population structure. We then generated a sites

exclusion list, focusing on: a) Hardy-Weinberg equilibrium (sites removed if exact test p-value was below the threshold of 1e−8); b) Heterozygosity rate distribution (removed sites with values greater than 3 standard deviations of the mean); c) MAF mismatch when compared with SNP array data; d) Non Reference Discordance rate (NRDR), defined as the ratio between the sum of concordant calls of the alternative allele in WGS and array data and the sum of all discordant calls of the alternative allele in WGS and array data (cut off value for removal of 3 standard deviations of the mean).

We removed 5,552 , 2,577 and 2,502 sites from CAR, FVG and VBI respectively.

We excluded samples using the following parameters: a) Singleton number, b) Heterozygosity rate and c) Non Reference Discordance rate. We removed one sample from the FVG cohort for an excess of singletons (~100000 singletons counted). We calculated also the heterozygosity rate for each sample and removed all samples with values exceeding a threshold of 3 SD from the average value for each population: one sample was removed from the CAR cohort, one sample from the FVG cohort and 4 samples from the VBI cohort. Finally, we calculate the samples' non-reference discordance rate and removed all individuals with an NRDR greater than 5%: 8 samples from the CAR cohort, 1 sample from FVG cohort and 5 samples from the VBI cohort.

Reference imputation panel

We selected a 'highly reliable' subset of variants to include in our reference panel.

In order to avoid mismatches between the INGI datasets, we split all multi-allelic variant sites in different vcf records and performed INDELs normalization with bcftools norm to prepare the data. Data from 1000G Project phase 3 and UK10K project (The UK10K Consortium, 2015) were processed in the same way.

SNPs and INDELs from the INGI WGS data to be included in the reference panel were selected with the following criteria: a) all sites with Alternative Allele count (AC) >= 2 and Read depth (DP) >= 5; b) all the singleton sites (AC = 1) either shared at least between two INGI cohorts or which were known sites or present at least in one of the external resources selected (UK10K and 1000G Project Phase 3).

To build the Italian reference dataset, a 'core' INGI panel was created merging data from the different INGI cohorts, using the method implemented by the IMPUTE2

559 software (Howie et al., 2011). The data were then added to the 1000G phase 3
560 reference panel to obtain a final reference (INGI+1000G also called IGRP1.0).
561 The imputation test was performed on chromosome 2 genotypes in different cohorts: a)
562 INGI cohorts; b) a cohort of 567 unselected outbred samples from North Western Italy
563 (NW-ITALY); c) three cohorts from Croatia (VIS - 960 samples, KORCULA - 1812
564 samples and SPLIT - 466 samples)
565 Imputation metrics across the different panels were compared for each population. We
566 assessed the $\mathbf{r^2}$ metric, which estimates the correlation between the true genotype and
567 the imputed genotype and the IMPUTE **info score** parameter, which provides a
568 measure of the observed statistical information associated with the allele frequency
569 estimate for each variant (Marchini and Howie, 2010). We removed from each INGI
570 cohort all the samples represented in the reference panel.
571
572 **Genome Wide Association Studies (GWAS)**
573 GWA studies on Red Blood Cells indexes (MCH, HGB, MCHC, RBC - normalized with
574 natural logarithm , HCT and MCV) were performed in each population separately, using
575 age and gender as covariate in an additive model, once using 1000G imputation and
576 once IGRP1.0. The analyses were carried out using the mixed linear models as
577 implemented in R ABEL packages (Aulchenko et al., 2007). Genomic kinship was used
578 to take into account the relatedness. Variants with info score<=0.4 were excluded if the
579 MAF was>=1%. For rare variants (MAF 0.1%-1%), a more stringent Info Score cut-off
580 (>=0.8) was used (Pistis et al., 2015). Meta-analysis was performed using the software
581 METAL (Willer et al., 2010) and heterogeneity Cochran Q test was performed. After
582 meta-analysis, the variants that were not present with the same direction in at least two
583 of the three cohorts were excluded. Variants with significant p-value (<0.05) for
584 heterogeneity test were also excluded. Bonferroni correction was applied: the thresholds
585 were P=6.23e-9 for 1000G and 4.69e-9 for IGRP1.0. The positions are referred to the
586 build 37. Manhattan plots were generated with the R library qqman (Turner, 2017) and
587 hudson package (Lucas, 2018) .
588
589 **Population Structure**

Principal component analysis (PCA) was carried out to define the genetic structure of our population using PLINK (Daly et al., 2007). PCA was carried out after removing markers in high LD (r2>0.4), using the function --indep-pairwise 200 50 0.4 and with MAF <0.02. Runs of homozygosity (ROH) and inbreeding coefficient we estimated as well using PLINK using the command --homozyg and --het . Pairwise Fst between worldwide populations was calculated using the software 4p (Benazzo et al., 2015) .
The same dataset was used for tree graph analyses implemented in Treemix (Pickrell and Pritchard, 2012). The analysis of ancestral component was performed using ADMIXTURE v 1.2 (Alexander et al., 2009) using the European population plus one African reference (YRI) one East Asian (CHB) and one South Asian (GIH). Cross validation error procedure was implemented to select the best cluster solution.

Natural Selection

Evidence of positive selection was estimated for each population using iHS statistic (Voight et al., 2006) implemented in selscan program (Szpiech and Hernandez, 2014), we used only markers with MAF>0.05, furthemore we adopted a conservative approach for genes under putative positive selection: we selected only genes with at least 20 markers with standardized |iHS|>=2.

Deleterious Variants

After the exclusion of multiallelic variants, we subdivided all variant in bins according to their CADD score and frequency. The following minor allele frequency classes were created: between 1-2 allele count, 3-5-allele count, 5-10AC and more 10 AC, thus the variants were binned in the following CADD categories 0-5, 5-15,15-20 >20. We then applied the DVxy statistic as described in Xue et al., using as reference the TSI population from 1000 Genomes. In addition, we estimated the ratio of of private and shared DV variants (variants enriched).

Human Knockouts

To identify HKO, we considered only deleterious variants in protein coding genes: we first selected variants with high impact as defined by VEP (i.e. frameshift, splice

acceptor variant, splice donor variant, stop gained, stop lost, start lost, transcript ablation, transcript amplification) and among those we further selected for CADD score >=20. A total of 12,231 variants (8,832 SNV and 3,399 indels) were selected and 5,916 had a CADD score >=20. Among this subset of variants, those presenting at least one homozygous individual in one population were defined putative HKO. After filtration for total KO, the average number of HKO per individual was 20 (12-31), in agreement with previous determinations (Narasimhan et al., 2016). HKO's were classified as TOTAL when the variant was predicted as LOF in all Ensembl database transcript, otherwise they were classified as PARTIAL, even though this approach is highly conservative, as some PARTIAL loci could still affect the functional transcripts. Overlaps of HKOs between populations were analysed using the R package "VennDiagram" (Chen, 2018) (https://cran.r-project.org/package=VennDiagram).

