Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations

Citation for published version:
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, Daniela, T, Wilson, J, Durbin, R, Tyler-Smith, C & Zeggini, E 2017, 'Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations', *Nature Communications*. https://doi.org/10.1038/ncomms15927

Digital Object Identifier (DOI):
10.1038/ncomms15927

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Nature Communications

Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations

Yali Xue[1*], Massimo Mezzavilla[1,2*], Marc Haber[1*], Shane McCarthy[1*] , Yuan Chen[1], Vagheesh Narasimhan[1], Arthur Gilly[1], Qasim Ayub[1], Vincenza Colonna[1,3], Lorraine Southam[1,4], Christopher Finan[1], Andrea Massaia[1,5], Himanshu Chheda[6], Priit Palta[6,7], Graham Ritchie[1,8,9], Jennifer Asimit[1], George Dedoussis[10], Paolo Gasparini[11], Aarno Palotie[1,6,12-16], Samuli Ripatti[1,6,17], Nicole Soranzo[1,18], Daniela Toniolo[19], James F. Wilson[9,20], Richard Durbin[1], Chris Tyler-Smith[1], Eleftheria Zeggini[1]

[1]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambs. CB10 1SA, UK.
[2]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, University of Trieste, 34137 Trieste, Italy.
[3]Consiglio Nazionale delle Ricerche, Istituto di Genetica e Biofisica "Adriano Buzzati-Traverso", via Pietro Castellino 111, 80131 Napoli, Italy.
[4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
[5] National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK
[6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, 00290 Helsinki, Finland.
[7]Estonian Genome Center, University of Tartu, 23B Riia Street, 51010 Tartu, Estonia.
[8]European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambs. CB10 1SD, UK.
[9]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.
[10]Department of Nutrition and Dietetics, Harokopio University Athens, Athens, Eleftheriou Venizelou 70, Kallithea 176 76, Greece.
[11]Medical Genetics, DSM, University of Trieste and IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) Burlo Garofolo Children Hospital, Via dell'Istria, 65, 34137, Trieste, Italy.
[12]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA.
[13]Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02114, USA.
[14]The Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA 02114, USA.
[15]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114, USA.
[16]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA.
[17]Department of Public Health, University of Helsinki, Helsinki FI-00014, Finland.
[18]Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK.

[19]Division of Genetics and Cell Biology, San Raffaele Scientific Institute, via Olgettina 60, 20132, Milan, Italy.
[20]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland, UK.

* Equal contribution to the work

Correspondence should be addressed to Y. X. (ylx@sanger.ac.uk) or E. Z. (Eleftheria@sanger.ac.uk).

Keywords: isolates, genetic drift, functional variation, relaxation of purifying selection, association study strategy

Abstract

The genetic features of isolated populations can boost power in complex-trait association studies, and an in-depth understanding of how their genetic variation has been shaped by their demographic history can help leverage these advantageous characteristics. Here, we perform a comprehensive investigation using 3059 newly-generated low-depth whole-genome sequences from eight European isolates and two matched general populations, together with published data from the 1000 Genomes Project and UK10K. Sequencing data give deeper and richer insights into population demography and genetic characteristics than genotype-chip data, distinguishing related populations more effectively and allowing their functional variants to be studied more fully. We demonstrate relaxation of purifying selection in the isolates, leading to enrichment of rare and low-frequency functional variants, using novel statistics, *DVxy* and *SVxy*. We also develop an isolation-index (*Isx*) that predicts the overall level of such key genetic characteristics and can thus help guide population choice in future complex-trait association studies.

Introduction

Population variation in disease susceptibility has been shaped by environment, demography and evolutionary history. Isolated populations (isolates) have generally experienced bottlenecks and strong genetic drift, so by chance some deleterious rare variants have increased in frequency while some neutral rare variation is lost, both helpful characteristics for the discovery of novel rare variant signals underpinning complex traits[1-3]. Studies to date have focused on individual isolates and have identified several disease-associated signals[4-12]. However, isolates differ in the time when they became isolated, their initial population size, the level of gene flow from outside and other historical demographic factors, and consequently also differ in their power for association studies[2]. We thus generate and analyze low-depth (4x-10x) whole-genome sequences (WGS) from eight cohorts drawn from isolated European populations and compare each isolate with the closest non-isolated (general) population, for which we also generate or access WGS data. We then investigate empirically how these historical differences influence the population-genetic properties of isolates, and frame these insights in terms of their consequences for study design in complex trait association studies.

Results

Samples, sequencing and QC. The dataset includes newly-generated low-depth (4x-10x) WGS from eight cohorts drawn from isolated European populations: one each from Kuusamo in Finland (FIK) and Crete in Greece (GRM[13]), four from Friuli-Venezia Giulia in Italy (IF1, IF2, IF3 and IF4[14]), and one each from Val Borbera in

Italy (IVB[15]) and the Orkney Islands in the UK (UKO[16]); and the closest non-isolated (general) population: Finland (FIG[9]), Greece (GRG), together with publicly available data for Italy (ITG[17]) and UK (UKG[18]) (Fig. 1a and Supplementary Table 1). We generated a superset of variants called in these cohorts and all 26 population samples in the 1000 Genomes Project Phase 3[17], and performed multi-sample genotype calling across all 9375 samples (3059 from the current study, 2353 from the 1000 Genomes Project Phase 3 release, and 3781 from UK10K). Both individual population and amalgamated genotype call data, which have greater than 99% concordance with genotyping data (Supplementary Table 2), are available to the scientific community (Data availability).

