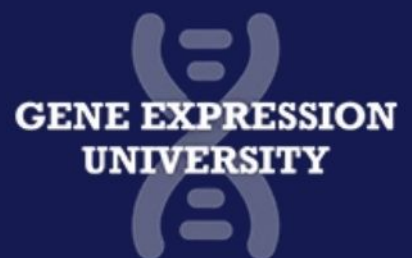




December 1st, 12pm EST

**THE ULTIMATE
VIRTUAL SEMINAR
SERIES IN GENE
EXPRESSION STUDIES**




[Register Now>>](#)

**applied
biosystems**
by Thermo Fisher Scientific

ThermoFisher
SCIENTIFIC

WILEY

Functional and clinical implications of genetic structure in 1686 Italian exomes

Giovanni Birolo¹  | Serena Aneli¹  | Cornelia Di Gaetano¹  |
Giovanni Cugliari¹  | Alessia Russo¹  | Alessandra Allione¹  |
Elisabetta Casalone¹  | Elisa Giorgio¹  | Elvezia M. Paraboschi^{2,3}  |
Diego Ardissino⁴  | Stefano Duga^{2,3}  | Rosanna Asselta^{2,3}  |
Giuseppe Matullo¹ 

¹Department of Medical Sciences, University of Turin, Turin, Italy

²Department of Biomedical Sciences, Humanitas University, Rozzano, Milan, Italy

³Humanitas Clinical and Research Center-IRCCS, Rozzano, Milan, Italy

⁴Division of Cardiology, Azienda Ospedaliero-Universitaria di Parma, Parma, Italy

Correspondence

Giovanni Birolo, Serena Aneli, and Giuseppe Matullo, Department of Medical Sciences, University of Turin, Turin, Italy.

Email: giovanni.birolo@unito.it (G. B.); serena.aneli@unito.it (S. A.) and giuseppe.matullo@unito.it (G. M.)

Funding information

Associazione Italiana per la Ricerca sul Cancro, Grant/Award Number: IG 2018 Id.21390; Compagnia di San Paolo and the Italian Institute for Genomic Medicine, IIGM, Grant/Award Number: 2017-2018 to Dr. Matullo; Ministero dell'Istruzione, dell'Università e della Ricerca, Grant/Award Number: D15D18000410001

Abstract

To reconstruct the phenotypical and clinical implications of the Italian genetic structure, we thoroughly analyzed a whole-exome sequencing data set comprised of 1686 healthy Italian individuals. We found six previously unreported variants with remarkable frequency differences between Northern and Southern Italy in the *HERC2*, *OR52R1*, *ADH1B*, and *THBS4* genes. We reported 36 clinically relevant variants (submitted as pathogenic, risk factors, or drug response in ClinVar) with significant frequency differences between Italy and Europe. We then explored putatively pathogenic variants in the Italian exome. On average, our Italian individuals carried 16.6 protein-truncating variants (PTVs), with 2.5% of the population having a PTV in one of the 59 American College of Medical Genetics (ACMG) actionable genes. Lastly, we looked for PTVs that are likely to cause Mendelian diseases. We found four heterozygous PTVs in haploinsufficient genes (*KAT6A*, *PTCH1*, and *STXBP1*) and three homozygous PTVs in genes causing recessive diseases (*DPYD*, *FLG*, and *PYGM*). Comparing frequencies from our data set to other public databases, like gnomAD, we showed the importance of population-specific databases for a more accurate assessment of variant pathogenicity. For this reason, we made aggregated frequencies from our data set publicly available as a tool for both clinicians and researchers (<http://nigdb.cineca.it>; NIG-ExIT).

KEYWORDS

genetic frequency database, genomic medicine, Italian population, pathogenic variants, rare variants, whole-exome sequencing

1 | INTRODUCTION

Despite the large amount of genomic data published in the last few years, identifying functionally important variations for the interpretation of personalized disease-risk profiles remains challenging.

The vast majority of coding variants, which are predicted to harbor most of the disease-causing variants are rare, and large-scale sequencing data sets are needed to adequately detect them and estimate their frequencies (Gibson, 2012; Keinan & Clark, 2012). Frequency databases, including rare variants, are essential for identifying the genetic causes of Mendelian disorders and, through gene-based burden testing approaches, understanding the complex

Giovanni Birolo and Serena Aneli contributed equally to this study.

genetic bases of common diseases (Gibson, 2012). Moreover, while the most common genetic variation is shared worldwide, rare variants, due to their recent origin, tend to be more geographically clustered in specific populations (Tennessen et al., 2012). Therefore, the availability of large population-specific data sets built from high-quality sequencing data is crucial for evaluating the role of rare variations in disease susceptibility and for exploring fine-scale genetic structure in a population. In the last few years, international projects and consortia have collected and made publicly available several data sets of human DNA sequence variation, such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), the Exome Aggregation Consortium (ExAC; Lek et al., 2016), and the Genome Aggregation Database (gnomAD; Karczewski et al., 2020). While in the 1000 Genomes Project, the genotypes of all individuals are available, ExAC and gnomAD can only be consulted in an aggregated manner (with some stratification), thus making it impossible to access country-specific genetic variation or individual genotypes.

This is a major inconvenience when studying the genetic variation in populations with a high degree of genetic structure, such as the Italian one. The Italian population has higher genetic variability than the other European populations with a well-defined internal structure (Fiorito et al., 2016; Raveane et al., 2019). This is mainly due to the geographic location of Italy, which is separated from continental Europe by the Alps and enclosed by the Mediterranean Sea. The Mediterranean Sea itself played a major role in the dispersion and admixture of human groups by acting for millennia as a barrier separating the African and the European continent and then turning into a bridge as first Bronze Age seafarers started to cross the open water (Antonio et al., 2019).

The genetic structure of the Italian population has been deeply investigated using single-nucleotide polymorphism (SNP) array data, which are mainly comprised of genome-wide common genetic markers (Di Gaetano et al., 2012; Fiorito et al., 2016; Parolo et al., 2015; Raveane et al., 2019; Sazzini et al., 2016). In this context, the exploitation of rarer genetic variation could provide new insight into more recent demographic events. In addition to the well-known studies exploring human genetic variation worldwide from whole-exome data, many groups (Dopazo et al., 2016; Kwak et al., 2017; Van Hout et al., 2020) have worked on national sequencing data sets with three main goals in mind: (i) studying the genetic structure of a population by also exploiting lower frequency variants, (ii) understanding the distribution of putative pathogenic variation in healthy cohorts, and, ultimately, (iii) generating a catalog of local variability. Previous large sequencing-based studies, which contained Italian participants include the 1000 Genomes Project (with 107 Tuscans) and more recent studies that focused on specific isolates (Cocca et al., 2019; Nutile et al., 2019; Sidore et al., 2015; Xue et al., 2017).

In this study, we analyzed a whole-exome sequencing (WES) data set comprising 1686 healthy Italian individuals. This is the first large-scale study based on sequencing data of the nonisolated Italian population covering many different geographical regions of the country. WES provided a large amount of both common and rare functional

variants that allowed us to give a more functional and clinical picture of the Italian genetic structure than previous array-based studies.

We began with a general picture of the genetic patterns in Italy and their possible interpretation in terms of phenotypes and diseases. We then moved to more clinically oriented analyses using our Italian sample to show how the availability of population-specific databases of allele frequencies can increase accuracy in the identification of pathogenic variants. Finally, we explored the burden of putatively pathogenic variants in our data set by focusing specifically on the American College of Medical Genetics (ACMG) secondary findings genes (ACMG SF v2.0; Kalia et al., 2017) and genes involved in Mendelian diseases.

2 | MATERIALS AND METHODS

2.1 | Study sample, editorial policies, and ethical considerations

We obtained variant calls from WES data of 1751 healthy controls enrolled in the Italian Genetic Study on early-onset myocardial infarction (Atherosclerosis Thrombosis and Vascular Biology Italian Study Group, 2003) and already analyzed as part of the “Myocardial Infarction Genetics Consortium” (Migen; Do et al., 2015). All participants in the study provided written informed consent for genetic studies. The institutional review boards at the Broad Institute and at the Ethical Committee of the Ospedale Niguarda, Milan (Italy) approved the study protocol. The ancestry information comprises the place of birth of the individuals, their parents, and their grandparents: 1235 individuals had complete information, 174 lacked the birthplace of their grandparents, 339 lacked both parents and grandparents, and 3 had no birthplace themselves.

2.2 | Sequencing

The cluster amplification, sequencing, read-mapping, and variant calling were performed by the Broad Institute as described in Do et al. (2015). Samples were kept when the read depth was 20× or greater on at least 80% of the exome target.

2.3 | Data cleaning

To produce a data set of high-quality genotypes of unrelated individuals with reasonably certain Italian ancestry for analysis, we performed several filtering steps starting from the original multi-sample variant call format (VCF) file comprised of 1,373,696 variants and 1751 individuals.

We began by removing 26 individuals with reportedly non-Italian ancestry (even partial), leaving 1725 individuals. Then, the reported sex of the subjects was cross-checked with sex inferred from variant calls (with bcftools+guess-ploidy, v1.5) and eight

individuals with discordant sex were removed (most likely misreported in the database), leaving 1717 individuals.

For genotypes, low-quality genotype calls, specifically calls with low read depth ($DP < 10$), calls with very high read depth ($DP > 180$, which is three standard deviations more than the mean depth of 60), and calls with low confidence ($GQ < 20$) were set to missing. Genotype calls in nonautosomal regions in males were converted to hemizygous. Then, we applied a second low-quality variant filter removing variants with the following criteria: missing genotypes for more than 10% of individuals (88,492 variants), extreme deviations ($p < 10^{-10}$) from Hardy–Weinberg equilibrium (2811 variants), and location in low complexity regions of the genome as described in X. Li and Kahveci (2006; 4613 variants). Approximately 93% of the variants (1,187,119) remained after the above filters.

A second filtering step on individuals was applied using the remaining variants: We verified that all individuals had $< 5\%$ of missing genotypes (thus removing no individual). Related individuals (up to second degree) were inferred from their genotypes and only one relative was kept, leaving 1688 individuals.

The last filtering step on individuals was the outlier removal according to principal component analysis (PCA). PCA was performed with PLINK v1.9 (Purcell et al., 2007) using variants with major allele frequency at least 0.2% and pruning by linkage disequilibrium (LD) at $r^2 = 0.2$. Outliers, defined as samples with Euclidean distance (computed from the first two PCs) $> 10 SDs$ from the mean position of all samples, were removed iteratively, recomputing the PCA until no more outliers were detected. Two samples were removed in one iteration as genetic outliers.

Finally, variants, where all individuals were homozygous for the reference allele, were removed leaving a data set of 669,718 variants for 1686 unrelated individuals of Italian ancestry. This data set was used in all analyses looking at the Italian population as a whole.