Acknowledgements

We would like to thank the people of the Friuli Venezia Giulia Region and of Carlantino for the everlasting support. We thank the inhabitants of the Val Borbera that made this study possible, the local administrations, the Tortona and Genova archdiocese and the ASL-22, Novi Ligure (Al) for support. We also thank Clara Camaschella for data collection supervision and organization of the clinical data collection, Fiammetta Viganò for technical help, Corrado Masciullo for building the analysis platform.

Funding

Fort FVG and CAR cohorts: Project co-financed by the European Regional Development Fund under the Regional Operational Programme of Friuli Venezia Giulia - Objective "Regional Competitiveness and Employment" 2007/2013, Telethon Foundation (GGP09037), Fondo Trieste (2008), Regione FVG (L.26.2008), and Italian Ministry of Health (RC16/06, ART. 13 D.LGS 297/99) (to PG). For VBI cohort:The research was supported by funds from Compagnia di San Paolo, Torino, Italy; Fondazione Cariplo, Italy and Ministry of Health, Ricerca Finalizzata 2008 and CCM

2010, and Telethon, Italy to DT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

No competing interests declared

References

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

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**:1655–1664. doi:10.1101/gr.094052.109

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinforma Oxf Engl* **23**:1294–1296. doi:10.1093/bioinformatics/btm108

Benazzo A, Panziera A, Bertorelle G. 2015. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol Evol* **5**:172–175. doi:10.1002/ece3.1261

Bergen SE, O'Dushlaine CT, Ripke S, Lee PH, Ruderfer DM, Akterin S, Moran JL, Chambert KD, Handsaker RE, Backlund L, Ösby U, McCarroll S, Landen M, Scolnick EM, Magnusson PKE, Lichtenstein P, Hultman CM, Purcell SM, Sklar P, Sullivan PF. 2012. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry* **17**:880–886. doi:10.1038/mp.2012.73

BridgeBuilder efficiently remaps BAM/SAM reads to a new reference by first building a "bridge" reference, first mapping to that bridge, and then remapping only a subset of reads to the fu.. 2015. . Wellcome Trust Sanger Institute - Human Genetics Informatics.

683 Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing-
684 Data Inference for Whole-Genome Association Studies By Use of Localized
685 Haplotype Clustering. *Am J Hum Genet* **81**:1084–1097. doi:10.1086/521987
686 Chasman DI, Paré G, Mora S, Hopewell JC, Peloso G, Clarke R, Cupples LA, Hamsten
687 A, Kathiresan S, Mälarstig A, Ordovas JM, Ripatti S, Parker AN, Miletich JP,
688 Ridker PM. 2009. Forty-Three Loci Associated with Plasma Lipoprotein Size,
689 Concentration, and Cholesterol Content in Genome-Wide Analysis. *PLOS Genet*
690 **5**:e1000730. doi:10.1371/journal.pgen.1000730
691 Chen H. 2018. VennDiagram: Generate High-Resolution Venn and Euler Plots.
692 Chheda H, Palta P, Pirinen M, McCarthy S, Walter K, Koskinen S, Salomaa V, Daly M,
693 Durbin R, Palotie A, Aittokallio T, Ripatti S. 2017. Whole-genome view of the
694 consequences of a population bottleneck using 2926 genome sequences from
695 Finland and United Kingdom. *Eur J Hum Genet*. doi:10.1038/ejhg.2016.205
696 Daly M, Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira M, Bender D, Maller J,
697 Sklar P, Debakker P. 2007. PLINK: A Tool Set for Whole-Genome Association
698 and Population-Based Linkage Analyses. *Am J Hum Genet* **81**:559–575.
699 doi:10.1086/519795
700 Danjou F, Zoledziewska M, Sidore C, Steri M, Busonero F, Maschio A, Mulas A, Perseu
701 L, Barella S, Porcu E, Pistis G, Pitzalis M, Pala M, Menzel S, Metrustry S,
702 Spector TD, Leoni L, Angius A, Uda M, Moi P, Thein SL, Galanello R, Abecasis
703 GR, Schlessinger D, Sanna S, Cucca F. 2015. Genome-wide association
704 analyses based on whole-genome sequencing in Sardinia provide insights into
705 regulation of hemoglobin levels. *Nat Genet* **advance online publication**.
706 doi:10.1038/ng.3307
707 Davies G, Lam M, Harris SE, Trampush JW, Luciano M, Hill WD, Hagenaars SP, Ritchie
708 SJ, Marioni RE, Fawns-Ritchie C, Liewald DCM, Okely JA, Ahola-Olli AV, Barnes
709 CLK, Bertram L, Bis JC, Burdick KE, Christoforou A, DeRosse P, Djurovic S,
710 Espeseth T, Giakoumaki S, Giddaluru S, Gustavson DE, Hayward C, Hofer E,
711 Ikram MA, Karlsson R, Knowles E, Lahti J, Leber M, Li S, Mather KA, Melle I,
712 Morris D, Oldmeadow C, Palviainen T, Payton A, Pazoki R, Petrovic K, Reynolds
713 CA, Sargurupremraj M, Scholz M, Smith JA, Smith AV, Terzikhan N, Thalamuthu