General description of the variants in the isolates. We identified approximately 12.2 million (M) variants with minor allele frequency (MAF) ≤2% (rare), 5.5M with MAF >2-≤5% (low-frequency) and 8.3M variants with MAF >5% (common) across the 10 populations newly sequenced here (eight isolates, GRG and FIG). Of these, 10.5%, 0.7% and 0.3%, respectively, are novel (Table 1 and Supplementary Table 3). As expected, most of the isolates have lower numbers of variant sites per genome than their closest general population (Supplementary Fig.1, Supplementary Table 5). We find ~188,000-~513,000 variants that are common with MAF >5.5% in each isolate but with MAF ≤1.4% in its closest general population (Table 1); ~30,000-122,000 of these per isolate have frequency ≤1.4% in all the general samples studied, among which ~150-~700 in coding regions and ~500-~2800 genome-wide are deleterious (Supplementary Table 4). These common and low-frequency variants are thus useful markers for whole-genome association studies in these populations and some of them (if absent from the general population) could potentially lead to novel association signals. They include known examples such as rs76353203 (R19X) in *APOC3* in GRM, which is associated with high-density lipoprotein and triglyceride levels[6].

Population-genetic analyses in the isolates. Previous population-genetic studies of isolates have, with some exceptions[11,19], been based on common variants found on genotyping arrays, and have illustrated general characteristics such as low genetic diversity and longer shared haplotypes[9,13-15,19,20]. Rare variants discovered from sequencing are on average more recent in origin than common variants[21] and therefore more powerful for distinguishing closely-related populations and more informative about recent demographic history. We find that isolates are, as expected, genetically close to their matched general population in principal component analyses (PCA), ADMIXTURE[22] and TreeMix[23] using common variants (Fig. 1b, Supplementary Figs. 2-5 and Supplementary Table 6), but PCA using rare and low-frequency variants, as found previously[24], distinguishes them more clearly from the general population and also from other isolates, particularly among the Italian samples (Fig. 1c, Supplementary Fig. 2). The majority of sharing of variants present just twice across all samples of 36 individuals from each population (f_2 variants[21]) takes place within the same population, and the isolates generally share more with their closest general population than with other populations. This latter trend, however, is not apparent for IF1-IF4, who show little sharing with any other population, pointing to a greater level of isolation and lower level of gene flow with their general population (Fig. 1d, lower triangle and Supplementary Fig. 7), which is confirmed by *f3*-statistics[25] comparing with a worldwide population panel of HGDP-CEPH samples using common SNPs (Supplementary Fig. 6). f_3-f_{10} variant sharing demonstrates sharing by ITG and IVB with both Greek and UK populations (Fig. 1d, upper triangle and Supplementary Fig. 7), potentially indicative of their more ancient heritage.

Population demographic history. All populations studied here, both isolates and general, appear to have shared a comparable effective population size (*Ne*)history before 20 thousand years ago (KYA) based on the multiple sequentially Markovian coalescent (MSMC) method[26] (Supplementary Fig. 9). The isolates diverged from their general populations within the last ~5000 years based on LD estimations[27] (Supplementary Table 7 and Supplementary Fig. 8) and yet had sharp decreases in their population sizes in more recent times as estimated using inferred long segments of identity by descent (IBD)[28] (Fig. 1e, f and Supplementary Fig.10). Different isolates also split from their respective general populations at different times. For example, IF1-IF4 split from ITG ~4-5 KYA, while most other isolates split from their general populations within the last ~1,000 years (Supplementary Table 7).

The different demographic histories of different isolates should lead to different genetic characteristics. To summarize these features in a single quantitative measure that can be calculated from genotype data, as well as sequence data, we developed an isolation index (*Isx*) which combines information on the divergence time from the general population (*Tdg*), *Ne* and migration rate (*M*), such that early-divergence-time isolates with small *Ne* and low *M* have a high *Isx* value (Fig. 2a and Supplementary Fig. 11). The different isolates show different *Isx* values: IF1, IF2, IF3 and IF4 have the highest, while IVB has the lowest (Supplementary Table 8). *Isx* values are highly correlated with other population-genetic characteristics (e.g. Fig. 2b, c, Supplementary Table 11), such as genome-wide pairwise F_{ST} between isolates and their matching general population (reflecting the genetic drift of the isolates) (Supplementary Fig. 12), the total length and number of runs of homozygosity (ROH) (Supplementary Fig. 13), inbreeding coefficient (F) (Supplementary Fig. 14) and length of LD (Supplementary Figs. 15-16 and Supplementary Table 9, 10). All these characteristics are correlated, but the pairwise correlation coefficients show that *Isx* is a slightly better overall predictor of the other measures than any single existing measure (Fig. 2c, Supplementary Fig. 17 and Supplementary Table 11); moreover, it is potentially more robust to confounding factors as it is calculated from three demographic parameters, while the others are all based on single measurements.

Purifying selection analyses. Several lines of evidence suggest relaxed purifying selection in the isolates due to their reduced *Ne*, although as expected we do not detect substantially increased genetic load per genome using the *Rxy* statistic[29] based on all of the variants in the genomes (Fig. 3a and Supplementary Table 12).

First, we see different levels of enrichment of low-frequency functional variants in isolates (Fig. 3b and c, Supplementary Tables 13 and14, Supplementary Figs. 18a) quantified by a new statistic, *DVxy-coding*, developed here (DV: drifted variants). *DVxy-coding* measures the ratio of functional coding variants (missense plus loss-of-function (LoF)) in isolates compared to the closest general population (and vice-versa), adjusted for the corresponding ratios of intergenic variants in order to correct for the effect of genetic drift. We applied this only to a subclass of DVs, defined as low-frequency (2-5%, the best choice according to the sample size we have) in any isolate, yet at least three-fold higher than in the closest general population (and vice versa). We find that *DVxy-coding* is >1 in all isolates and <1 in all general populations (Fig. 3c, Supplementary Fig. 18a and Supplementary Table 13). We also calculated a similar *DVxy-wg* statistic by stratifying whole-genome variants according to their combined annotation dependent depletion (CADD) score (0-5, neutral variants; 5-10, mildly deleterious; 10-20, deleterious; and >20, highly deleterious; these cut-off choices balance the number of variants in each bin to allow us comparable statistical power among all bins, although the conclusions are robust to the particular cut-off values chosen and different bins (Supplementary Figs. 18b and Supplementary Fig. 19)). The *DVxy-wg* values are differentiated for variants with CADD score of 10-20 and significantly so (assessed using the jack-knife bootstrap method) for ones with CADD scores >20, with *DVxy-wg* values >1 in all isolates and <1 in all general populations (Fig. 3b, Supplementary Fig. 18b and Supplementary Table 14). This demonstrates enrichment of low-frequency functional variants, both coding and genome-wide with CADD score >10, in the isolated populations. Moreover, both *DVxy-coding* and *DVxy-wg* values are correlated with *Isx*, suggesting that different isolation characteristics lead to different levels of enrichment of functional variants.