To perform comparisons between different macro-areas within Italy, we selected individuals with a well-defined ancestry at the macro-area level removing those with uncertain or likely misreported ancestry. From the previous data set of 1686 individuals, we selected the 1197 individuals who had information about all four grandparents' birthplaces available. We assigned them to a macro-area if all their grandparents were born there. We performed further iterative PCA-based outlier removal on individuals with more than 3.5 SDs from the center of their cluster, removing 43 individuals in seven iterations, leaving 1154 individuals. This data set was used in all analyses comparing different macro-areas. When comparing single administrative regions, we furthermore selected those individuals whose grandparents were born in the same region (Table S1).

2.4 | Variant annotation and interpretation

Variants were annotated by the software *vcfanno* (Pedersen et al., 2016), version 0.3.0, with the following databases: dbNSFP (X. Liu et al., 2016), version 3.5a; ClinVar (Landrum et al., 2016), release 20200615; and gnomAD exomes, version 2.0.1.

Functional annotation was performed with SnpEff (Cingolani et al., 2012), version 4.3t, with respect to the canonical RefSeq transcript, except for ACMG SF v2.0 actionable genes for which we used the transcript that occurred most frequently in the ClinVar annotations of pathogenic variants. Variants were labeled as protein-truncating variants (PTVs) when their allele frequency in the whole data set was below 5% and their reported effect in the SnpEff annotation was one of the following: *frameshift_variant*, *splice_acceptor_variant*, *splice_donor_variant*, or *stop_gained*. Missense variants were evaluated with seven pathogenicity predictors: MutPred (B. Li et al., 2009), VEST 3 (Carter et al., 2013), REVEL (Ioannidis et al., 2016), fathmm with rankscore at least 0.73, M-CAP (Jagadeesh et al., 2016), and MetaSVM (Dong et al., 2015) from dbNSFP and CADD (version 1.4; Kircher et al., 2014) with score at least 25. Missense variants were considered damaging (DMG) when at least five predictors out of seven supported this conclusion and their allele frequency in the whole data set was $< 5\%$. Variants that were annotated as “pathogenic” or “likely pathogenic” without any other conflicting annotation and whose allele frequency in the whole data set was $< 5\%$ were labeled CLNPAT variants.

To highlight variants with a putative role in pharmacogenetics, we annotated them with the specialized public repository PharmGKB (Whirl-Carrillo et al., 2012).

We then computed the number of “dominant” genes with at least one heterozygous PTV and the number of “recessive” genes with a homozygous PTV, and we annotated these genes with the OMIM database (McKusick, 1998). Finally, we selected genes harboring PTVs and with a probability of being loss-of-function intolerant (pLI) equal to 1. As above, we annotated these genes with OMIM. The genotypes reported in Section 3.7 are of good quality, with coverage > 20 and balanced allele depth for heterozygous variants. We further applied Loss-Of-Function Transcript Effect Estimator (LOFTEE), which is an Ensembl Variant Effect Predictor (VEP) plugin used to identify high-confidence loss-of-function variants.

2.5 | Exploring the genetic structure with coding variants

PCAs on the whole data set and on macro-areas using only individuals with a “well-defined” macro-area were performed with PLINK v1.9 (Purcell et al., 2007) with a major allele frequency of at least 0.2% and pruning by LD at $0.2 r^2$. On the same data set, we inferred pairwise fixation index (F_{ST}) estimates among macro-areas and among Italian administrative regions using the *smartpca* software implemented in the EINGESOFT package (Patterson et al., 2006), which computes the Hudson's F_{ST} estimator.

To investigate demographic events from rare variations in Italian macro-areas, we computed the allele frequency spectrum in each Italian macro-area and administrative region and tested for differences. To avoid bias caused by the different sizes of our regional subpopulations, we performed random subsampling without replacement of the

individuals, producing 1000 subsamples of 10 unrelated individuals for each subpopulation with at least 10 samples. Allele counts were computed separately in each subsample for the variants that were observed in that subsample, thus producing counts ranging from 1 to 10 (as we counted the minor allele for subsamples of 10 individuals). This process yielded 1000 estimates of the allele frequency spectrum (with 10 frequency bins) for each subpopulation. Each frequency bin was analyzed separately using the 1000 subsamples to estimate the distribution of values for each subpopulation. This method has the desirable property of producing estimates whose median is independent of the size of the subpopulation. However, the smaller subpopulations showed a reduced variance as the subsamples have high overlap and are thus not independent enough. Thus, the allele frequency spectrum was computed independently for each subsample of a subpopulation. We then compared the distribution of these uniformly sized subsamples for each allele count using a Wilcoxon rank-sum test and Bonferroni's correction in the R programming language environment (R Core Team, 2017). Note that this is not a bootstrap: We are subsampling individuals and not variants (as one would normally do to estimate a distribution of variants), and we are subsampling without replacement because sampling individuals more than once would entail having related individuals in the samples, which would, in turn, produce completely skewed allele counts. In particular, sampling individuals with replacement would cause very rare variants occurring only in one individual to be counted more than once. This would happen more when subsampling from smaller subpopulations producing a very strong bias, where rare variants were shifted toward higher frequency bins, completely skewing the allele frequency spectrum distribution and making comparisons impossible.

Finally, we retained individuals whose four grandparents came from the same macro-area, we filtered out variants with a minor allele frequency lower than 5% and we computed the long-term effective population size for each of the four Italian macro-areas using the NeON R-package (Mezzavilla & Ghirotto, 2015).

2.6 | Genetic comparison within Italy and between Italy and Europe

Differences in allele frequency between macro-areas were tested with Fisher's exact test. We only tested Northern versus Southern Italy (622 and 305 individuals, respectively): We excluded both Sardinia and Central Italy because of their reduced sample size of 20 and 76 individuals, respectively and, for Central Italy, also because of its intermediate position in the North-South cline. We only tested variants with an allele frequency >1% in the whole data set as with our sample sizes we did not have the power to test lower frequency variants. We considered significant all variants passing the 0.01 significance threshold after Bonferroni's multiple test correction. We then computed single-locus F_{ST} estimates to confirm the genetic signals retrieved with Fisher's exact test.

Then, we tested for allele frequency differences between the Italian data set and non-Finnish Europeans from gnomAD, using the χ^2 contingency test instead of Fisher's exact test, due to the much higher sample sizes in this comparison. For the same reason, we tested

all variants and we used a stricter p value threshold of 0.01 (after Bonferroni correction).

In both frequency comparisons (Northern vs. Southern Italy and Italy vs. non-Finnish Europeans), we further investigated the most significant variants. We searched for them in the Genome-Wide Association Studies (GWAS) catalog (downloaded in March 2019; Denny et al., 2013; MacArthur et al., 2017), keeping only associations with a $p < 5E-08$. We also performed a gene enrichment analysis on the genes harboring these variants with the online tool Enrichr (Chen et al., 2013; Kuleshov et al., 2016), selecting the significant enrichments in the following databases: "dbGaP" (Mailman et al., 2007; Tryka et al., 2013), "GWAS catalog 2019" (Buniello et al., 2019), "Jensen disease," "Kyoto Encyclopedia of Genes and Genomes" (KEGG; Kanehisa & Goto, 2000), and "GO Biological Process."

2.7 | Burden test on genes related to hypertension

We gathered an extensive list of 3085 genes and genomic regions that were reported to be associated with the phenotype "Essential hypertension" in the literature from studies based on a genome-wide array, exome array, and sequencing.

Hypertensive status was available for 1670 individuals, of which 150 were affected. We also retrieved systolic and diastolic blood pressure values (SBP and DBP, respectively). We computed the burden of PTV and DMG variants (allele frequency <5%) on genes with at least nine variants (>0.5% of incidence).

According to the δ values (percentage of differences between hypertension and nonhypertension outcome), a gene-based matrix was selected for association analyses. For the genetic risk score (GRS) calculation, three subdatasets were considered:

- (i) $\delta \geq 1$ and $\delta \leq -1$, gene matrix = 101;
- (ii) $\delta \geq 1$ (at risk score), gene matrix = 70;
- (iii) $\delta \leq -1$ (protective score), gene matrix = 31.

The normality assumption of the data was evaluated with the Shapiro-Wilk test; homoscedasticity and autocorrelation of the variables were assessed using the Breusch-Pagan and Durbin-Watson tests.

Logistic regression for hypertension and linear regression for SBP and DBP were performed to test for genetic association. The models were adjusted for potential confounders, including age, gender, body mass index, smoking status, alcohol consumption, and population substructure via the top two PCs. Results were reported as estimates (differences of means or variation at one unit increase of GRS, considering dichotomic and continuous distribution, respectively) at 95% confidence interval. The level of significance was set at $p < 0.05$. Statistical analyses were conducted using R (version 3.0.3; R Core Team, 2017).

We searched the variants within the 101 genes on the GeneAtlas database (Canela-Xandri et al., 2018) using the keyword "I10 Essential (primary) hypertension." We considered significant those variants with a p value after multiple test corrections

of $9.06E-12$. The results of the association in 84,640 cases and 367,624 controls in UK Biobank for 103 SNPs were retrieved.

3 | RESULTS

We analyzed our WES data set of healthy Italian subjects with the dual goal of exploring the genetic patterns of the Italian population and their importance in clinical genetics. After quality control procedures (Section 2), the data set comprised 1686 unrelated individuals, with 669,718 observed variants. As expected for a sequencing data set with this sample size, most variants are low frequency: 92.2% have an allele frequency <5% in our data set (88.4% less than 1% and 76.4% less than 0.1%) and 60.5% are singletons, that is, only observed in a single sample. Functionally, 35.4% of the variants are missense, followed by intronic (29.5%) and synonymous variants (21.7%), while PTVs (“frameshift indel,” “stop gained,” and “essential splice variant”) accounted for 1.95% of the total variants (Table S2). Given our sample size, we expect to observe at least once 96% of the variants with an allele frequency >0.1% in the general population, and virtually all variants with a frequency >0.2% (Supporting Information Materials and Figure S1). Note that throughout the text, variants are referred to as “rare,” “low-frequency,” and “common” without implying any specific frequency threshold. Explicit thresholds are given as required as in the previous paragraph.

Following the guidelines set in previous studies (Fiorito et al., 2016; Sazzini et al., 2016), we split the 20 Italian administrative regions into four macro-areas (Figure 1a and Table S1): Northern Italy, Central Italy, Southern Italy, and Sardinia. We assigned individuals to a macro-area or to a region only when it was the shared birthplace of all four grandparents. Additionally, we removed those who did not cluster well with their macro-area in PCA (Section 2). The sample sizes for macro-areas and regions are in Table S1.