714 A, Trompet S, Lee SJ van der, Ware EB, Windham BG, Wright MJ, Yang J, Yu J,
715 Ames D, Amin N, Amouyel P, Andreassen OA, Armstrong NJ, Assareh AA, Attia
716 JR, Attix D, Avramopoulos D, Bennett DA, Böhmer AC, Boyle PA, Brodaty H,
717 Campbell H, Cannon TD, Cirulli ET, Congdon E, Conley ED, Corley J, Cox SR,
718 Dale AM, Dehghan A, Dick D, Dickinson D, Eriksson JG, Evangelou E, Faul JD,
719 Ford I, Freimer NA, Gao H, Giegling I, Gillespie NA, Gordon SD, Gottesman RF,
720 Griswold ME, Gudnason V, Harris TB, Hartmann AM, Hatzimanolis A, Heiss G,
721 Holliday EG, Joshi PK, Kähönen M, Kardia SLR, Karlsson I, Kleineidam L,
722 Knopman DS, Kochan NA, Konte B, Kwok JB, Hellard SL, Lee T, Lehtimäki T, Li
723 S-C, Liu T, Koini M, London E, Longstreth WT, Lopez OL, Loukola A, Luck T,
724 Lundervold AJ, Lundquist A, Lyytikäinen L-P, Martin NG, Montgomery GW,
725 Murray AD, Need AC, Noordam R, Nyberg L, Ollier W, Papenberg G, Pattie A,
726 Polasek O, Poldrack RA, Psaty BM, Reppermund S, Riedel-Heller SG, Rose RJ,
727 Rotter JI, Roussos P, Rovio SP, Saba Y, Sabb FW, Sachdev PS, Satizabal CL,
728 Schmid M, Scott RJ, Scult MA, Simino J, Slagboom PE, Smyrnis N, Soumaré A,
729 Stefanis NC, Stott DJ, Straub RE, Sundet K, Taylor AM, Taylor KD, Tzoulaki I,
730 Tzourio C, Uitterlinden A, Vitart V, Voineskos AN, Kaprio J, Wagner M, Wagner
731 H, Weinhold L, Wen KH, Widen E, Yang Q, Zhao W, Adams HHH, Arking DE,
732 Bilder RM, Bitsios P, Boerwinkle E, Chiba-Falek O, Corvin A, Jager PLD, Debette
733 S, Donohoe G, Elliott P, Fitzpatrick AL, Gill M, Glahn DC, Hägg S, Hansell NK,
734 Hariri AR, Ikram MK, Jukema JW, Vuoksimaa E, Keller MC, Kremen WS, Launer
735 L, Lindenberger U, Palotie A, Pedersen NL, Pendleton N, Porteous DJ,
736 Räikkönen K, Raitakari OT, Ramirez A, Reinvang I, Rudan I, Rujescu D, Schmidt
737 R, Schmidt H, Schofield PW, Schofield PR, Starr JM, Steen VM, Trollor JN,
738 Turner ST, Duijn CMV, Villringer A, Weinberger DR, Weir DR, Wilson JF,
739 Malhotra A, McIntosh AM, Gale CR, Seshadri S, Mosley TH, Bressler J, Lencz T,
740 Deary IJ. 2018. Study of 300,486 individuals identifies 148 independent genetic
741 loci influencing general cognitive function. *Nat Commun* **9**:2098.
742 doi:10.1038/s41467-018-04362-x
743 Delaneau O, Howie B, Cox AJ, Zagury J-F, Marchini J. 2013. Haplotype Estimation
744 Using Sequencing Reads. *Am J Hum Genet* **93**:687–696.

745 doi:10.1016/j.ajhg.2013.09.002
746 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
747 Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM,
748 Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A
749 framework for variation discovery and genotyping using next-generation DNA
750 sequencing data. *Nat Genet* **43**:491–498. doi:10.1038/ng.806
751 Esko T, Mezzavilla M, Nelis M, Borel C, Debniak T, Jakkula E, Julia A, Karachanak S,
752 Khrunin A, Kisfali P, Krulisova V, Aušrelé Kučinskiené Z, Rehnström K, Traglia M,
753 Nikitina-Zake L, Zimprich F, Antonarakis SE, Estivill X, Glavač D, Gut I, Klovins J,
754 Krawczak M, Kučinskas V, Lathrop M, Macek M, Marsal S, Meitinger T, Melegh
755 B, Limborska S, Lubinski J, Paolotie A, Schreiber S, Toncheva D, Toniolo D,
756 Wichmann H-E, Zimprich A, Metspalu M, Gasparini P, Metspalu A, D'Adamo P.
757 2013. Genetic characterization of northeastern Italian population isolates in the
758 context of broader European genetic diversity. *Eur J Hum Genet* **21**:659–665.
759 doi:10.1038/ejhg.2012.229
760 Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, Ntritsos G,
761 Dimou N, Cabrera CP, Karaman I, Ng FL, Evangelou M, Witkowska K, Tzanis E,
762 Hellwege JN, Giri A, Edwards DRV, Sun YV, Cho K, Gaziano JM, Wilson PWF,
763 Tsao PS, Kovesdy CP, Esko T, Mägi R, Milani L, Almgren P, Boutin T, Debette S,
764 Ding J, Giulianini F, Holliday EG, Jackson AU, Li-Gao R, Lin W-Y, Luan J,
765 Mangino M, Oldmeadow C, Prins BP, Qian Y, Sargurupremraj M, Shah N,
766 Surendran P, Thériault S, Verweij N, Willems SM, Zhao J-H, Amouyel P, Connell
767 J, Mutsert R de, Doney ASF, Farrall M, Menni C, Morris AD, Noordam R, Paré G,
768 Poulter NR, Shields DC, Stanton A, Thom S, Abecasis G, Amin N, Arking DE,
769 Ayers KL, Barbieri CM, Batini C, Bis JC, Blake T, Bochud M, Boehnke M,
770 Boerwinkle E, Boomsma DI, Bottinger EP, Braund PS, Brumat M, Campbell A,
771 Campbell H, Chakravarti A, Chambers JC, Chauhan G, Ciullo M, Cocca M,
772 Collins F, Cordell HJ, Davies G, Borst MH de, Geus EJ de, Deary IJ, Deelen J, M
773 FDG, Demirkale CY, Dörr M, Ehret GB, Elosua R, Enroth S, Erzurumluoglu AM,
774 Ferreira T, Frånberg M, Franco OH, Gandin I, Gasparini P, Giedraitis V, Gieger
775 C, Girotto G, Goel A, Gow AJ, Gudnason V, Guo X, Gyllensten U, Hamsten A,