We also investigated the relaxation of purifying selection by assessing functional (missense) singleton variants (SV) pooled for all of the genes that have at least one singleton missense or synonymous variant in a pair of populations (one isolate and its general population), correcting with pooled synonymous variants (*SVxy* statistic,). We find a substantial deviation from 1 for functional singletons in all of the isolates (Fig. 3d and Supplementary Table 15), with *SVxy* values positively correlating with *Isx* (Fig. 2c and Supplementary Fig. 20). We also find that the proportion of relaxed essential genes[30] with *SVxy* >1 in isolates is significantly higher than in the general population (Supplementary Table 15). Such rare and low-frequency drifted functional variants, measured by both DVxy and *DVxy*, are particularly relevant for boosting the power of association studies[6].

Positive selection analyses. We do not find convincing evidence for positive selection in any isolate using deltaDAF[31], PCAdapt[32] or SDS[33], although we do identify some highly differentiated variants (Supplementary Fig. 21 and Supplementary Tables 16,17), including in the protein-coding genes *ALK*, *SPNS2*, *SLC39A11* and *ACSS2*, which can nevertheless be accounted for by drift. Interestingly, we also find six highly-differentiated variants shared between different isolates from Italy, IF2, IF3 and IF4, but interpret them as likely to result from drift or positive selection for the ancestral allele in the ITG (Supplementary Table 17). We find that the SDS method has little power in our samples because of their small size, and failed to detect selection even at the lactose tolerance SNP in the UKO, a known strong signal of recent selection (Supplementary Fig. 22).

Discussion

Isolated populations have special characteristics that can be leveraged to increase the power of association studies, as several previous studies have shown[19,34]. Nevertheless, only a small proportion of functional variants have increased in frequency in any one isolate, so multiple isolates must be investigated to reveal the full diversity of associated variants. Here, we probed an extended allele frequency spectrum of variants potentially underpinning human complex disease through the analysis of whole-genome sequence data in multiple isolates matched to nearby non-isolated populations, capturing common, low-frequency and rare variants. We quantified different levels of isolation resulting from different demographic histories and have demonstrated that the *Isx* statistic, calculated even from SNP-chip data, reliably captures these relevant features. This study provides a systematic evaluation of the genetic characteristics of multiple European isolates and for the first time empirically demonstrates enrichment of rare functional variants across multiple isolates. With the advent of large-scale whole-genome sequencing, studies in isolates are poised to continue as major contributors to our understanding of complex disease etiology.

Methods

Dataset and variant calling: The dataset includes 3059 whole-genome low-depth sequences generated at The Wellcome Trust Sanger Institute using the Illumina Genome Analyzer II and Illumina HiSeq 2000 platforms, as well as 100 high-depth sequences from the Illumina HiSeq X Ten (Fig. 1a and Supplementary Table 1). Informed consent was obtained from all subjects and the study was approved by the HMDMC (Human Materials and Data Management Committee) of the Welcome Trust Sanger Institute. The multi-sample genotype calling across all of the low-coverage sequencing data from the current study, as well as 2353 from the 1000 Genomes Project Phase 3 release, and 3781 from UK10K (a total of 9375) was performed with the defined site selection criteria (Supplementary Note). Genotype likelihoods were calculated with samtools/bcftools (0.2.0-rc9) and then genotypes were called and phased using Beagle v4 (r1274)[35]. We assessed the performance of the genotype calling from the low coverage data using the available genotype chip data for a subset of the cohorts consisting of 4665 individuals, and calculated the discordance rates on chromosome 20 separately for the categories REF-REF, REF-ALT and ALT-ALT.

The sample sizes are very different across these collections, and we used three different standard-sized subsets of the samples for different analyses: (1) the whole dataset; (2) the sample-size-matched dataset, obtained either by randomly selecting samples from general population to match the isolated population (for example, we randomly select 377 from FIG to match FIK), or by randomly selecting a subset of the isolated population to match the general population (for example, we randomly select 108 IVB to match the general population ITG); (3) the minimum-sample-size dataset of 36 individuals per population. By doing this, we maximize the use of the data for different analyses, and we specify which dataset is used for each analysis. The sequencing depth is also different across different populations, within a 2.5-fold range (apart from GRG, in which variants were called differently, details in Supplementary Notes), and we allowed for these differences when interpreting the results.

Variant counts: We first re-annotated all variants using the Variant Effect Predictor (VEP) annotation from Ensembl 76 with the "- pick" option, which gives one annotation per variant. We then performed variant counting at both the population and individual level, stratifying by functional categories and frequency bins. These counts were either plotted in figures or summarized as median values in tables. We carried out these analyses using both the sample-size-matched dataset and the minimum-sample-size dataset.