3.1 | Exploring the genetic structure with exome variants

The genetic structure of Italy is clearly discerned from the PCA (Figures 1b and S2A). The main visible features are the North-South cline and the Sardinian isolate, confirming the distinctive genetic profile (mirroring the geographic shape of the country) shown by high-density arrays in previous studies (Di Gaetano et al., 2012; Fiorito et al., 2016; Parolo et al., 2015; Raveane et al., 2019; Sazzini et al., 2016).

While we could not discern any inner structure within macro-areas from the PCA (Figure S2), by plotting the values of the first principal component for each region, we observed some stratification even at the regional level (Figure 1c). This was confirmed by pairwise Wilcoxon testing (Table S3).

To provide an additional measure of population differentiation, we estimated the pairwise fixation index (F_{ST} ; Section 2) between regions (Figure S3). Again, we observed a strong differentiation between macro-areas. However, a finer dissection of the Italian population was not

possible due to the low genetic distance between regions in the same macro-area and the relatively high standard errors of the estimates, especially for regions with small sample sizes (Figure S3 and Tables S1, S4, S5).

We also took advantage of the availability of rare variation in our data set to look for differences in the allele frequency spectrum at the regional level, that is, the distribution of the allele frequency of the variants in the subpopulations (Section 2). Significant differences were detected between all regions (except Lombardy and Emilia Romagna) in the low-frequency end of the spectrum, with Southern Italy having the most low-frequency variants (regions are plotted in Figure 1d, macro-areas in Figure S4 and nominal p values are reported in Table S6).

Finally, we inferred the effective population size changes of the four macro-areas from 5000 to 30,000 years ago (Figure S5).

3.2 | Allele frequency differences between Northern and Southern Italy

After evaluating the genetic structure of Italy as a whole, we delved into details examining which variants and genes present the highest degree of differentiation between Northern and Southern Italy (622 and 305 individuals, respectively), focusing on the possible phenotypic implications of the structure we observed. Other macro-areas were not examined because of their small sample size.

Allele frequencies in the North and South are highly correlated (Pearson $r = 0.998$; $p < 2.2E-16$; Figure 1e), with a maximum difference of 17%. We tested for significant differences between all variants with a frequency of at least 1% in our data set and reported in Table 1 the six variants with a $p < 0.01$ after Bonferroni correction for multiple testing. They are also the same six variants with the highest single-locus F_{ST} estimate. The full list is available in Supporting Information File S1. From the test results, we observed a genomic inflation factor of 1.77 (Figure S6) as expected when comparing groups with population stratification.

The strongest signal is rs1129038, located in the 3'-UTR of the *HERC2* gene. The derived allele T is enriched in Northern Italy, and it is associated with eye and hair color and skin pigmentation (Morgan et al., 2018). Its homozygous occurrence is highly predictive of blue eye color (Eiberg et al., 2008), but it is also associated with pigmentation-related diseases like melanoma and vitiligo (Jin et al., 2012).

The second strongest signal comprises three very close (<400 bp apart) and previously unreported missense variants in the *OR52R1* gene, an olfactory receptor reported as a segregating pseudogene by RefSeq.

The third signal is rs1229984, a missense variant in the *ADH1B* gene encoding the beta subunit of class I alcohol dehydrogenase (ADH), which is involved in ethanol metabolism. The derived allele T (which, for this variant, is the reference allele) protects against alcoholism by metabolizing alcohol to acetaldehyde more efficiently than the ancestral allele C leading to elevated acetaldehyde levels that make drinking unpleasant (Polimanti & Gelernter, 2017). On the contrary, the C allele is associated with alcohol-related diseases, alcohol dependence

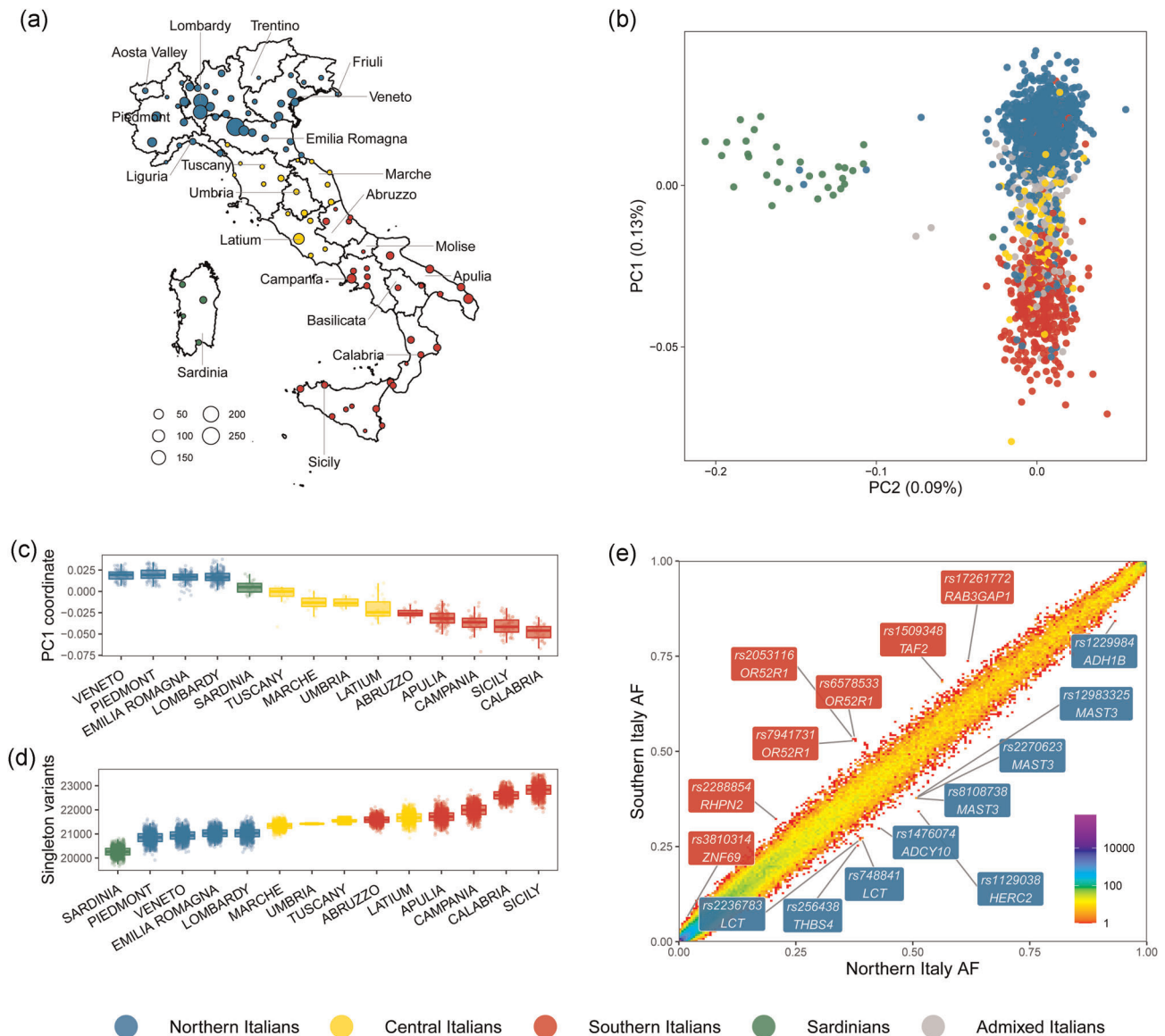


FIGURE 1 Genetic structure of the Italian population and allele frequency (AF) differences between Northern and Southern Italy. (a) The administrative province of origin of the individuals in our data set. The size of the circles shows the number of individuals. (b) Scatterplot of the first two components of the principal component analysis (PCA). (c) Strip and boxplot of the first principal component of each individual grouped by Italian regions. (d) Strip and boxplot of the number of low-frequency variants in 1000 resamplings of 10 individuals from each region. Only regions with at least 10 individuals have been included in (c) and (d). (e) Density plot of all variants by their alternative allele frequency in Northern and Southern Italy. Labels (variant ID and gene) are shown for the 16 variants with a $p < 0.05$ after multiple test correction. Label color denotes the higher frequency of the allele in the Northern (blue, 622 individuals) or Southern (red, 305 individuals) Italian population

(Sanchez-Roige et al., 2019), cancers (Lesseur et al., 2016), and, not surprisingly, to the trait “regular attendance at a pub or social club” (Day et al., 2018).

The fourth signal, the intronic SNP rs256438 in the *THBS4* gene, is closer to the background (Figure 1e). The variant was inconclusively associated with serum thyroid-stimulating hormone levels (Malinowski et al., 2014) and significantly associated with abnormalities of ocular refraction (Hysi et al., 2020).

We also examined the remaining top 1% of the most significant variants to look for associations in the GWAS catalog and performed a gene enrichment analysis. Both analyses yielded signals related to pigmentation, immune, and cardiovascular traits

and diseases. Of particular interest are rs16891982 (*SLC45A2* gene), rs16903574 (*OTULINL* gene), and rs3131379 (*MSH5* gene), which were associated with pigmentation, allergic diseases (asthma, hay fever, eczema), and systemic lupus erythematosus, respectively (Supporting Information Materials and File S2).

3.3 | Allele frequency differences between Italy and Europe

The allele frequency differences we observed along the Italian peninsula are part of the larger North-South European clines

TABLE 1 Variants with significant allele frequency differences between Northern and Southern Italy (Bonferroni's multiple test correction with $p < 0.01$)

Variant	dbSNP ID	HGVSc	Gene symbol	Effect	Northern Italy AF	Southern Italy AF	North-western European AF	Southern European AF	African AF	East Asian AF	Nominal p value
15:28356859:C>T	rs1129038	NM_004667.5:c.*50G>A	HERC2	3'-UTR	0.51	0.34	0.77	0.43	0.11	0.0008	5.76E-12
11:4824878:A>C	rs2053116	NM_001005177.3:c.733T>G	OR52R1	Missense	0.37	0.53	0.34	0.44	0.51	0.02	7.14E-11
11:4825010:T>A	rs6578533	NM_001005177.3:c.601A>T	OR52R1	Missense	0.37	0.53	0.33	0.44	0.51	0.017	1.04E-10
11:4825225:A>G	rs7941731	NM_001005177.3:c.386T>C	OR52R1	Missense	0.37	0.53	0.34	0.44	0.51	0.018	5.28E-10
4:100239319:T>C	rs1229984	NM_000668.5:c.143A>G	ADH1B	Missense	0.93	0.84	0.97	0.90	0.98	0.26	1.28E-08
5:79366249:T>G	rs256438	NM_003248.5:c.1452+16 T>G	THBS4	Intronic	0.38	0.25	0.36	0.35	0.22	0.21	5.19E-08

Note: Column "Variant" reports the variant chromosome, position, reference, and alternative alleles. Column "dbSNP ID" reports the SNP ID according to the dbSNP database. Column "HGVSc" is the standard HGVSc coding DNA variant nomenclature, with the annotated transcript as the prefix. Columns "North-western European," "Southern European," "African," and "East Asian" report the gnomAD allele frequency in Europe, Africa, and East Asia. "North-western European" and "Southern European" are subsets of "Non-Finnish European." Abbreviations: AF, allele frequency; SNP, single-nucleotide polymorphism.

reported in gnomAD. In the case of *HERC2* and *OR52R1*, these clines extend to Africa. In contrast, for *ADH1B*, the derived allele is enriched only in Southern Europe, while the ancestral allele is almost fixed in both Northern Europe and Africa. Only East Asia sports a (much) higher frequency of the derived allele (Table 1).