776 Harris TB, Harris SE, Hartman CA, Havulinna AS, Hicks AA, Hofer E, Hofman A,
777 Hottenga J-J, Huffman JE, Hwang S-J, Ingelsson E, James A, Jansen R, Jarvelin
778 M-R, Joehanes R, Johansson Å, Johnson AD, Joshi PK, Jousilahti P, Jukema
779 JW, Jula A, Kähönen M, Kathiresan S, Keavney BD, Khaw K-T, Knekt P, Knight
780 J, Kolcic I, Kooner JS, Koskinen S, Kristiansson K, Kutalik Z, Laan M, Larson M,
781 Launer LJ, Lehne B, Lehtimäki T, Liewald DCM, Lin L, Lind L, Lindgren CM, Liu
782 Y, Loos RJF, Lopez LM, Lu Y, Lyytikäinen L-P, Mahajan A, Mamasoula C,
783 Marrugat J, Marten J, Milaneschi Y, Morgan A, Morris AP, Morrison AC, Munson
784 PJ, Nalls MA, Nandakumar P, Nelson CP, Niiranen T, Nolte IM, Nutile T,
785 Oldehinkel AJ, Oostra BA, O'Reilly PF, Org E, Padmanabhan S, Palmas W,
786 Palotie A, Pattie A, Penninx BWJH, Perola M, Peters A, Polasek O, Pramstaller
787 PP, Nguyen QT, Raitakari OT, Ren M, Rettig R, Rice K, Ridker PM, Ried JS,
788 Riese H, Ripatti S, Robino A, Rose LM, Rotter JI, Rudan I, Ruggiero D, Saba Y,
789 Sala CF, Salomaa V, Samani NJ, Sarin A-P, Schmidt R, Schmidt H, Shrine N,
790 Siscovick D, Smith AV, Snieder H, Sõber S, Sorice R, Starr JM, Stott DJ,
791 Strachan DP, Strawbridge RJ, Sundström J, Swertz MA, Taylor KD, Teumer A,
792 Tobin MD, Tomaszewski M, Toniolo D, Traglia M, Trompet S, Tuomilehto J,
793 Tzourio C, Uitterlinden AG, Vaez A, Most PJ van der, Duijn CM van, Vergnaud A-
794 C, Verwoert GC, Vitart V, Völker U, Vollenweider P, Vuckovic D, Watkins H, Wild
795 SH, Willemsen G, Wilson JF, Wright AF, Yao J, Zemunik T, Zhang W, Attia JR,
796 Butterworth AS, Chasman DI, Conen D, Cucca F, Danesh J, Hayward C, Howson
797 JMM, Laakso M, Lakatta EG, Langenberg C, Melander O, Mook-Kanamori DO,
798 Palmer CNA, Risch L, Scott RA, Scott RJ, Sever P, Spector TD, Harst P van der,
799 Wareham NJ, Zeggini E, Levy D, Munroe PB, Newton-Cheh C, Brown MJ,
800 Metspalu A, Hung AM, O'Donnell CJ, Edwards TL, Psaty BM, Tzoulaki I, Barnes
801 MR, Wain LV, Elliott P, Caulfield MJ. 2018. Genetic analysis of over 1 million
802 people identifies 535 new loci associated with blood pressure traits. *Nat Genet*
803 **50**:1412. doi:10.1038/s41588-018-0205-x
804 Ferreira MAR, Hottenga J-J, Warrington NM, Medland SE, Willemsen G, Lawrence RW,
805 Gordon S, de Geus EJC, Henders AK, Smit JH, Campbell MJ, Wallace L, Evans
806 DM, Wright MJ, Nyholt DR, James AL, Beilby JP, Penninx BW, Palmer LJ, Frazer

IH, Montgomery GW, Martin NG, Boomsma DI. 2009. Sequence Variants in Three Loci Influence Monocyte Counts and Erythrocyte Volume. *Am J Hum Genet* **85**:745–749. doi:10.1016/j.ajhg.2009.10.005

Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**:435–444. doi:10.1038/ng.3247

Hatzikotoulas K, Gilly A, Zeggini E. 2014. Using population isolates in genetic association studies. *Brief Funct Genomics* **13**:371–377. doi:10.1093/bfgp/elu022

Hoffmann TJ, Choquet H, Yin J, Banda Y, Kvale MN, Glymour M, Schaefer C, Risch N, Jorgenson E. 2018. A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci. *Genetics* **210**:499–515. doi:10.1534/genetics.118.301479

Howie B, Marchini J, Stephens M. 2011. Genotype Imputation with Thousands of Genomes. *G3 Genes Genomes Genet* **1**:457–470. doi:10.1534/g3.111.001198

Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**:e1000529. doi:10.1371/journal.pgen.1000529

Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, Hamamsy T, Lek M, Samocha KE, Cummings BB, Birnbaum D, Daly MJ, MacArthur DG. 2017. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **45**:D840–D845. doi:10.1093/nar/gkw971

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general

framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**:310–315. doi:10.1038/ng.2892

Kullo IJ, Ding K, Jouni H, Smith CY, Chute CG. 2010. A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLOS ONE* **5**:e13011. doi:10.1371/journal.pone.0013011

Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Linnér RK, Fontana MA, Kundu T, Lee C, Li H, Li R, Royer R, Timshel PN, Walters RK, Willoughby EA, Yengo L, Alver M, Bao Y, Clark DW, Day FR, Furlotte NA, Joshi PK, Kemper KE, Kleinman A, Langenberg C, Mägi R, Trampush JW, Verma SS, Wu Y, Lam M, Zhao JH, Zheng Z, Boardman JD, Campbell H, Freese J, Harris KM, Hayward C, Herd P, Kumari M, Lencz T, Luan J, Malhotra AK, Metspalu A, Milani L, Ong KK, Perry JRB, Porteous DJ, Ritchie MD, Smart MC, Smith BH, Tung JY, Wareham NJ, Wilson JF, Beauchamp JP, Conley DC, Esko T, Lehrer SF, Magnusson PKE, Oskarsson S, Pers TH, Robinson MR, Thom K, Watson C, Chabris CF, Meyer MN, Laibson DI, Yang J, Johannesson M, Koellinger PD, Turley P, Visscher PM, Benjamin DJ, Cesarini D. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**:1112. doi:10.1038/s41588-018-0147-3

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma Oxf Engl* **27**:2987–2993. doi:10.1093/bioinformatics/btr509

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**:589–595. doi:10.1093/bioinformatics/btp698

Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, Hou C, Abrams DJ, Chang D, Gao F, Guo Y, Wei Z, Connolly JJ, Cardinale CJ, Bakay M, Glessner JT, Li D, Kao C, Thomas KA, Qiu H, Chiavacci RM, Kim CE, Wang F, Snyder J, Richie MD, Flatø B, Førre Ø, Denson LA, Thompson SD, Becker ML, Guthery SL, Latiano A, Perez E, Resnick E, Russell RK, Wilson DC, Silverberg MS, Annese V, Lie BA, Punaro M, Dubinsky MC, Monos DS, Strisciuglio C, Staiano A, Miele E, Kugathasan S, Ellis JA, Munro JE, Sullivan KE, Wise CA, Chapel H,