Population-genetic analyses: We used the whole dataset for the analyses in this section, unless otherwise specified. Principal component analyses (PCAs) were performed separately with common variants or rare variants using EIGENSTRAT v.501[36]. Shared ancestry between the populations studied here was evaluated using ADMIXTURE v1.22[22]. The relationships between the populations studied here, combined with worldwide populations from the HGDP-CEPH panel[37], were also examined using ancestry graph analyses implemented in TreeMix v.1.12[23]. We also used formal test of f3-satitisitcs[25] to investigate population mixture in the history of the populations studied here, as well as worldwide populations from the HGDP-CEPH panel. Rare f_2 variants (with only two copies of the alternative allele in the minimum-sample-size dataset) and moderately rare $f_{3\text{-}10}$ variants (3-10 copies of the alternative allele in the same dataset) are particularly informative for investigating recent human history[21]. We investigated the sharing pattern of these two types of variant by summing all f_2 variants or any random two alleles of the $f_{3\text{-}10}$ variants shared by pairs of individuals. We plotted the results as a heat map using the image[1] function from the base R package (https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/image.html). Variants were aggregated by pair of individuals using the 'count' function of the plyr package, then arranged in matrix form and colorized using 'colorRampPalette' from the colorspace package (https://cran.r-project.org/web/packages/colorspace/index.html). Runs of homozygosity (ROH), inbreeding coefficient (F) as well as the length of LD-blocks were calculated in PLINK, and finally genome-wide F_{ST} values between isolates and their general populations were calculated with the software 4P[38] using the minimum-sample-size dataset.

Demographic inferences: LD-based[39-41] demographic inference was performed in the NeON R package[27] using the minimum-sample-size dataset; the median and confidence interval were estimated using the 50th, 5th and 95th percentiles of the distribution of long-term *Ne* in each time interval. We used the multiple sequentially Markovian coalescent (MSMC) method[26] to infer demographic changes before 20,000 years ago using four individual sequences from each population. In order to account for some loss of heterozygous sites in the low-depth data, we used a slow mutation rate of 0.8 x 10^{-8} mutations per nucleotide per generation and a longer generation time of 33 years. We then estimated more recent demographic changes (from the present to ~9,000 years ago) using IBDNe[28] with the minimum-sample-size dataset. We used IBDseq[42] to detect IBD segments in sequence data from chromosome 2 in all populations. We then used IBDNe with the default parameters and a minimum IBD segment length of 2 centiMorgan (cM) units. We assumed a generation time of 29 years.

Isolation index: In order to quantify the different isolation levels of different isolates, we developed an index that combines three demographic parameters: (a) *Tdg*, (b) *Ne*, and (c) the level of private isolate ancestry (*M*). We call this estimate the Isolation index (*Isx*). It is defined as:

$$Isx = \frac{\log(Tdg(100 * M)^2)}{\log(Ne)}$$

Both *Tdg* and *Ne* were inferred from the LD-based method using the NeON R package[27]. *M* is difficult to estimate directly from SNP genotype data, so here we estimated the difference of shared ancestral components between an isolate and its general population from ADMIXTURE analysis. We ran ADMIXTURE with only one isolate and it closest general population using K=2. We then estimated the difference in the means of ancestry between the isolate and its general population. The *M* parameter was defined as Delta Ancestry.

***Rxy* analysis:** Rxy statistics[29] between each pair of populations (an isolate and its closest general population) for different functional categories were calculated using the matched-sample-size data for missense and LoF variants, including stop gain, splice donor and acceptor variants, using synonymous variants as controls (we did not use intragenic variants as control because of the ascertainment in the ITG which has high-depth exome sequences and low depth for the rest of the genome). We also calculated *Rxy* statistics for variants with CADD scores[43] greater than 10 and 20, using variants with CADD scores less than 5 as controls. The mean and standard deviation for each Rxy value were obtained from 100 bootstraps.

***DVxy* analysis:** A new statistic, *DVxy*, was developed to quantify the enrichment of low-frequency functional variants in the isolates using both the matched-sample-size and minimum-sample-size datasets. It calculates the proportion of functional variants in each isolate compared with its general population, correcting for genetic drift at the same time. We calculated *DVxy* specifically for the subset of variants with DAF 2-5% in the isolate, and at least three times lower in its closest general population, or vice-versa. We called these variants "drifted variants" (DV). *DVxy* was calculated for both coding regions and whole genomes.

For coding variants, we defined missense or missense plus LoF variants as functional variants. We counted the number of functional DVs and neutral (intergenic) DVs in each isolate (population *x*) and the corresponding general population (population *y*). The ratio between the fraction of DV variants from the isolated population (corrected by the count of intergenic variants) and the corresponding fraction of DV variants from its general population was defined as the *DVxy* statistic. If *DVxy* is equal to 1, there is no enrichment for the functional DVs in the isolate; less than 1 indicates depletion, and greater than 1 indicates enrichment.

$$DVxy\_coding = {\frac{\%DVx\ missense}{\%DVx\ intergenic}}\Bigg/{\frac{\%DVy\ missense}{\%DVy\ intergenic}}$$

For the whole genome, we used different CADD score cut-offs and bins. We calculated a DV statistic by stratifying the variants according to their CADD scores (0-5, neutral variants; 5-10, mildly deleterious; 10-20, deleterious; and greater than 20, highly deleterious) for each isolate and its closest general population. We finally calculated a ratio of the fraction of DV variants (from each class) between the isolate and its general population, and vice-versa. The following formula shows the *DVxy-wg* calculation for variants with CADD score between *i* (isolate) and *j* (general population).

$$DVxy_{CADD(ij)} = {\%\ DVx\ (CADD\ i - j)}\Big/{\%DVy(CADD\ i - j)}$$

The 95% confidence interval for each calculation was obtained by randomly sampling data from 20 chromosomes 100 times.