We tested all the variants for significant allele frequency differences between the whole Italian population from our data set and the gnomAD non-Finnish European population. We found 19,561 variants passing the 0.01 significance threshold after Bonferroni's multiple test correction (Supporting Information File S1). Of these variants, 35 have been reported in ClinVar as linked to disease or drug response: 21 with higher and 14 with lower frequency in Italy than in Europe. We reported the most interesting ones in Table 2, while the full list is in Table S7.

Out of 35 variants, 8 have been submitted as pathogenic or likely pathogenic. Among those, we found several variants, with an allele frequency <0.1%, that are at least three times more frequent in Italy: two of them have been evaluated as "Pathogenic" for phenylketonuria (rs76212747 in the *PAH* gene; Guldberg et al., 1998) and thalassemia (rs11549407 in the *HBB* gene; Richards et al., 2015) and showed a frequency five and seven times higher in Italy than in the non-Finnish European gnomAD data set, respectively. Three of them (rs200635937 in the gene *GPR161*, rs770171865 in *PSAP*, and rs769409705 in *SLC34A1*) have been submitted as "Likely Pathogenic" for the medical conditions pituitary stalk interruption syndrome, sphingolipid activator protein 1 deficiency, and infantile hypercalcemia, respectively. In contrast, we found one variant in the gene *FLG* (coding for profilaggrin), submitted as pathogenic for atopic dermatitis (Richards et al., 2015) and ichthyosis vulgaris (Smith et al., 2006), whose frequency is seven times lower in Italy (0.24% vs. 1.6%).

Another 13 variants were submitted as "risk factors" or "protective." Among those with lower frequency in Italy, we found three variants that are risk factors for myocardial infarction and one protective variant for alcohol dependence (in *ADH1C*). Another statistically significant signal with higher frequency in Italy (42% vs. 26%) was a missense variant in *TLR1*, which is a risk factor for leprosy.

The remaining 14 variants were submitted as "drug response." Again, some have a higher frequency in Italy, like rs57913007 (5.1% vs. 2.9%), which is linked to response to tramadol (analgesic). Others have a lower frequency in Italy, like rs11676382 (5.5% vs. 9.3%), which is linked to the efficacy of warfarin (anticoagulant). Still, other variants were linked to cisplatin (chemotherapeutic) toxicity, phenylthiocarbamide tasting, and nicotine toxicity.

Finally, we explored the 19,561 variants, whose allele frequencies were significantly different between Italian and non-Finnish European individuals, both by the direct association in the GWAS catalog and by gene enrichments, thus obtaining a vast number of phenotypes and diseases (Supporting Information Materials and File S3). We can group most of them in five broad categories: pigmentation (e.g., hair and skin color, tan response, and skin cancer), cardiovascular (both as susceptibility to diseases and related phenotypes), immune diseases (e.g., rheumatoid arthritis, type 1 diabetes mellitus), cancer, and neurological disorders (such as Alzheimer's disease).

TABLE 2 A selection of the significantly different variants between our Italian data set and European non-Finnish populations from gnomAD reported to be “pathogenic” or “likely pathogenic,” “risk factor,” “protective,” or “drug response” in the ClinVar database (the full list is in Table S7)

Variant	dbSNP ID	HGVSc	Gene symbol	Effect	ClinVar disease	ClinVar significance	Northern Italy AF	Southern Italy AF	Italy AF	Non-Finnish European AF	Nominal p value	Frequency Ratio
1:15228586-1:G>A	rs61816761	NM_002016.1:-c.1501C>T	FLG	Stop gained	Atopic dermatitis, ichthyosis vulgaris	Pathogenic/likely pathogenic	0.0033	0.0017	0.0024	0.01666	2.60E-10	0.14
2:85777633-:C>G	rs11676382	NM_000821.6:-c.2084+45-G>C	GGCX	Intronic	Warfarin response (dosage)	Drug response	0.0596	0.0492	0.0555	0.09312	1.30E-13	0.60
4:10026078-9:T>C	rs698	NM_000669.4:-c.1048A>G	ADH1C	Missense	Alcohol dependence	Protective	0.3052	0.2328	0.2849	0.40205	2.18E-42	0.71
3:14187449-:G>T	rs2228001	NM_004628.4:-c.2815C>A	XPC	Missense	Cisplatin response (toxicity/ADR)	Drug response	0.5450	0.4902	0.5229	0.59232	8.59E-16	0.88
7:14167334-5:C>G	rs713598	NM_176817.4:-c.145G>C	TAS2R38	Missense	Phenylthiocarbamide tasting	Drug response	0.4751	0.5213	0.4697	0.40570	1.19E-13	1.16
15:7888292-5:G>A	rs16969968	NM_000745.3:-c.1192G>A	CHRNA5	Missense	Nicotine response (toxicity/ADR)	Drug response	0.3955	0.4279	0.4144	0.34798	2.05E-15	1.19
17:4836381-:C>T	rs6065	NM_000173.6:-c.482C>T	GP1BA	Missense	Aspirin response (efficacy)	Drug response	0.1083	0.1056	0.1126	0.08297	1.34E-09	1.36
4:38799710-:T>C	rs4833095	NM_003263.3:-c.743A>G	TLR1	Missense	Leprosy 5	Risk factor	0.3969	0.4836	0.4254	0.26536	4.50E-94	1.60
4:69973921-:C>T	rs57913007	NM_001074.2:-c.1191C>T	UGT2B7	Synonymous	Tramadol response	Drug response	0.0448	0.0505	0.0509	0.02886	1.93E-13	1.76
12:1032467-01:A>G	rs76212747	NM_000277.1:-c.734T>C	PAH	Structural inter-action	Phenylketonuria, hyperphenylalaninaemia	Pathogenic	0.0072	0.0016	0.0053	0.00095	1.55E-13	5.62
11:5248004-:G>A	rs11549407	NM_000518.4:-c.118C>T	HBB	Stop gained	Heinz body anemia, HbSS disease, alpha thalassemia, susceptibility to malaria, familial erythrocytosis-6, beta thalassemia,	Pathogenic	0.0024	0.0066	0.0047	0.00067	1.50E-15	7.06

(Continues)

TABLE 2 (Continued)

Variant	dbSNP ID	HGVSc	Gene symbol	Effect	ClinVar disease	ClinVar significance	Northern Italy AF	Southern Italy AF	Italy AF	Non-Finnish European AF	Nominal <i>p</i> value	Frequency Ratio
1:16807409-3:A>T	rs200635937	N-	GPR161	Missense	methemoglobinemia, fetal hemoglobin quantitative trait locus 1	Likely pathogenic	0.0016	0.0000	0.0027	0.00038	6.04E-09	7.09
		M_001267-609.1:c.56-T>A			Pituitary stalk interruption syndrome							
10:7358784-6:G>T	rs770171865	N-	PSAP	Missense	Sphingolipid activator protein 1 deficiency	Likely pathogenic	0.0008	0.0016	0.0012	0.00006	1.33E-08	18.93
		M_001042-465.2:c.645C>A										
5:17681349-3:G>T	rs769409705	NM_003052.4:c.458G>T	SLC34A1	Missense	SLC34A1-related disorders, infantile hypercalcemia	Likely pathogenic	0.0024	0.0016	0.0012	0.00004	1.60E-10	26.49

Note: The nominal *p* value (column "Nominal *p* value") and the frequency ratio (column "Frequency ratio") between Italy AF (alternative allele frequency in our Italian data set) and non-Finnish European AF (alternative allele frequency in non-Finnish European individuals from gnomAD) are reported. Column "Variant" reports the variant chromosome, position, reference, and alternative alleles. Column "dbSNP ID" reports the SNP ID according to the dbSNP database. Column "HGVSc" is the standard HGVS coding DNA variant nomenclature, with the annotated transcript as a prefix. Abbreviations: ADR, adverse drug reaction; AF, allele frequency; SNP, single-nucleotide polymorphism.

3.4 | Importance of the reference population for assessing pathogenicity

When assessing the pathogenicity of variants, it is often appropriate to assume that variants that frequently occur in healthy individuals are not pathogenic, at least not with high penetrance. This is normally done by checking allele and genotype frequencies in public databases (Eilbeck et al., 2017). We explored the risk of mis-assessing the pathogenicity of a variant by using frequencies estimated from an insufficient number of individuals or a population that is not a perfect match for the affected individual.

We compared the allele frequencies estimated from our Italian sample with those estimated from commonly used reference data sets: the Tuscans and non-Finnish Europeans in the 1000 Genomes Project (phase 3, “KGP_TSI” and “KGP_NFE,” respectively) and the non-Finnish European in gnomAD (“GND_NFE”). Frequency thresholds used in pathogenicity assessment are disease- and phenotype-specific. We chose a few commonly used thresholds and counted how many variants in our data set fall on different sides of each threshold for each of the reference data sets in Table S8. For instance, if we decide to assess variants with an allele frequency >1% as nonpathogenic, there are 3782 false pathogenic candidate variants whose frequency is greater than the threshold in our sample but below it in gnomAD. While most of them are very close to the threshold in both data sets, a few are quite different: 590 are less than 0.5% and 33 less than 0.1% in gnomAD, while still being above 1% in our sample. Frequency discrepancies as large as those in variants satisfying other pathogenicity criteria could cause an incorrect assessment of pathogenicity even under careful scrutiny. For instance, we reported all variants that were reclassified in ClinVar from nonbenign (i.e., “Uncertain significance,” “Conflicting interpretations of pathogenicity,” “Likely pathogenic,” or “Pathogenic” in the August 5, 2018 release) to benign (i.e., “Benign” in the June 15, 2020 release), and whose allele frequency is at least 1.5 times higher in our data set than in the non-Finnish Europeans from gnomAD (Table S9). The higher frequency in our data set could have been useful in detecting these misclassifications, especially in genes responsible for dominant diseases. For instance, rs34136999, rs61734190, and rs121912749 were from 5 to 20 times more frequent in our Italian data set than in gnomAD, and they were reclassified as benign after being submitted as uncertain significance, conflicting interpretations, and pathogenic for serious dominant disease, such as Lynch syndrome, Angelman syndrome, and spherocytosis, respectively.