Cunningham-Rundles C, Grant SFA, Orange JS, Sleiman PMA, Behrens EM, Griffiths AM, Satsangi J, Finkel TH, Keinan A, Prak ETL, Polychronakos C, Baldassano RN, Li H, Keating BJ, Hakonarson H. 2015. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med* **21**:1018–1027. doi:10.1038/nm.3933

Liu C, Kraja AT, Smith JA, Brody JA, Franceschini N, Bis JC, Rice K, Morrison AC, Lu Y, Weiss S, Guo X, Palmas W, Martin LW, Chen Y-DI, Surendran P, Drenos F, Cook JP, Auer PL, Chu AY, Giri A, Zhao W, Jakobsdottir J, Lin L-A, Stafford JM, Amin N, Mei H, Yao J, Voorman A, CHD Exome+ Consortium, ExomeBP Consortium, GoT2DGenes Consortium, T2d-Genes Consortium, Larson MG, Grove ML, Smith AV, Hwang S-J, Chen H, Huan T, Kosova G, Stitziel NO, Kathiresan S, Samani N, Schunkert H, Deloukas P, Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia, Li M, Fuchsberger C, Pattaro C, Gorski M, CKDGen Consortium, Kooperberg C, Papanicolaou GJ, Rossouw JE, Faul JD, Kardia SLR, Bouchard C, Raffel LJ, Uitterlinden AG, Franco OH, Vasan RS, O'Donnell CJ, Taylor KD, Liu K, Bottinger EP, Gottesman O, Daw EW, Giulianini F, Ganesh S, Salfati E, Harris TB, Launer LJ, Dörr M, Felix SB, Rettig R, Völzke H, Kim E, Lee W-J, Lee I-T, Sheu WH-H, Tsosie KS, Edwards DRV, Liu Y, Correa A, Weir DR, Völker U, Ridker PM, Boerwinkle E, Gudnason V, Reiner AP, van Duijn CM, Borecki IB, Edwards TL, Chakravarti A, Rotter JI, Psaty BM, Loos RJF, Fornage M, Ehret GB, Newton-Cheh C, Levy D, Chasman DI. 2016. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat Genet* **48**:1162–1170. doi:10.1038/ng.3660

Lucas A. 2018. An R package for creating mirrored Manhattan plots: anastasia-lucas/hudson.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow

C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**:823–828. doi:10.1126/science.1215040

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**:499–511. doi:10.1038/nrg2796

Martinelli-Boneschi F, Colombi M, Castori M, Devigili G, Eleopra R, Malik RA, Ritelli M, Zoppi N, Dordoni C, Sorosina M, Grammatico P, Fadavi H, Gerrits MM, Almomani R, Faber CG, Merkies ISJ, Toniolo D, Cocca M, Doglioni C, Waxman SG, Dib-Hajj SD, Taiana MM, Sassone J, Lombardi R, Cazzato D, Zauli A, Santoro S, Marchi M, Lauria G. n.d. COL6A5 variants in familial neuropathic chronic itch. *Brain*. doi:10.1093/brain/aww343

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki A-E, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PIW, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R, the Haplotype Reference Consortium. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**:1279–1283. doi:10.1038/ng.3643

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**:1297–1303. doi:10.1101/gr.107524.110

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinforma Oxf Engl* **26**:2069–2070. doi:10.1093/bioinformatics/btq330

Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates C, Bellary S, Bockett NA, Giorda K, Griffiths CJ, Hemingway H, Jia Z, Kelly MA, Khawaja HA, Lek M, McCarthy S, McEachan R, O'Donnell-Luria A, Paigen K, Parisinos CA, Sheridan E, Southgate L, Tee L, Thomas M, Xue Y, Schnall-Levin M, Petkov PM, Tyler-Smith C, Maher ER, Trembath RC, MacArthur DG, Wright J, Durbin R, Heel DA van. 2016. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**:474–477. doi:10.1126/science.aac8624

Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, Surakka I, Mathis MR, Yamazaki M, Crawford RD, Gabrielsen ME, Skogholt AH, Holmen OL, Lin M, Wolford BN, Dey R, Dalen H, Sulem P, Chung JH, Backman JD, Arnar DO, Thorsteinsdottir U, Baras A, O'Dushlaine C, Holst AG, Wen X, Hornsby W, Dewey FE, Boehnke M, Kheterpal S, Mukherjee B, Lee S, Kang HM, Holm H, Kitzman J, Shavit JA, Jalife J, Brummett CM, Teslovich TM, Carey DJ, Gudbjartsson DF, Stefansson K, Abecasis GR, Hveem K, Willer CJ. 2018. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* **50**:1234. doi:10.1038/s41588-018-0171-3

Perry JRB, 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, Paternoster L, Porcu E, Smith AV, Stolk L, Teumer A, Tšernikova N, Tikkanen E, Ulivi S, Wagner EK, Amin N, Bierut LJ, Byrne EM, Hottenga J-J, Koller DL, Mangino M, Pers TH, Yerges-Armstrong LM, Hua Zhao J, Andrulis IL,