***SVxy* analysis:** We further investigated the relaxation of purifying selection in the isolated populations using singleton variants. Here, we also used the minimum-sample-size dataset. Another new statistic, *SVxy*, was developed to measure the ratio of missense vs synonymous singletons per gene in each population, as well as the ratio of the sum of singletons in all genes which have at least one singleton in the pair of the populations (one isolate and one general population). We counted the number of missense singletons and synonymous singletons per gene in each population, and *SVgene* was calculated as:

SVgene = (SV missense count +1)/ (SV synonymous count +1)

SVgene >1 indicates relaxation of purifying selection; *SVgene* = 1 indicates neutrality; and *SVgene* <1 indicates purifying selection.

We then divided the gene list into essential genes[30] and non-essential genes (the rest), and calculated a statistic, G_{SV}, for each population, defined as:

G_{SV} = percentage of essential genes with *SVgene* >1/percentage of non-essential genes with *SVgene* >1

We finally calculated a statistic, *SVxy*, which is the ratio of *SVpop* of each isolate to *SVpop* of its general population. *SVpop* for each isolate and its general population was calculated using all genes which have at least one singleton in the pair of the populations and defined as *SVpop* = Σ (SV missense counts)/Σ(SV synonymous counts).

We used the same annotation as in the variant counts. We calculated a confidence interval for each estimate using bootstrapping of 80% of the genes 100 times.

Correlation analyses: We calculated pair-wise correlation coefficients between the *Isx* values, population-genetic measurements ROH, F, F_{ST}, and number and length of LD blocks, as well as the newly-developed statistics *DVxy* and SVxy using the Pearson correlation in R.

Positive selection analyses: We calculated genome-wide pairwise derived allele frequency differences (deltaDAF) for each pair of populations (an isolate and its general population) as described previously[31] using the matched-sample-size dataset. We also carried out PCAdapt analyses[32] for each pair of populations using the whole dataset. Both analyses look for high derived allele frequency variants in the isolates, and will not be affected by sample size. Finally, we ran the singleton density score (SDS) method[33] using the whole UKO and UKG datasets, which have the largest sample sizes for both isolate and its general population, and thus the greatest power for this method.

Data availability:

Amalgamated genotype calls across all populations studied are available through the European Genome/Phenome Archive (EGAD00001002014) with Data Access Agreement described in the Supplementary Information.

References

1. Zeggini, E. Using genetically isolated populations to understand the genomic basis of disease. *Genome Med* **6**, 83 (2014).
2. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief Funct Genomics* **13**, 371-7 (2014).
3. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
4. Pollin, T.I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702-5 (2008).
5. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* **44**, 1326-9 (2012).
6. Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 (2013).
7. Huyghe, J.R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197-201 (2013).
8. Li, A.H. *et al.* Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat Genet* **47**, 640-2 (2015).
9. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 (2014).
10. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190-3 (2014).
11. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**, 1272-81 (2015).
12. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-8 (2014).
13. Panoutsopoulou, K. *et al.* Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun* **5**, 5345 (2014).
14. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* **21**, 659-65 (2013).
15. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur J Hum Genet* **21**, 89-94 (2013).

16. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* **40**, 437-42 (2008).

17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

18. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

19. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).

20. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am J Hum Genet* **83**, 359-72 (2008).

21. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528 (2014).

22. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-64 (2009).

23. Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).

24. O'Connor, T.D. *et al.* Rare variation facilitates inferences of fine-scale population structure in humans. *Mol Biol Evol* **32**, 653-60 (2015).

25. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-494 (2009).

26. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-25 (2014).

27. Mezzavilla, M. & Ghirotto, S. Neon: An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. *J Comput Sci Syst Biol* **8**, 37-44 (2015).

28. Browning, S.R. & Browning, B.L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet* **97**, 404-18 (2015).

29. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-31 (2015).

30. Blomen, V.A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-6 (2015).

31. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol* **15**, R88 (2014).

32. Duforet-Frebourg, N., Bazin, E. & Blum, M.B.G. Genome scans for detecting local adaptation using a Bayesian factor model. *Mol Biol Evol* **31**, 2483-2495 (2014).

33. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).

34. Zoledziewska, M., Sidore, C. & Chiang, C.W. Height-reducing variants and selection for short stature in Sardinia. *Nat Genet* **47**, 1352-1356 (2015).

35. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).

36. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).

37. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-4 (2008).

38. Benazzo, A., Panziera, A. & Bertorelle, G. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol Evol* **5**, 172-175 (2014).

39. Hill, W.G. Estimation of effective population size from data on linkage disequilibrium. *Genetical Res.* **38**, 209-216 (1981).

40. Hayes, B.J., Visscher, P.M., McPartlan, H.C. & Goddard, M.E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**, 635-43 (2003).

41. Tenesa, A. *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**, 520-6 (2007).

42. Browning, B.L. & Browning, S.R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* **93**, 840-51 (2013).

Author contributions

Y.X., C.T.-S., R.D. and E.Z.: design and supervision of the project. G.D., P.G., A.P., S.R., N.S., D.T. and J.F.W.: population liaison, sampling and DNA provision. N.S., JF.W. and R.D: comments and approval of the manuscript on behalf of the population consortia. Y.X., M.M. and M.H.: statistical method development. M.M., M.H., S.M., V.N., A.G., Q.A., V.C., L.S., C.F., G.R., H.C. and P.R. and J.A.: population-genetic analyses, statistical analyses and data interpretation. Y.C. and A.M.: bioinformatics support. S.M., N.S. and R.D: data processing and QC. Y.X., M.M., M.H., S.M., V.C., C.T-.S. and E.Z.: manuscript drafting. All authors: approval of the final version of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

We thank all study participants for making this work possible. Our work was supported by the Wellcome Trust (098051). We also thank the UK10K Consortium and SISu Consortium for making their data available to this study; detailed acknowledgements for these contributions are included in the Supplementary Information.