3.5 | Putatively pathogenic variants

We examined how other common pathogenicity assessment criteria behave in our healthy Italian cohort. We employed three different and complementary methods for assessing the pathogenicity of a variant: PTV, missense variants predicted to be damaging (DMG, see Section 2), and variants submitted as pathogenic or likely pathogenic in the ClinVar database (CLNPAT). After excluding variants with an allele frequency >5% in our data set, we obtained 12,852 PTV,

23,682 DMG, and 1308 CLNPAT variants in the whole data set. We refer to these variants as putatively pathogenic (PP) variants.

We related the functional effect and pathogenicity of variants to their rarity measured as the ratio of singleton variants to the total number of variants in each category (Figure 2a). We also computed the pathogenic burden in our healthy Italian individuals showing that the average number of PP variants in an individual is 16.6 for PTV, 23.2 for DMG and 1.8 for CLNPAT variants (Figure 2b–d).

As expected, drug and xenobiotics metabolism, as well as olfactory transduction pathways from KEGG, accumulate the highest number of functionally disrupting variants (Supporting Information Materials and Figure S7), while several metabolic and the “ABC transporters” pathways were the most prone to accumulate DMG variations (Figure S8).

Finally, we verified that our sets of PTV and DMG PP variants could be a valid “proxy” for pathogenicity by relating the genetic burden of such variants in selected genes (File S5) with three blood pressure phenotypes, which were available in our data set (hypertension, SBP, and DBP). The resulting GRs were significantly associated ($p < 0.05$) with the target phenotypes in all models, thus indirectly validating our assessment of variant pathogenicity (Section 2, Supporting Information Materials, Tables S10 and S11).

3.6 | PP variants in ACMG SF genes

To explore PP variants in a more clinically relevant context, we focused on the 59 medically actionable genes recommended by the ACMG for reporting of incidental findings (ACMG SF v2.0; Kalia et al., 2017).

We found that PTVs are only half (0.52) as frequent in the ACMG genes than in the whole genome (Table S12). In contrast, DMG and CLNPAT variants are 2.20 and 5.97 times more frequent in the ACMG genes, respectively, while missense variants are only slightly less abundant (0.95).

Most of the PTV (68%) and DMG (75%) variants in the ACMG genes had already been submitted to ClinVar. Unsurprisingly, 22 out of 29 submitted PTVs are classified as pathogenic or likely pathogenic, while most of the submitted DMG variants are of uncertain significance or with conflicting interpretations (the rest are evenly split between the benign and pathogenic classes; Table S13).

We computed the prevalence of PP variants in the ACMG genes in our healthy Italian population (Table 3). Here, PTV and DMG variants were further restricted, removing those that were reported as benign or likely benign in ClinVar at least once. We see 2.5% of the population carrying PTVs in the ACMG medically actionable genes (Table 3).

3.7 | PP variants in Mendelian diseases

We also looked for variants that, when occurring in an individual affected by a matching phenotype, would be likely diagnosed as causative. We considered two subsets of our PP variants with a high likelihood of being causative of Mendelian disorders (Supporting Information File S4). The first subset consists of heterozygous PTVs with an allele frequency

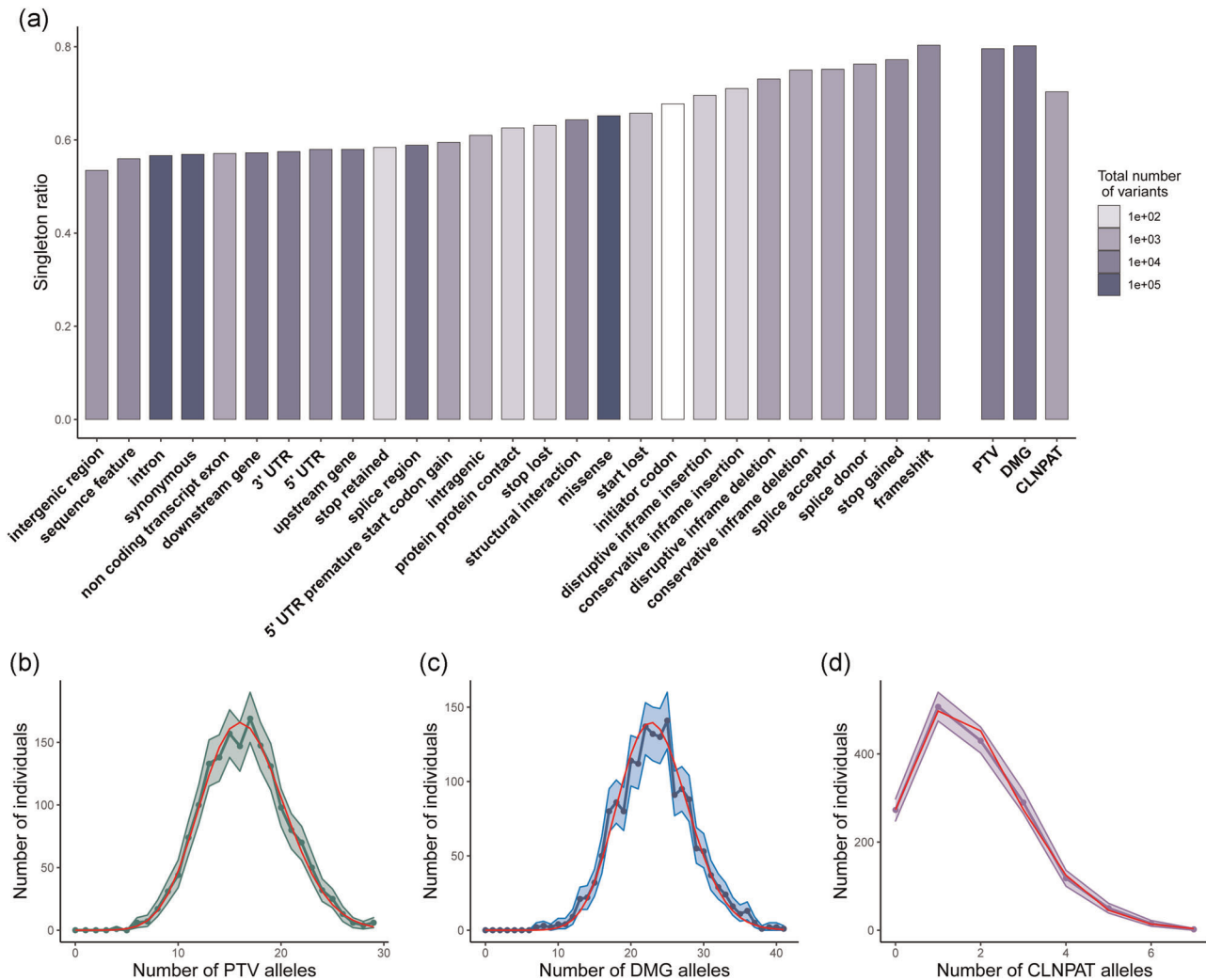


FIGURE 2 Evaluation of pathogenic variation. (a) Ratio of singleton to the total number of variants for each effect and pathogenic category. (b–d) Distribution of the burden of each pathogenic category per individual counted as the number of variants observed. The shaded area shows the 5%–95% confidence interval estimated by bootstrapping. The red line shows the expected Poisson distribution as a reference

<0.1% in loss-of-function intolerant genes for a total of 80 candidates “dominant” variants. The second subset consists of homozygous PTVs with an allele frequency of <1% in genes reported as “recessive” in OMIM (McKusick, 1998), for a total of 36 candidates “recessive” variants. Loss-of-function intolerant genes were selected as those with pLI equal to 1. pLI is a score introduced by Lek et al. (2016), which is frequently used for prioritizing candidate genes in practical diagnostics.

We manually reviewed these variants selecting those where the Mendelian disease mechanism of the relevant gene was compatible with a loss-of-function variant and assessing their ACMG pathogenicity class with Varsome (Kopanos et al., 2019; Table 4).

We found four heterozygous PTVs in haploinsufficient genes known to cause Mendelian diseases with a dominant model. They are very rare, appearing in a single individual in both our data set and the non-Finnish European from gnomAD (our data set is part of gnomAD). Three of these were pathogenic (Class 5) and one likely pathogenic (Class 4), affecting genes associated with epilepsy (*STXBP1*), mental retardation (*KAT6A*), and basal cell nevus syndrome (*PTCH1*).

We also found three homozygous PTVs in genes known to cause Mendelian conditions by a recessive or semidominant model. These variants have a higher frequency both in our data set and in gnomAD. However, they are well-known variants and are reported in ClinVar as causing milder conditions that may have gone undiagnosed: dihydropyrimidine dehydrogenase deficiency (increased toxicity of several chemotherapy drugs), ichthyosis vulgaris (common dominant skin disorder where homozygosity may cause a more extreme phenotype), and McArdle disease (a metabolic syndrome that can be diagnosed as late as the third or fourth decade).

All four frameshift variants are located in the last or second-to-last exons and prolong the open reading frame, thus making their effect on the protein function harder to interpret. There is evidence of frameshift variants increasing the final transcript length, allowing them to escape the nonsense-mediated messenger RNA decay pathway leading to the loss of function through different mechanisms (Carvalho et al., 2009; Kausar et al., 2019; Patronas et al., 2012). However, we cannot rule out the possibility of these variants being

TABLE 3 Prevalence of PP variants in ACMG genes in the Italian population

PP category	Individuals	Ratio (%)
PTV	82	4.9
PTV without benign	42	2.5
DMG	437	25.9
DMG without benign	134	7.9
CLNPAT	74	4.4

Note: Column “Individuals” shows the number of individuals carrying at least one alternative allele of at least one variant in the corresponding PP category. Column “Ratio” is the ratio of individuals carrying the variant in our Italian sample.

Abbreviations: ACMG, American College of Medical Genetics; PP, putatively pathogenic; PTV, protein-truncating variant.

gain-of-function, dominant-negative, or even without any pathogenic consequences.

4 | DISCUSSION

The genetic structure of the Italian population has already been investigated usually using SNP array data mainly comprised of genome-wide common genetic markers (Di Gaetano et al., 2012; Fiorito et al., 2016; Parolo et al., 2015; Raveane et al., 2019; Sazzini et al., 2016). Other well-known studies explored human genetic variation worldwide from whole-exome data (ExAC/gnomAD). This is the first study to explore the Italian genetic structure from WES data of a sizable number of individuals.