962 Anton-Culver H, Atsma F, Bandinelli S, Beckmann MW, Benitez J, Blomqvist C,
963 Bojesen SE, Bolla MK, Bonanni B, Brauch H, Brenner H, Buring JE, Chang-
964 Claude J, Chanock S, Chen J, Chenevix-Trench G, Collée JM, Couch FJ, Couper
965 D, Coviello AD, Cox A, Czene K, D'adamo AP, Davey Smith G, De Vivo I,
966 Demerath EW, Dennis J, Devilee P, Dieffenbach AK, Dunning AM, Eiriksdottir G,
967 Eriksson JG, Fasching PA, Ferrucci L, Flesch-Janys D, Flyger H, Foroud T,
968 Franke L, Garcia ME, García-Closas M, Geller F, de Geus EEJ, Giles GG,
969 Gudbjartsson DF, Gudnason V, Guénel P, Guo S, Hall P, Hamann U, Haring R,
970 Hartman CA, Heath AC, Hofman A, Hooning MJ, Hopper JL, Hu FB, Hunter DJ,
971 Karasik D, Kiel DP, Knight JA, Kosma V-M, Kutalik Z, Lai S, Lambrechts D,
972 Lindblom A, Mägi R, Magnusson PK, Mannermaa A, Martin NG, Masson G,
973 McArdle PF, McArdle WL, Melbye M, Michailidou K, Mihailov E, Milani L, Milne
974 RL, Nevanlinna H, Neven P, Nohr EA, Oldehinkel AJ, Oostra BA, Palotie A,
975 Peacock M, Pedersen NL, Peterlongo P, Peto J, Pharoah PDP, Postma DS,
976 Pouta A, Pylkäs K, Radice P, Ring S, Rivadeneira F, Robino A, Rose LM,
977 Rudolph A, Salomaa V, Sanna S, Schlessinger D, Schmidt MK, Southey MC,
978 Sovio U, Stampfer MJ, Stöckl D, Storniolo AM, Timpson NJ, Tyrer J, Visser JA,
979 Vollenweider P, Völzke H, Waeber G, Waldenberger M, Wallaschofski H, Wang
980 Q, Willemsen G, Winqvist R, Wolffenbuttel BHR, Wright MJ, Australian Ovarian
981 Cancer Study, The GENICA Network, kConFab, The LifeLines Cohort Study, The
982 InterAct Consortium, Early Growth Genetics (EGG) Consortium, Boomsma DI,
983 Econs MJ, Khaw K-T, Loos RJF, McCarthy MI, Montgomery GW, Rice JP,
984 Streeten EA, Thorsteinsdottir U, van Duijn CM, Alizadeh BZ, Bergmann S,
985 Boerwinkle E, Boyd HA, Crisponi L, Gasparini P, Gieger C, Harris TB, Ingelsson
986 E, Järvelin M-R, Kraft P, Lawlor D, Metspalu A, Pennell CE, Ridker PM, Snieder
987 H, Sørensen TIA, Spector TD, Strachan DP, Uitterlinden AG, Wareham NJ,
988 Widen E, Zygmunt M, Murray A, Easton DF, Stefansson K, Murabito JM, Ong
989 KK. 2014. Parent-of-origin-specific allelic associations among 106 genomic loci
990 for age at menarche. *Nature* **514**:92–97. doi:10.1038/nature13545
991 Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. 2016. Detection and
992 interpretation of shared genetic influences on 42 human traits. *Nat Genet*

48:709–717. doi:10.1038/ng.3570

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genet* **8**:e1002967. doi:10.1371/journal.pgen.1002967

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. 2015. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* **23**:975–983. doi:10.1038/ejhg.2014.216

Sazzini M, Gnecchi Ruscone GA, Giuliani C, Sarno S, Quagliariello A, De Fanti S, Boattini A, Gentilini D, Fiorito G, Catanoso M, Boiardi L, Croci S, Macchioni P, Mantovani V, Di Blasio AM, Matullo G, Salvarani C, Franceschi C, Pettener D, Garagnani P, Luiselli D. 2016. Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. *Sci Rep* **6**:32513. doi:10.1038/srep32513

Schormair B, Zhao C, Bell S, Tilch E, Salminen AV, Pütz B, Dauvilliers Y, Stefani A, Högl B, Poewe W, Kemlink D, Sonka K, Bachmann CG, Paulus W, Trenkwalder C, Oertel WH, Hornyak M, Teder-Laving M, Metspalu A, Hadjigeorgiou GM, Polo O, Fietze I, Ross OA, Wszolek Z, Butterworth AS, Soranzo N, Ouwehand WH, Roberts DJ, Danesh J, Allen RP, Earley CJ, Ondo WG, Xiong L, Montplaisir J, Gan-Or Z, Perola M, Vodicka P, Dina C, Franke A, Tittmann L, Stewart AFR, Shah SH, Gieger C, Peters A, Rouleau GA, Berger K, Oexle K, Di Angelantonio E, Hinds DA, Müller-Myhsok B, Winkelmann J, Balkau B, Ducimetière P, Eschwège E, Rancière F, Alhenc-Gelas F, Gallois Y, Girault A, Fumeron F, Marre M, Roussel R, Bonnet F, Bonnefond A, Cauchi S, Froguel P, Cogneau J, Born C, Caces E, Cailleau M, Lantieri O, Moreau J, Rakotozafy F, Tichet J, Vol S, Agee M, Alipanahi B, Auton A, Bell RK, Bryc K, Elson SL, Fontanillas P, Furlotte NA, Hinds DA, Hromatka BS, Huber KE, Kleinman A, Litterman NK, McIntyre MH, Mountain JL, Northover CA, Pitts SJ, Sathirapongsasuti JF, Sazonova OV,

1024 Shelton JF, Shringarpure S, Tian C, Tung JY, Vacic V, Wilson CH. 2017.
1025 Identification of novel risk loci for restless legs syndrome in genome-wide
1026 association studies in individuals of European ancestry: a meta-analysis. *Lancet*
1027 *Neurol* **16**:898–907. doi:10.1016/S1474-4422(17)30327-7
1028 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001.
1029 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**:308–311.
1030 Steensel MAM van, Steijlen PM, Bladergroen RS, Vermeer M, Geel M van. 2005. A
1031 missense mutation in the type II hair keratin hHb3 is associated with monilethrix.
1032 *J Med Genet* **42**:e19–e19. doi:10.1136/jmg.2004.021030
1033 Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y,
1034 Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo
1035 Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP,
1036 Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X,
1037 Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding
1038 L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer E-W,
1039 McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM,
1040 Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt
1041 EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker
1042 JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T,
1043 Sebat J, Batzer MA, McCarroll SA, The 1000 Genomes Project Consortium, Mills
1044 RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO.
1045 2015. An integrated map of structural variation in 2,504 human genomes. *Nature*
1046 **526**:75–81. doi:10.1038/nature15394
1047 Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang
1048 T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK,
1049 Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR,
1050 Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS,
1051 Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS. 2018. Genomic atlas of
1052 the human plasma proteome. *Nature* **558**:73. doi:10.1038/s41586-018-0175-2
1053 Szpiech ZA, Hernandez RD. 2014. selscan: An Efficient Multithreaded Program to
1054 Perform EHH-Based Scans for Positive Selection. *Mol Biol Evol* **31**:2824–2827.