Figure legends

Fig. 1. General characteristics and demographic history of isolated and matched general populations. a. Geographical locations of samples. The base map was plotted in R using the mapdata package and circles were added using Photoshop. b. PCA using common variants. c. PCA using low-frequency variants. d. Sharing of rare variants within and between populations. Upper left triangle: f_2 variants; lower right triangle $f_{3\text{-}f10}$ variants. e. Effective population size (Ne) inferred from IBDNe for UKO and UKG during the past 9 KY. f. The lowest Ne inferred by IBDNe for all populations for the past 3KY, plotted as a function of the time at which it occurred.

Fig. 2. Isolation index (*Isx*) and its correlation with other genetic measures. a. Information summarized in *Isx*. b. Example of the correlation between *Isx* and other statistics, here *DVxy-coding*. c. Summary of the correlations between *Isx* and other population-genetic statistics. All the correlation coefficients are high and statistically significant.

Fig. 3. Purifying selection in the isolates and general populations. a. *Rxy*-missense statistic in each isolate, showing no evidence for increased genetic load in the isolates. The mean and standard deviation for each Rxy value from 100 bootstraps are shown. b. *DVxy-wg* (*DVxy*-whole genome) statistic in isolates and general populations, stratified by CADD score, showing enrichment of highly-functional low-frequency variants. c. *DVxy-coding* statistic in isolates and general populations, showing enrichment of low-frequency missense variants in isolates. d. SVxy-missense statistic in each isolate, showing relaxation of purifying selection in isolates in singletons. The standard errors for both *DVxy* and *SVxy* were calculated by randomly sampling data from 20 chromosomes 100 times. All of these analyses are based on the minimum-sample-size dataset (36 individuals from each population).

Tables:

Table 1. Summary of variants discovered in this study

POP	n	average depth	MAF ≤2%		MAF >2-≤5%		MAF >5%		Novel common SNPs in isolate*	Novel common SNPs in isolate**
			total	novel %	total	novel %	total	novel %		
FIK	377	4x	4,066,373	10.90	1,553,076	1.20	6,025,077	0.70	190,527	70,579
FIG	1564	6x	6,548,833	11.80	1,540,915	0.80	6,053,704	0.70	n.a.	n.a.
GRM	249	4x	5,129,513	7.20	1,447,981	1.10	6,111,923	0.80	513,272	49,884
GRG* **	99	10-30x	3,757,110	n.a.	1,321,955	n.a.	5,842,537	n.a.	n.a.	n.a.
IF1	60	4-10x	1,456,881	1.30	1,420,929	1.30	5,890,714	0.80	320,191	119,157
IF2	45	4-10x	1,063,098	1.30	1,554,145	1.00	6,001,568	0.80	273,694	94,496
IF3	47	4-10x	961,059	1.30	1,455,284	1.10	6,068,304	0.80	299,603	107,281
IF4	36	4-10x	1,030,673	1.30	1,124,789	1.10	6,001,625	0.80	308,356	122,254
IVB	222	6x	4,857,767	1.60	1,396,799	0.80	6,112,476	0.80	188,972	30,284
UKO	397	4x	5,963,416	11.70	1,471,782	0.80	6,047,383	0.80	193,300	36,512
Total	3096		12,218,797	10.50	5,503,179	0.70	8,301,524	0.30		

'Novel' variants are those not found in 1000 Genomes Project Phase 3 or UK10K project. *Variants that are common (minor allele frequency, MAF ≥5.6%, alternative allele count ≥ 4) in an isolated population but not common (MAF <1.4%, alternative allele count ≤ 1) in its closest general population. **Variants that are common (MAF ≥5.6%, alternative allele count ≥ 4) in an isolated population but not (MAF <1.4%, alternative allele count ≤ 1) in *any* of the general populations. ***Different variant calling procedure in this population.

Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations

Yali Xue[1*], Massimo Mezzavilla[1,2*], Marc Haber[1*], Shane McCarthy[1*] , Yuan Chen[1], Vagheesh Narasimhan[1], Arthur Gilly[1], Qasim Ayub[1], Vincenza Colonna[1,3], Lorraine Southam[1,4], Christopher Finan[1], Andrea Massaia[1,5], Himanshu Chheda[6], Priit Palta[6,7], Graham Ritchie[1,8,9], Jennifer Asimit[1], George Dedoussis[10], Paolo Gasparini[11], Aarno Palotie[1,6,12-16], Samuli Ripatti[1,6,17], Nicole Soranzo[1,18], Daniela Toniolo[19], James F. Wilson[9,20], Richard Durbin[1], Chris Tyler-Smith[1], Eleftheria Zeggini[1]

[1]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambs. CB10 1SA, UK.
[2]Institute for Maternal and Child Health, IRCCS Burlo Garofolo, University of Trieste, 34137 Trieste, Italy.
[3]Consiglio Nazionale delle Ricerche, Istituto di Genetica e Biofisica "Adriano Buzzati-Traverso", via Pietro Castellino 111, 80131 Napoli, Italy.
[4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
[5] National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK
[6]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, 00290 Helsinki, Finland.
[7]Estonian Genome Center, University of Tartu, 23B Riia Street, 51010 Tartu, Estonia.
[8]European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambs. CB10 1SD, UK.
[9]MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK.
[10]Department of Nutrition and Dietetics, Harokopio University Athens, Athens, Eleftheriou Venizelou 70, Kallithea 176 76, Greece.
[11]Medical Genetics, DSM, University of Trieste and IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) Burlo Garofolo Children Hospital, Via dell'Istria, 65, 34137, Trieste, Italy.
[12]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA.
[13]Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA 02114, USA.
[14]The Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA 02114, USA.
[15]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114, USA.
[16]Department of Neurology, Massachusetts General Hospital, Boston, MA 02114, USA.
[17]Department of Public Health, University of Helsinki, Helsinki FI-00014, Finland.
[18]Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK.