We corroborated the previously observed genetic structure of Italy with the north-south cline of the mainland and Sicily as well as the outlying Sardinian isolate (Di Gaetano et al., 2014). This shows that WES data also provides enough information to highlight the Italian macro-areas both in PCA, F_{ST} , and in the allele frequency spectrum. Northern Italy appears more genetically homogeneous than Central and Southern Italy, while the higher quantity of rare variants in the Southern Italian regions suggests a higher effective population size (Lao et al., 2008; Marth et al., 2004). This is an independent confirmation of previous results obtained by LD-based methods (Fiorito et al., 2016).

Many of the differences in variant frequencies between Northern and Southern Italy occur in genes involved in greater European latitudinal clines like skin/hair pigmentation (*HERC2*, *SLC45A2*) and lactose tolerance (*LCT*), where the associated phenotypes are also known to follow the same cline (Donnelly et al., 2012). Pigmentation and skin diseases caused by UV light exposure are also a recurrent theme in GWAS direct associations and gene enrichments both in the Northern–Southern Italy comparison and in the Italy–Europe comparison. The rs1229984 variant in the *ADH1B* gene follows a similar yet different distribution: The derived allele is almost absent in Northern Europe but more frequent in Spain and Italy, especially in Southern Italy. However, it is also rare in Africa

and worldwide except in Eastern Asia, where it is the major allele and shows signs of recent positive selection (Polimanti & Gelernter, 2017). Notably, we also found an independent signal at rs698 in *ADH1C* in the Italy–Europe frequency comparison, which relates to alcohol metabolism. Other differences we observed are harder to link to a phenotype, like the differences in olfactory receptor genes that are generally considered to be under low selective pressure for their tolerance of loss-of-function variants (Karczewski et al., 2020). This makes the strong signal we found in the *OR52R1* gene that has no known phenotypic association hard to interpret. *HERC2*, *OR52R1*, and *ADH1C* were previously reported as genes containing non-specified, differentially frequent variants between Italian macro-areas (Sazzini et al., 2016).

Differences in frequencies among populations and public databases also have an impact on the assessment of variant pathogenicity in disease-affected individuals. An often-employed criterion is excluding variants that have a high allele frequency in the healthy population. This is done based on the assumption that purifying selection has curbed the frequency of high penetrance pathogenic variants (Richards et al., 2015). However, public frequency databases may not be fully representative of the real population of the affected individual. In that case, a variant that is common in the real population might be rare in the reference population and thus erroneously be considered as pathogenic.

What does this entail for Italy when publicly available European reference databases are employed? We observed that a large sample size seems to be the most important factor to increase accuracy: The 107 Tuscans from the 1000 Genomes Project, although the most closely related to our sample, are too few to produce accurate frequencies for low-frequency variants. The non-Finnish European populations from the 1000 Genomes (~400 individuals) and gnomAD (~56,000) produce more accurate frequency estimates for our Italian sample. However, even with gnomAD, which provides the best results, we found evidence of a possible misclassification of variants. Note that our data set is part of gnomAD and while this is undoubtedly a bias, its effect should not be very significant as our Italian sample amounts to just 3% of the entire non-Finnish European gnomAD sample. This is also the reason why we did not compare our sample to the Southern Europe gnomAD subpopulation, where the overlap is much greater.

In conclusion, gnomAD provides very good frequency estimates for Italian individuals. However, the researcher/clinician should be aware of the small likelihood that variants could be misclassified by relying exclusively on it. When possible, we recommend the use of a complementary population-specific database of frequencies estimated from at least a thousand individuals. Conceivably, frequency databases of this kind could be used to detect previously misclassified variants in pathogenicity databases like ClinVar, as we showed in an example. For these purposes, we made the aggregated frequencies from our Italian sample publicly available from the website of the Italian partnership called Network for Italian Genomes (<http://nigdb.cineca.it>, NIG-ExIT).

We then investigated the potential functional and pathogenic role of variants in our data set, producing three classes of PP variants: PTV, DMG, and CLNPAT. Care must be used when speaking about pathogenicity in a cohort of reportedly healthy individuals without manifest

TABLE 4 Heterozygous PTVs in “dominant” and haploinsufficient genes and homozygous PTVs in “recessive” genes

Variant	dbSNP ID	GT	Gene symbol	Effect	Exon or intron	HGVS	Consequence	OMIM disease/OMIM accession number	Heredity class	ACMG class	LOFTEE (%)	Italy AF (%)	Non-Finnish European AF (%)
8-41789794-A-A	rs747125100	HET	KAT6A	Frameshift	17/17	NP_006757.2:p.Ser1982fs	Extended protein (+10 aa)	Mental retardation, autosomal dominant 32/616268	AD, LoF	P	LC	0.03	0.001
9-98209349-GT-G	rs769175073	HET	PTCH1	Frameshift	23/24	NP_001077071.1:p.Leu1331fs	Extended protein (+3 aa)	Basal cell nevus syndrome/109400	AD, LoF	LP	HC	0.03	0.001
9-130446718-A-G	rs776851083	HET	STXBP1	Frameshift	19/20	NP_003456.1:p.Val593fs	Extended protein (+21 aa)	Epileptic encephalopathy, early infantile, 4/612164	AD, LoF	P	HC	0.03	0.001
9-130446725-CT-C	rs759537033	HET	STXBP1	Frameshift	19/20	NP_003456.1:p.Phe595fs	Extended protein (+21 aa)	Epileptic encephalopathy, early infantile, 4/612164	AD, LoF	P	HC	0.03	0.001
1-97915614-C-T	rs3918290	HOM	DPYD	Splice donor	14/22	NM_000110.3:c.1905+1G>A		Dihydropyrimidine dehydrogenase deficiency/612779	AR	P	HC	0.33	0.55
1-1522858661-G-A	rs61816761	HOM	FLG	Stop gained	3/3	NP_002007.1:p.Arg501Ter	Truncated protein	Ichthyosis vulgaris/146700	SD	P	LC	0.24	1.67
11-64527223-G-A	rs116987552	HOM	PYGM	Stop gained	1/20	NP_005600.1:p.Arg50Ter	Truncated protein	McArdle disease/232600	AR, LoF	P	HC	0.12	0.24

Note: We report the alternative allele frequencies from our data set of Italian individuals (Italy AF) and the non-Finnish European samples from gnomAD (non-Finnish European AF). Column “GT” indicates the genotype of the individuals carrying the variant. Column “LOFTEE” indicates the LOFTEE (VEP plugin) classification of the PTVs in HC (high-confidence) or LC (low-confidence). Column “ACMG class” reports the variant classification according to the ACMG Guidelines (P: Class 5, pathogenic; LP: Class 4, likely pathogenic; VUS: Class 3, variant of uncertain significance; B: Class 1, benign), as reported by VarSome as of November 2020 (VarSome classification engine is frequently updated). The complete list of PTVs in haploinsufficient genes and homozygous PTVs can be found in Supporting Information File S4. Abbreviations: ACMG, American College of Medical Genetics; AD, autosomal dominant; AF, allele frequency; AR, autosomal recessive; GT, genotype; HET, heterozygous genotype; HGVS, sequence variant nomenclature from the Human Genome Variation Society; HOM, homozygous genotype; LoF, loss-of-function; PTV, protein-truncating variant; SD, semidominant.

pathogenic phenotypes. Without a specific disease hypothesis, criteria for assessing variant pathogenicity must necessarily be very broad and generic. For instance, as we wanted to also include risk factors and low-penetrance variants causing mild phenotypes, we purposely chose a rather high allele frequency threshold of 5%. Another limit is that without a disease, we do not have a list of linked genes; thus, we look at variants in the whole exome even though many genes are not linked to any disease. For all these reasons, we call these sets of variants *putatively* pathogenic to stress that they are enriched in variants that are likely to be pathogenic.

Indeed, we saw that categories of variants understood to have greater effect and/or deleteriousness have higher ratios of singletons (Figure 2a), especially those included in the PTV category. Also, the high ratio of PTV and DMG categories is consistent with our interpretation of deleteriousness. On the contrary, CLNPAT has a much lower ratio but this can be explained by the fact that singleton variants from our data set are less likely to have been observed, diagnosed, and submitted to ClinVar than the other more common variants.

We also gave some estimates on the pathogenetic burden of our Italian individuals finding that they carry, on average, 16.6 PTVs, 23.2 DMG, and 1.8 CLNPAT variants. The higher number of DMG with respect to PTVs can be explained by their weaker effect because in many genes, even a very disruptive missense variant is unlikely to be as disruptive as a PTV. The much lower number of CLNPAT variants is likely due to the ClinVar database focus on genes that are relevant to disease, while PTVs and DMG variants are found more evenly in the whole exome. Also relevant is that most of the variants in our data set are low frequency and again less likely to be in ClinVar. Conversely, all variants were tested for being PTV and DMG. Note that these interpretations are not population-specific: For instance, we believe DMG variants to be more common than PTVs in most populations.

Our estimate of individual PTV burden is roughly comparable with estimates provided by other studies. For instance, in Van Hout et al. (2020), the authors examined individuals from the UK Biobank and found 15 and 24 loss-of-function variants (LOF) when considering LOF in all transcripts and any transcript, respectively. However, differences in the selection of PTV versus LOF variants, choice of transcripts, frequency cut-offs, and variant calling make an unbiased comparison impossible.

To perform a sort of “validation” of our work on pathogenic variation by linking it to a phenotype, we took advantage of the availability of hypertensive status and blood pressure values for our cohort. As suggested in Russo et al. (2018), we followed a burden-based approach to link pathogenic variation (PTV and DMG) to hypertension, an intermediate phenotype associated with an increased risk of cardiovascular disorders. Although the number of individuals affected by hypertension in our sample was too low to discover new associations, we showed that our results are consistent with known associations, thus indirectly validating our assessment of variant pathogenicity.

Previous studies (Karczewski et al., 2020; Lek et al., 2016) showed that many genes are quite tolerant of variants causing loss-of-function. As a consequence, PP variants in those genes are unlikely to actually be pathogenic, thus explaining part of the burden of PP variants that

we observed in our and other healthy cohorts. When focusing on the ACMG genes, our three classes of PP variants behave quite differently. The lower proportion of PTVs found in the ACMG genes shows that the ACMG genes are, as a whole, less tolerant to PTVs than the rest of the exome. In contrast, the enrichment of DMG and CLNPAT may seem counter-intuitive. In the case of DMG variants, which are missense variants predicted to be deleterious, this is likely due to the fact that the gene or some related feature (e.g., genetic position, conservation score, etc.) is considered in the prediction producing a positive bias in clinically relevant genes (missense variants themselves are not enriched in the ACMG genes). In the case of CLNPAT, the likely explanation is a greater interest within the clinical community in the ACMG genes and thus a greater representation in ClinVar.