doi:10.1093/molbev/msu211

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, Melander O, Johnson T, Li X, Guo X, Li M, Shin Cho Y, Jin Go M, Jin Kim Y, Lee J-Y, Park T, Kim K, Sim X, Twee-Hee Ong R, Croteau-Chonka DC, Lange LA, Smith JD, Song K, Hua Zhao J, Yuan X, Luan J, Lamina C, Ziegler A, Zhang W, Zee RYL, Wright AF, Witteman JCM, Wilson JF, Willemsen G, Wichmann H-E, Whitfield JB, Waterworth DM, Wareham NJ, Waeber G, Vollenweider P, Voight BF, Vitart V, Uitterlinden AG, Uda M, Tuomilehto J, Thompson JR, Tanaka T, Surakka I, Stringham HM, Spector TD, Soranzo N, Smit JH, Sinisalo J, Silander K, Sijbrands EJG, Scuteri A, Scott J, Schlessinger D, Sanna S, Salomaa V, Saharinen J, Sabatti C, Ruokonen A, Rudan I, Rose LM, Roberts R, Rieder M, Psaty BM, Pramstaller PP, Pichler I, Perola M, Penninx BWJH, Pedersen NL, Pattaro C, Parker AN, Pare G, Oostra BA, O'Donnell CJ, Nieminen MS, Nickerson DA, Montgomery GW, Meitinger T, McPherson R, McCarthy MI, McArdle W, Masson D, Martin NG, Marroni F, Mangino M, Magnusson PKE, Lucas G, Luben R, Loos RJF, Lokki M-L, Lettre G, Langenberg C, Launer LJ, Lakatta EG, Laaksonen R, Kyvik KO, Kronenberg F, König IR, Khaw K-T, Kaprio J, Kaplan LM, Johansson Å, Jarvelin M-R, Cecile J. W. Janssens A, Ingelsson E, Igl W, Kees Hovingh G, Hottenga J-J, Hofman A, Hicks AA, Hengstenberg C, Heid IM, Hayward C, Havulinna AS, Hastie ND, Harris TB, Haritunians T, Hall AS, Gyllensten U, Guiducci C, Groop LC, Gonzalez E, Gieger C, Freimer NB, Ferrucci L, Erdmann J, Elliott P, Ejebe KG, Döring A, Dominiczak AF, Demissie S, Deloukas P, de Geus EJC, de Faire U, Crawford G, Collins FS, Chen YI, Caulfield MJ, Campbell H, Burtt NP, Bonnycastle LL, Boomsma DI, Boekholdt SM, Bergman RN, Barroso I, Bandinelli S, Ballantyne CM, Assimes TL, Quertermous T, Altshuler D, Seielstad M, Wong TY, Tai E-S, Feranil AB, Kuzawa CW, Adair LS, Taylor Jr HA, Borecki IB, Gabriel SB, Wilson JG, Holm H, Thorsteinsdottir U, Gudnason V, Krauss RM, Mohlke KL, Ordovas JM, Munroe PB, Kooner JS, Tall AR, Hegele

RA, Kastelein JJP, Schadt EE, Rotter JI, Boerwinkle E, Strachan DP, Mooser V, Stefansson K, Reilly MP, Samani NJ, Schunkert H, Cupples LA, Sandhu MS, Ridker PM, Rader DJ, van Duijn CM, Peltonen L, Abecasis GR, Boehnke M, Kathiresan S. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**:707–713. doi:10.1038/nature09270

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

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**:799–816. doi:10.1038/nature05874

The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**:82–90. doi:10.1038/nature14962

Turner S. 2017. qqman: Q-Q and Manhattan Plots for GWAS Data.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. *PLOS Biol* **4**:e72. doi:10.1371/journal.pbio.0040072

Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinforma Oxf Engl* **26**:2190–2191. doi:10.1093/bioinformatics/btq340

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, Mägi R, Mihailov E, Porcu E, Randall JC, Scherag A, Vinkhuyzen AAE, Westra H-J, Winkler TW, Workalemahu T, Zhao JH, Absher D, Albrecht E, Anderson D, Baron J, Beekman M, Demirkan A, Ehret GB, Feenstra B, Feitosa MF, Fischer K, Fraser RM, Goel A, Gong J, Justice AE, Kanoni S, Kleber ME, Kristiansson K, Lim U, Lotay V, Lui JC, Mangino M, Leach IM, Medina-Gomez C, Nalls MA, Nyholt DR, Palmer CD, Pasko D, Pechlivanis S, Prokopenko I, Ried JS, Ripke S, Shungin D, Stancáková A, Strawbridge RJ, Sung YJ, Tanaka T, Teumer A, Trompet S, van der Laan SW, van Setten J, Van Vliet-Ostaptchouk JV, Wang Z, Yengo L, Zhang W, Afzal U, Ärnlöv J, Arscott GM, Bandinelli S, Barrett A, Bellis

C, Bennett AJ, Berne C, Blüher M, Bolton JL, Böttcher Y, Boyd HA, Bruinenberg M, Buckley BM, Buyske S, Caspersen IH, Chines PS, Clarke R, Claudi-Boehm S, Cooper M, Daw EW, De Jong PA, Deelen J, Delgado G, Denny JC, Dhonukshe-Rutten R, Dimitriou M, Doney ASF, Dörr M, Eklund N, Eury E, Folkersen L, Garcia ME, Geller F, Giedraitis V, Go AS, Grallert H, Grammer TB, Gräßler J, Grönberg H, de Groot LCPGM, Groves CJ, Haessler J, Hall P, Haller T, Hallmans G, Hannemann A, Hartman CA, Hassinen M, Hayward C, Heard-Costa NL, Helmer Q, Hemani G, Henders AK, Hillege HL, Hlatky MA, Hoffmann W, Hoffmann P, Holmen O, Houwing-Duistermaat JJ, Illig T, Isaacs A, James AL, Jeff J, Johansen B, Johansson Å, Jolley J, Juliusdottir T, Junttila J, Kho AN, Kinnunen L, Klopp N, Kocher T, Kratzer W, Lichtner P, Lind L, Lindström J, Lobbens S, Lorentzon M, Lu Y, Lyssenko V, Magnusson PKE, Mahajan A, Maillard M, McArdle WL, McKenzie CA, McLachlan S, McLaren PJ, Menni C, Merger S, Milani L, Moayyeri A, Monda KL, Morken MA, Müller G, Müller-Nurasyid M, Musk AW, Narisu N, Nauck M, Nolte IM, Nöthen MM, Oozageer L, Pilz S, Rayner NW, Renstrom F, Robertson NR, Rose LM, Roussel R, Sanna S, Scharnagl H, Scholtens S, Schumacher FR, Schunkert H, Scott RA, Sehmi J, Seufferlein T, Shi J, Silventoinen K, Smit JH, Smith AV, Smolonska J, Stanton AV, Stirrups K, Stott DJ, Stringham HM, Sundström J, Swertz MA, Syvänen A-C, Tayo BO, Thorleifsson G, Tyrer JP, van Dijk S, van Schoor NM, van der Velde N, van Heemst D, van Oort FVA, Vermeulen SH, Verweij N, Vonk JM, Waite LL, Waldenberger M, Wennauer R, Wilkens LR, Willenborg C, Wilsgaard T, Wojczynski MK, Wong A, Wright AF, Zhang Q, Arveiler D, Bakker SJL, Beilby J, Bergman RN, Bergmann S, Biffar R, Blangero J, Boomsma DI, Bornstein SR, Bovet P, Brambilla P, Brown MJ, Campbell H, Caulfield MJ, Chakravarti A, Collins R, Collins FS, Crawford DC, Cupples LA, Danesh J, de Faire U, den Ruijter HM, Erbel R, Erdmann J, Eriksson JG, Farrall M, Ferrannini E, Ferrières J, Ford I, Forouhi NG, Forrester T, Gansevoort RT, Gejman PV, Gieger C, Golay A, Gottesman O, Gudnason V, Gyllensten U, Haas DW, Hall AS, Harris TB, Hattersley AT, Heath AC, Hengstenberg C, Hicks AA, Hindorff LA, Hingorani AD, Hofman A, Hovingh GK, Humphries SE, Hunt SC, Hypponen E, Jacobs KB,