[19]Division of Genetics and Cell Biology, San Raffaele Scientific Institute, via Olgettina 60, 20132, Milan, Italy.
[20]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland, UK.

* Equal contribution to the work

Correspondence should be addressed to Y. X. (ylx@sanger.ac.uk) or E. Z. (Eleftheria@sanger.ac.uk).

Keywords: isolates, genetic drift, functional variation, relaxation of purifying selection, association study strategy

Abstract

The genetic features of isolated populations can boost power in complex-trait association studies, and an in-depth understanding of how their genetic variation has been shaped by their demographic history can help leverage these advantageous characteristics. Here, we perform a comprehensive investigation using 3059 newly-generated low-depth whole-genome sequences from eight European isolates and two matched general populations, together with published data from the 1000 Genomes Project and UK10K. Sequencing data give deeper and richer insights into population demography and genetic characteristics than genotype-chip data, distinguishing related populations more effectively and allowing their functional variants to be studied more fully. We demonstrate relaxation of purifying selection in the isolates, leading to enrichment of rare and low-frequency functional variants, using novel statistics, *DVxy* and *SVxy*. We also develop an isolation-index (*Isx*) that predicts the overall level of such key genetic characteristics and can thus help guide population choice in future complex-trait association studies.

Introduction

Population variation in disease susceptibility has been shaped by environment, demography and evolutionary history. Isolated populations (isolates) have generally experienced bottlenecks and strong genetic drift, so by chance some deleterious rare variants have increased in frequency while some neutral rare variation is lost, both helpful characteristics for the discovery of novel rare variant signals underpinning complex traits[1-3]. Studies to date have focused on individual isolates and have identified several disease-associated signals[4-12]. However, isolates differ in the time when they became isolated, their initial population size, the level of gene flow from outside and other historical demographic factors, and consequently also differ in their power for association studies[2]. We thus generate and analyze low-depth (4x-10x) whole-genome sequences (WGS) from eight cohorts drawn from isolated European populations and compare each isolate with the closest non-isolated (general) population, for which we also generate or access WGS data. We then investigate empirically how these historical differences influence the population-genetic properties of isolates, and frame these insights in terms of their consequences for study design in complex trait association studies.

Results

Samples, sequencing and QC. The dataset includes newly-generated low-depth (4x-10x) WGS from eight cohorts drawn from isolated European populations: one each from Kuusamo in Finland (FIK) and Crete in Greece (GRM[13]), four from Friuli-Venezia Giulia in Italy (IF1, IF2, IF3 and IF4[14]), and one each from Val Borbera in

Italy (IVB[15]) and the Orkney Islands in the UK (UKO[16]); and the closest non-isolated (general) population: Finland (FIG[9]), Greece (GRG), together with publicly available data for Italy (ITG[17]) and UK (UKG[18]) (Fig. 1a and Supplementary Table 1). We generated a superset of variants called in these cohorts and all 26 population samples in the 1000 Genomes Project Phase 3[17], and performed multi-sample genotype calling across all 9375 samples (3059 from the current study, 2353 from the 1000 Genomes Project Phase 3 release, and 3781 from UK10K). Both individual population and amalgamated genotype call data, which have greater than 99% concordance with genotyping data (Supplementary Table 2), are available to the scientific community (Data availability).

General description of the variants in the isolates. We identified approximately 12.2 million (M) variants with minor allele frequency (MAF) ≤2% (rare), 5.5M with MAF >2-≤5% (low-frequency) and 8.3M variants with MAF >5% (common) across the 10 populations newly sequenced here (eight isolates, GRG and FIG). Of these, 10.5%, 0.7% and 0.3%, respectively, are novel (Table 1 and Supplementary Table 3). As expected, most of the isolates have lower numbers of variant sites per genome than their closest general population (Supplementary Fig.1, Supplementary Table 5). We find ~188,000-~513,000 variants that are common with MAF >5.5% in each isolate but with MAF ≤1.4% in its closest general population (Table 1); ~30,000-122,000 of these per isolate have frequency ≤1.4% in all the general samples studied, among which ~150-~700 in coding regions and ~500-~2800 genome-wide are deleterious (Supplementary Table 4). These common and low-frequency variants are thus useful markers for whole-genome association studies in these populations and some of them (if absent from the general population) could potentially lead to novel association signals. They include known examples such as rs76353203 (R19X) in *APOC3* in GRM, which is associated with high-density lipoprotein and triglyceride levels[6].

Population-genetic analyses in the isolates. Previous population-genetic studies of isolates have, with some exceptions[11,19], been based on common variants found on genotyping arrays, and have illustrated general characteristics such as low genetic diversity and longer shared haplotypes[9,13-15,19,20]. Rare variants discovered from sequencing are on average more recent in origin than common variants[21] and therefore more powerful for distinguishing closely-related populations and more informative about recent demographic history. We find that isolates are, as expected, genetically close to their matched general population in principal component analyses (PCA), ADMIXTURE[22] and TreeMix[23] using common variants (Fig. 1b, Supplementary Figs. 2-5 and Supplementary Table 6), but PCA using rare and low-frequency variants, as found previously[24], distinguishes them more clearly from the general population and also from other isolates, particularly among the Italian samples (Fig. 1c, Supplementary Fig. 2). The majority of sharing of variants present just twice across all samples of 36 individuals from each population (f_2 variants[21]) takes place within the same population, and the isolates generally share more with their closest general population than with other populations. This latter trend, however, is not apparent for IF1-IF4, who show little sharing with any other population, pointing to a greater level of isolation and lower level of gene flow with their general population (Fig. 1d, lower triangle and Supplementary Fig. 7), which is confirmed by f3-statistics[25] comparing with a worldwide population panel of HGDP-CEPH samples using common SNPs (Supplementary Fig. 6). f_3-f_{10} variant sharing demonstrates sharing by ITG and IVB with both Greek and UK populations (Fig. 1d, upper triangle and Supplementary Fig. 7), potentially indicative of their more ancient heritage.