We found that 2.5% of Italian individuals carry PTVs in ACMG genes. Different estimates for the frequency of loss-of-function variants in the ACMG genes have been provided in other studies: 1% in Olfson et al. (2015), 2.6%–4.9% in Shah et al. (2018) and 2% in Van Hout et al. (2020). However, when comparing the individual PTV burden in the whole exome, every estimate was computed differently; thus, discrepancies between these values are likely to be due more to methodological than biological reasons.

When investigating variants that could cause Mendelian diseases, we were much more restrictive, considering only PTVs and manually checking the genes, inheritance model, associated disease, and the pathogenicity class. The variants we found were variants that, in the presence of disease, would be diagnosed as causative in a clinical setting. Unfortunately, we have been unable to validate these variants or to confirm the disease status of the individuals carrying these variants. Thus, there are three possible reasons for these individuals to have been included in a healthy cohort: (i) the variant has been mis-called and is not present; (ii) the variant, even though satisfying the ACMG standard, is not actually pathogenic; or (iii) the phenotype was undetected or unreported in the recruitment. While we cannot exclude reason (i), it seems unlikely that it is the only relevant explanation as these genotypes appear to be of high quality and we expect that validation would confirm them. Reason (iii) is the most likely explanation for the homozygous PTVs in the recessive genes because they cause an adverse response to chemotherapy (*DPYD*), a mild disease (*FLG*), and a muscle disease (*PYGM*) often diagnosed later in life. As these individuals are included in the healthy subset of gnomAD individuals, this shows that one cannot exclude the presence of other phenotypes like these in public databases. On the contrary, reason (iii) is less likely for heterozygous PTVs in haploinsufficient dominant genes, such as the *KAT6A* gene, which is linked to evident morphological features and mental retardation. Thus, for the variants in *KAT6A*, *PTCH1*, and *STXBP1*, we suspect reason (ii) to be the more likely explanation. This means that even with careful scrutiny, the standard diagnostic criteria may still produce some false positives.

In conclusion, we believe large whole-exome or even whole-genome sequencing data sets to be very relevant to many fields in genetics, especially for highly structured populations like the Italian one. They are also instrumental to a more comprehensive approach to clinical genetics that uses population genetics as a lens to better

understand the interplay between polymorphisms, genetic susceptibility, and pathogenic variation. A great amount of whole-exome, clinical exome, and genome sequencing data is routinely produced in Italy both for clinical and research purposes. Collecting it in a comprehensive Italian-specific database, and comparing variant frequencies with other population databases, would be a valuable resource in many research and clinical contexts. With this in mind, we make available our data set in aggregated form on the site <http://nigdb.cineca.it> (under the name NIG-ExIT), as a new reference for genetic frequencies in Italy and its macro-areas.

ACKNOWLEDGMENTS

We thank all the volunteers who participated in this study. We thank CINECA and C3S (<http://c3s.unito.it>) for having provided the computational resources. We thank Dr. Beth O. Van Emburgh and Eric Van Emburgh for revising the manuscript. This study was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca—MIUR project “Dipartimenti di Eccellenza 2018–2022” (no. D15D18000410001) to the Department of Medical Sciences, University of Torino (Dr. Giuseppe Matullo); the Compagnia di San Paolo and the Italian Institute for Genomic Medicine, IIGM (fka Human Genetics Foundation-Torino, HuGeF), Turin, Italy (grant 2017–2018 to Dr. Giuseppe Matullo); and the AIRC—Associazione Italiana per la Ricerca sul Cancro (IG 21390 to Dr. Giuseppe Matullo; SA fellowship).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Giovanni Birolo, Serena Aneli, Cornelia Di Gaetano, Giuseppe Matullo, Rosanna Asselta, and Stefano Duga conceived the study. Giovanni Birolo and Serena Aneli performed the analyses. Giovanni Cugliari performed the statistical analysis for the hypertension phenotype. Alessia Russo and Cornelia Di Gaetano performed the analyses on the hypertension phenotype. Elisa Giorgio manually reviewed the variants in the Mendelian diseases section. Giuseppe Matullo, Giovanni Birolo, and Serena Aneli coordinated all the analyses. Giovanni Birolo and Serena Aneli wrote the manuscript with input from all authors. Alessandra Allione, Elisabetta Casalone, Elvezia Maria Paraboschi, and Diego Ardissino contributed to this study by providing data and/or other resources. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

No sequencing data was generated for this study. The whole-exome data set analyzed during the current study is accessible in an aggregated form in a separate section of the NIG database (Network for Italian Genomes, <http://nigdb.cineca.it>; under the name NIG-ExIT, http://nigdb.cineca.it/search_exit.php). The genotypes of single individuals cannot be published due to informed consent limitations.

ORCID

Giovanni Birolo  <https://orcid.org/0000-0003-0160-9312>

Serena Aneli  <https://orcid.org/0000-0003-4303-2507>

Cornelia Di Gaetano  <https://orcid.org/0000-0001-8938-7093>

Giovanni Cugliari  <https://orcid.org/0000-0002-6080-0718>

Alessia Russo  <https://orcid.org/0000-0002-5494-2218>

Alessandra Allione  <https://orcid.org/0000-0001-9599-309X>

Elisabetta Casalone  <https://orcid.org/0000-0001-5392-0511>

Elisa Giorgio  <https://orcid.org/0000-0003-4076-4649>

Elvezia M. Paraboschi  <https://orcid.org/0000-0002-7935-798X>

Diego Ardissino  <https://orcid.org/0000-0003-0410-3528>

Stefano Duga  <https://orcid.org/0000-0003-3457-1410>

Rosanna Asselta  <https://orcid.org/0000-0001-5351-0619>

Giuseppe Matullo  <https://orcid.org/0000-0003-0674-7757>

REFERENCES

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.
- Antonio, M. L., Gao, Z., Moots, H. M., Lucci, M., Candilio, F., Sawyer, S., Oberreiter, V., Calderon, D., Devitofranceschi, K., Aikens, R. C., Aneli, S., Bartoli, F., Bedini, A., Cheronet, O., Cotter, D. J., Fernandes, D. M., Gasperetti, G., Grifoni, R., Guidi, A., ... Pritchard, J. K. (2019). Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*, *366*(6466), 708–714. <https://doi.org/10.1126/science.aay6826>
- Atherosclerosis, Thrombosis, and Vascular Biology Italian Study Group. (2003). No evidence of association between prothrombotic gene polymorphisms and the development of acute myocardial infarction at a young age. *Circulation*, *107*(8), 1117–1122.
- Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012.
- Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. *Nature Genetics*, *50*(11), 1593–1599.
- Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*, *14*, S3.
- Carvalho, M., Pino, M. A., Karchin, R., Beddor, J., Godinho-Netto, M., Mesquita, R. D., Rodarte, R. S., Vaz, D. C., Monteiro, V. A., Manoukian, S., Colombo, M., Ripamonti, C. B., Rosenquist, R., Suthers, G., Borg, A., Radice, P., Grist, S. A., Monteiro, A. N. A., & Billack, B. (2009). Analysis of a set of missense, frameshift, and in-frame deletion variants of BRCA1. *Mutation Research*, *660*(1–2), 1–11.
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*, 128.
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92.
- Cocca, M., Barbieri, C., Concas, M. P., Robino, A., Brumat, M., Gandin, I., Trudu, M., Sala, C. F., Vuckovic, D., Giroto, G., Matullo, G., Polasek, O., Kolčić, I., Gasparini, P., Soranzo, N., Toniolo, D., & Mezzavilla, M. (2019). A bird's-eye view of Italian genomic variation through whole-genome sequencing. *European Journal of Human Genetics*, *28*, 435–444.
- Day, F. R., Ong, K. K., & Perry, J. R. B. (2018). Elucidating the genetic basis of social interaction and isolation. *Nature Communications*, *9*(1), 2457.

- Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., ... Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12), 1102–1110.
- Di Gaetano, C., Fiorito, G., Ortu, M. F., Rosa, F., Guarrera, S., Pardini, B., Cusi, D., Frau, F., Barlassina, C., Troffa, C., Argiolas, G., Zaninello, R., Fresu, G., Glorioso, N., Piazza, A., & Matullo, G. (2014). Sardinians genetic background explained by runs of homozygosity and genomic regions under positive selection. *PLOS One*, 9(3), e91237.
- Di Gaetano, C., Voglino, F., Guarrera, S., Fiorito, G., Rosa, F., Di Blasio, A. M., Manzini, P., Dianzani, I., Betti, M., Cusi, D., Frau, Barlassina, C., Mirabelli, D., Magnani, C., Glorioso, N., Bonassi, S., Piazza, A., & Matullo, G. (2012). An overview of the genetic structure within the Italian population from genome-wide data. *PLOS One*, 7(9), e43759.
- Do, R., Stitzel, N. O., Won, H.-H., Jørgensen, A. B., Duga, S., Angelica Merlini, P., Kiezun, A., Farrall, M., Goel, A., Zuk, O., Guella, I., Asselta, R., Lange, L. A., Peloso, G. M., Auer, P. L., NHLBI Exome Sequencing Project, Girelli, D., Martinelli, N., Farlow, D. N., & Kathiresan, S. (2015). Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*, 518(7537), 102–106.
- Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, 24(8), 2125–2137. <https://doi.org/10.1093/hmg/ddu733>
- Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Barta, C., Lu, R.-B., Zhukova, O. V., Kim, J.-J., Siniscalco, M., New, M., Li, H., Kajuna, S. L. B., Manolopoulos, V. G., Speed, W. C., Pakstis, A. J., Kidd, J. R., & Kidd, K. K. (2012). A global view of the OCA2-HERC2 region and pigmentation. *Human Genetics*, 131(5), 683–696.
- Dopazo, J., Amadoz, A., Bleda, M., García-Alonso, L., Alemán, A., García-García, F., Rodríguez, J. A., Daub, J. T., Muntané, G., Rueda, A., Vela-Boza, A., López-Domingo, F. J., Florido, J. P., Arce, P., Ruiz-Ferrer, M., Méndez-Vidal, C., Arnold, T. E., Spleiss, O., Alvarez-Tejado, M., ... Antiñolo, G. (2016). 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. *Molecular Biology and Evolution*, 33(5), 1205–1218.
- Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. W., & Hansen, L. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics*, 123(2), 177–187.
- Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: Variant prioritization and Mendelian disease. *Nature Reviews Genetics*, 18(10), 599–612.
- Fiorito, G., Di Gaetano, C., Guarrera, S., Rosa, F., Feldman, M. W., Piazza, A., & Matullo, G. (2016). The Italian genome reflects the history of Europe and the Mediterranean basin. *European Journal of Human Genetics*, 24(7), 1056–1062.
- Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, 13(2), 135–145.
- Guldberg, P., Rey, F., Zschocke, J., Romano, V., François, B., Michiels, L., Ullrich, K., Hoffmann, G. F., Burgard, P., Schmidt, H., Meli, C., Riva, E., Dianzani, I., Ponzone, A., Rey, J., & Güttler, F. (1998). A European multicenter study of phenylalanine hydroxylase deficiency: Classification of 105 mutations and a general system for genotype-based prediction of metabolic phenotype. *The American Journal of Human Genetics*, 63(1), 71–79. <https://doi.org/10.1086/301920>
- Hysi, P. G., Choquet, H., Khawaja, A. P., Wojciechowski, R., Tedja, M. S., Yin, J., Simcoe, M. J., Patasova, K., Mahroo, O. A., Thai, K. K., Cumberland, P. M., Melles, R. B., Verhoeven, V. J. M., Vitart, V., Segre, A., Stone, R. A., Wareham, N., Hewitt, A. W., Mackey, D. A., ... Hammond, C. J. (2020). Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia. *Nature Genetics*, 52(4), 401–407.
- Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., Cannon-Albright, L. A., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., & Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4), 877–885. <https://doi.org/10.1016/j.ajhg.2016.08.016>
- Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, 48(12), 1581–1586.
- Jin, Y., Birlea, S. A., Fain, P. R., Ferrara, T. M., Ben, S., Riccardi, S. L., Cole, J. B., Gowan, K., Holland, P. J., Bennett, D. C., Luiten, R. M., Wolkerstorfer, A., van der Veen, J. P. W., Hartmann, A., Eichner, S., Schuler, G., van Geel, N., Lambert, J., Kemp, E. H., ... Spritz, R. A. (2012). Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature Genetics*, 44(6), 676–680.
- Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., Herman, G. E., Hufnagel, S. B., Klein, T. E., Korf, B. R., McKelvey, K. D., Ormond, K. E., Richards, C. S., Vlangos, C. N., Watson, M., Martin, C. L., & Miller, D. T. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine*, 19(2), 249–255.
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443.
- Kausar, M., Chew, E. G. Y., Ullah, H., Anees, M., Khor, C. C., Foo, J. N., Makitie, O., & Siddiqi, S. (2019). A novel homozygous frameshift variant in causes spondyloocular syndrome in a consanguineous Pakistani family. *Frontiers in Genetics*, 10, 144.
- Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082), 740–743.
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315.
- Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., & Massouras, A. (2019). VarSome: The human genomic variant search engine. *Bioinformatics*, 35(11), 1978–1980.
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma’ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1), W90–W97.
- Kwak, S. H., Chae, J., Choi, S., Kim, M. J., Choi, M., Chae, J.-H., Cho, E.-H., Hwang, T. J., Jang, S. S., Kim, J.-I., Park, K. S., & Bang, Y.-J. (2017). Findings of a 1303 Korean whole-exome sequencing study. *Experimental & Molecular Medicine*, 49(7):e356.
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), D862–D868.

- Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balasckakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvasi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., ... Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology*, *18*(16), 1241–1248.
- Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., ... MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291.
- Lesueur, C., Diergaarde, B., Olshan, A. F., Wunsch-Filho, V., Ness, A. R., Liu, G., Lacko, M., Eluf-Neto, J., Franceschi, S., Lagiou, P., Macfarlane, G. J., Richiardi, L., Boccia, S., Polesel, J., Kjaerheim, K., Zaridze, D., Johansson, M., Menezes, A. M., Curado, M. P., ... Brennan, P. (2016). Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nature Genetics*, *48*(12), 1544–1550.
- Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., Mooney, S. D., & Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, *25*(21), 2744–2750.
- Li, X., & Kahveci, T. (2006). A novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics*, *22*(24), 2980–2987.
- Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*, *37*(3), 235–241.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901.
- Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., ... Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, *39*(10), 1181–1186.
- Malinowski, J. R., Denny, J. C., Bielinski, S. J., Basford, M. A., Bradford, Y., Peissig, P. L., Carrell, D., Crosslin, D. R., Pathak, J., Rasmussen, L., Pacheco, J., Kho, A., Newton, K. M., Li, R., Kullo, I. J., Chute, C. G., Chisholm, R. L., Jarvik, G. P., Larson, E. B., ... Crawford, D. C. (2014). Genetic variants associated with serum thyroid stimulating hormone (TSH) levels in European Americans and African Americans from the eMERGE Network. *PLOS One*, *9*(12), e111301.
- Marth, G. T., Czabarka, E., Murvai, J., & Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, *166*(1), 351–372.
- McKusick, V. A. (1998). *Mendelian inheritance in man: A catalog of human genes and genetic disorders* (12th ed.). Johns Hopkins University Press. <https://omim.org/>
- Mezzavilla, M., & Ghirrotto, S. (2015). Neon: An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. *Journal of Computer Science & Systems Biology*, *8*(1), O37–O44. <https://doi.org/10.4172/jcsb.1000168>
- Morgan, M. D., Pairo-Castineira, E., Rawlik, K., Canela-Xandri, O., Rees, J., Sims, D., Tenesa, A., & Jackson, I. J. (2018). Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nature Communications*, *9*(1), 5271.
- Nutile, T., Ruggiero, D., Herzig, A. F., Tirozzi, A., Nappo, S., Sorice, R., Marangio, F., Bellenguez, C., Leutenegger, A. L., & Ciullo, M. (2019). Whole-exome sequencing in the isolated populations of Cilento from South Italy. *Scientific Reports*, *9*, 4059. <https://doi.org/10.1038/s41598-019-41022-6>
- Olfson, E., Cottrell, C. E., Davidson, N. O., Gurnett, C. A., Heusel, J. W., Stitzel, N. O., Chen, L.-S., Hartz, S., Nagarajan, R., Saccone, N. L., & Bierut, L. J. (2015). Identification of medically actionable secondary findings in the 1000 Genomes. *PLOS One*, *10*(9), e0135193.
- Parolo, S., Lisa, A., Gentilini, D., Di Blasio, A. M., Barlera, S., Nicolis, E. B., Boncoraglio, G. B., Parati, E. A., & Bione, S. (2015). Characterization of the biological processes shaping the genetic structure of the Italian population. *BMC Genetics*, *16*, 132.
- Patronas, Y., Horvath, A., Greene, E., Tsang, K., Bimpaki, E., Haran, M., Nesterova, M., & Stratakis, C. A. (2012). In vitro studies of novel PRKAR1A mutants that extend the predicted R1 α protein sequence into the 3'-untranslated open reading frame: Proteasomal degradation leads to R1 α haploinsufficiency and Carney complex. *The Journal of Clinical Endocrinology and Metabolism*, *97*(3), E496–E502.
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLOS Genetics*, *2*(12), e190.
- Pedersen, B. S., Layer, R. M., & Quinlan, A. R. (2016). Vcfanno: Fast, flexible annotation of genetic variants. *Genome Biology*, *17*(1), 118.
- Polimanti, R., & Gelernter, J. (2017). ADH1B: From alcoholism, natural selection, and cancer to the human phenome. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics*, *177*(2), 113–125.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575.
- R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016.
- Raveane, A., Aneli, S., Montinaro, F., Athanasiadis, G., Barlera, S., Birolo, G., Boncoraglio, G., Di Blasio, A. M., Di Gaetano, C., Pagani, L., Parolo, S., Paschou, P., Piazza, A., Stamatoyannopoulos, G., Angius, A., Brucato, N., Cucca, F., Hellenthal, G., Mulas, A., ... Capelli, C. (2019). Population structure of modern-day Italians reveals patterns of ancient and archaic ancestries in Southern Europe. *Science Advances*, *5*(9), eaaw3492.
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Reh, H. L., & ACMG Laboratory Quality Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, *17*(5), 405–424.
- Russo, A., Di Gaetano, C., Cugliari, G., & Matullo, G. (2018). Advances in the genetics of hypertension: The effect of rare variants. *International Journal of Molecular Sciences*, *19*(3), 688. <https://doi.org/10.3390/ijms19030688>
- Sanchez-Roige, S., Palmer, A. A., Fontanillas, P., Elson, S. L., 23andMe Research Team, the Substance Use Disorder Working Group of the Psychiatric Genomics Consortium, Adams, M. J., Howard, D. M., Edenberg, H. J., Davies, G., Crist, R. C., Deary, I. J., McIntosh, A. M., & Clarke, T.-K. (2019). Genome-Wide Association Study meta-analysis of the Alcohol Use Disorders Identification Test (AUDIT) in two population-based cohorts. *The American Journal of Psychiatry*, *176*(2), 107–118.
- Sazzini, M., Gnechi Ruscone, G. A., Giuliani, C., Sarno, S., Quagliarillo, A., De Fanti, S., Boattini, A., Gentilini, D., Fiorito, G., Catanoso, M., Boiardi, L., Croci, S., Macchioni, P., Mantovani, V., Di Blasio, A. M., Matullo, G., Salvarani, C., Franceschi, C., Pettener, D., ... Luiselli, D. (2016). Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula. *Scientific Reports*, *6*, 32513.
- Shah, N., Hou, Y.-C. C., Yu, H.-C., Sainger, R., Caskey, C. T., Venter, J. C., & Telenti, A. (2018). Identification of misclassified ClinVar variants via disease population prevalence. *American Journal of Human Genetics*, *102*(4), 609–619.
- Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziwska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F., Kwong, A.,

- Ortega Del Vecchio, V. D., Chiang, C. W. K., Bragg-Gresham, J., Pitzalis, M., Nagaraja, R., Tarrier, B., Brennan, C., Uzzau, S., ... Abecasis, G. R. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics*, *47*(11), 1272–1281.
- Smith, F. J. D., Irvine, A. D., Terron-Kwiatkowski, A., Sandilands, A., Campbell, L. E., Zhao, Y., Liao, H., Evans, A. T., Goudie, D. R., Lewis-Jones, S., Arseculeratne, G., Munro, C. S., Sergeant, A., O'Regan, G., Bale, S. J., Compton, J. G., DiGiovanna, J. J., Presland, R. B., Fleckman, P., & McLean, W. H. I. (2006). Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. *Nature Genetics*, *38*(3), 337–342.
- Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., ... Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, *337*(6090), 64–69.
- Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., & Feolo, M. (2013). NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Research*, *42*(D1), D975–D979.
- Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. D., Liu, D., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., Li, A. H., O'Dushlaine, C., Marcketta, A., Staples, J., Schurmann, C., Hawes, A., Maxwell, E., Barnard, L., Lopez, A., ... Baras, A. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature*, *586*(7831), 749–756.
- Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., Altman, R. B., & Klein, T. E. (2012). Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, *92*(4), 414–417.
- 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., ... Zeggini, E. (2017). Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nature Communications*, *8*, 15927.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Birolo G, Aneli S, Di Gaetano C, et al. Functional and clinical implications of genetic structure in 1686 Italian exomes. *Human Mutation*. 2021;42:272–289. <https://doi.org/10.1002/humu.24156>