Jarvelin M-R, Jousilahti P, Jula AM, Kaprio J, Kastelein JJP, Kayser M, Kee F, Keinanen-Kiukaanniemi SM, Kiemeney LA, Kooner JS, Kooperberg C, Koskinen S, Kovacs P, Kraja AT, Kumari M, Kuusisto J, Lakka TA, Langenberg C, Le Marchand L, Lehtimäki T, Lupoli S, Madden PAF, Männistö S, Manunta P, Marette A, Matise TC, McKnight B, Meitinger T, Moll FL, Montgomery GW, Morris AD, Morris AP, Murray JC, Nelis M, Ohlsson C, Oldehinkel AJ, Ong KK, Ouwehand WH, Pasterkamp G, Peters A, Pramstaller PP, Price JF, Qi L, Raitakari OT, Rankinen T, Rao DC, Rice TK, Ritchie M, Rudan I, Salomaa V, Samani NJ, Saramies J, Sarzynski MA, Schwarz PEH, Sebert S, Sever P, Shuldiner AR, Sinisalo J, Steinthorsdottir V, Stolk RP, Tardif J-C, Tönjes A, Tremblay A, Tremoli E, Virtamo J, Vohl M-C, The Electronic Medical Records and Genomics (eMERGE) Consortium, The MIGen Consortium, The PAGE Consortium, The LifeLines Cohort Study, Amouyel P, Asselbergs FW, Assimes TL, Bochud M, Boehm BO, Boerwinkle E, Bottinger EP, Bouchard C, Cauchi S, Chambers JC, Chanock SJ, Cooper RS, de Bakker PIW, Dedoussis G, Ferrucci L, Franks PW, Froguel P, Groop LC, Haiman CA, Hamsten A, Hayes MG, Hui J, Hunter DJ, Hveem K, Jukema JW, Kaplan RC, Kivimaki M, Kuh D, Laakso M, Liu Y, Martin NG, März W, Melbye M, Moebus S, Munroe PB, Njølstad I, Oostra BA, Palmer CNA, Pedersen NL, Perola M, Pérusse L, Peters U, Powell JE, Power C, Quertermous T, Rauramaa R, Reinmaa E, Ridker PM, Rivadeneira F, Rotter JI, Saaristo TE, Saleheen D, Schlessinger D, Slagboom PE, Snieder H, Spector TD, Strauch K, Stumvoll M, Tuomilehto J, Uusitupa M, van der Harst P, Völzke H, Walker M, Wareham NJ, Watkins H, Wichmann H-E, Wilson JF, Zanen P, Deloukas P, Heid IM, Lindgren CM, Mohlke KL, Speliotes EK, Thorsteinsdottir U, Barroso I, Fox CS, North KE, Strachan DP, Beckmann JS, Berndt SI, Boehnke M, Borecki IB, McCarthy MI, Metspalu A, Stefansson K, Uitterlinden AG, van Duijn CM, Franke L, Willer CJ, Price AL, Lettre G, Loos RJF, Weedon MN, Ingelsson E, O'Connell JR, Abecasis GR, Chasman DI, Goddard ME, Visscher PM, Hirschhorn JN, Frayling TM. 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**:1173–1186. doi:10.1038/ng.3097

Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, Gilly A, Ayub Q, Colonna V, Southam L, Finan C, Massaia A, Chheda H, Palta P, Ritchie G, Asimit J, Dedoussis G, Gasparini P, Palotie A, Ripatti S, Soranzo N, Toniolo D, Wilson JF, Durbin R, Tyler-Smith C, Zeggini E. 2017. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* **8**. doi:10.1038/ncomms15927

Yang Q, Kathiresan S, Lin J-P, Tofler GH, O'Donnell CJ. 2007. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BMC Med Genet* **8**:S12. doi:10.1186/1471-2350-8-S1-S12

Zhang Q, Marioni RE, Robinson MR, Higham J, Sproul D, Wray NR, Deary IJ, McRae AF, Visscher PM. 2018. Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Med* **10**:75. doi:10.1186/s13073-018-0585-7

Table 1. Final data release of WGS data for all the INGI cohorts.

INGI All samples				
	CAR	**FVG**	**VBI**	**INGI**
Samples	124	378	424	926
Females	66	220	249	535
Males	58	158	175	391
Average coverage	6.31	7.23	6.12	6.55
Sites	13,370,262	17,002,010	19,361,094	26,619,091
Multiallelic Sites	248,638	356,599	393,328	560,918
SNPs	12,208,629	15,521,313	17,830,208	24,557,366
INDELs	1,161,633	1,480,697	1,530,886	2,061,725
Sites MAF <= 1%	3,627,622	7,283,720	9,416,028	16,685,951
Sites 1% < MAF <= 5%	3,007,162	3,069,534	3,121,545	3,125,971
Sites MAF > 5%	6,735,478	6,648,756	6,823,521	7,123,064
Singletons SNPs	2,061,824	2,784,746	3,554,744	6,193,486
Singletons INDELs	92,372	131,275	133,156	273,679
Average Heterozygosity rate per sample	17.57%	13.27%	12.16%	13.34%
Average Derived allele count per sample	4,703,290	4,741,910	4,844,980	4,763,393
Average variations per sample	3,518,020	3,421,910	3,541,760	3,493,897
Average INDELs per sample	531,151	586,740	590,109	569,333
Average singleton per sample	17,285	7,671	8,646	6,925