Population demographic history. All populations studied here, both isolates and general, appear to have shared a comparable effective population size (*Ne*)history before 20 thousand years ago (KYA) based on the multiple sequentially Markovian coalescent (MSMC) method[26] (Supplementary Fig. 9). The isolates diverged from their general populations within the last ~5000 years based on LD estimations[27] (Supplementary Table 7 and Supplementary Fig. 8) and yet had sharp decreases in their population sizes in more recent times as estimated using inferred long segments of identity by descent (IBD)[28] (Fig. 1e, f and Supplementary Fig.10). Different isolates also split from their respective general populations at different times. For example, IF1-IF4 split from ITG ~4-5 KYA, while most other isolates split from their general populations within the last ~1,000 years (Supplementary Table 7).

The different demographic histories of different isolates should lead to different genetic characteristics. To summarize these features in a single quantitative measure that can be calculated from genotype data, as well as sequence data, we developed an isolation index (*Isx*) which combines information on the divergence time from the general population (*Tdg*), *Ne* and migration rate (*M*), such that early-divergence-time isolates with small *Ne* and low *M* have a high *Isx* value (Fig. 2a and Supplementary Fig. 11). The different isolates show different *Isx* values: IF1, IF2, IF3 and IF4 have the highest, while IVB has the lowest (Supplementary Table 8). *Isx* values are highly correlated with other population-genetic characteristics (e.g. Fig. 2b, c, Supplementary Table 11), such as genome-wide pairwise F_{ST} between isolates and their matching general population (reflecting the genetic drift of the isolates) (Supplementary Fig. 12), the total length and number of runs of homozygosity (ROH) (Supplementary Fig. 13), inbreeding coefficient (F) (Supplementary Fig. 14) and length of LD (Supplementary Figs. 15-16 and Supplementary Table 9, 10). All these characteristics are correlated, but the pairwise correlation coefficients show that *Isx* is a slightly better overall predictor of the other measures than any single existing measure (Fig. 2c, Supplementary Fig. 17 and Supplementary Table 11); moreover, it is potentially more robust to confounding factors as it is calculated from three demographic parameters, while the others are all based on single measurements.

Purifying selection analyses. Several lines of evidence suggest relaxed purifying selection in the isolates due to their reduced *Ne*, although as expected we do not detect substantially increased genetic load per genome using the *Rxy* statistic[29] based on all of the variants in the genomes (Fig. 3a and Supplementary Table 12).

First, we see different levels of enrichment of low-frequency functional variants in isolates (Fig. 3b and c, Supplementary Tables 13 and14, Supplementary Figs. 18a) quantified by a new statistic, *DVxy-coding*, developed here (DV: drifted variants). *DVxy-coding* measures the ratio of functional coding variants (missense plus loss-of-function (LoF)) in isolates compared to the closest general population (and vice-versa), adjusted for the corresponding ratios of intergenic variants in order to correct for the effect of genetic drift. We applied this only to a subclass of DVs, defined as low-frequency (2-5%, the best choice according to the sample size we have) in any isolate, yet at least three-fold higher than in the closest general population (and vice versa). We find that *DVxy-coding* is >1 in all isolates and <1 in all general populations (Fig. 3c, Supplementary Fig. 18a and Supplementary Table 13). We also calculated a similar *DVxy-wg* statistic by stratifying whole-genome variants according to their combined annotation dependent depletion (CADD) score (0-5, neutral variants; 5-10, mildly deleterious; 10-20, deleterious; and >20, highly deleterious; these cut-off choices balance the number of variants in each bin to allow us comparable statistical power among all bins, although the conclusions are robust to the particular cut-off values chosen and different bins (Supplementary Figs. 18b and Supplementary Fig. 19)). The *DVxy-wg* values are differentiated for variants with CADD score of 10-20 and significantly so (assessed using the jack-knife bootstrap method) for ones with CADD scores >20, with *DVxy-wg* values >1 in all isolates and <1 in all general populations (Fig. 3b, Supplementary Fig. 18b and Supplementary Table 14). This demonstrates enrichment of low-frequency functional variants, both coding and genome-wide with CADD score >10, in the isolated populations. Moreover, both *DVxy-coding* and *DVxy-wg* values are correlated with *Isx*, suggesting that different isolation characteristics lead to different levels of enrichment of functional variants.

We also investigated the relaxation of purifying selection by assessing functional (missense) singleton variants (SV) pooled for all of the genes that have at least one singleton missense or synonymous variant in a pair of populations (one isolate and its general population), correcting with pooled synonymous variants (*SVxy* statistic,). We find a substantial deviation from 1 for functional singletons in all of the isolates (Fig. 3d and Supplementary Table 15), with *SVxy* values positively correlating with *Isx* (Fig. 2c and Supplementary Fig. 20). We also find that the proportion of relaxed essential genes[30] with *SVxy* >1 in isolates is significantly higher than in the general population (Supplementary Table 15). Such rare and low-frequency drifted functional variants, measured by both *SVxy* and *DVxy*, are particularly relevant for boosting the power of association studies[6].

Positive selection analyses. We do not find convincing evidence for positive selection in any isolate using deltaDAF[31], PCAdapt[32] or SDS[33], although we do identify some highly differentiated variants (Supplementary Fig. 21 and Supplementary Tables 16,17), including in the protein-coding genes *ALK*, *SPNS2*, *SLC39A11* and *ACSS2*, which can nevertheless be accounted for by drift. Interestingly, we also find six highly-differentiated variants shared between different isolates from Italy, IF2, IF3 and IF4, but interpret them as likely to result from drift or positive selection for the ancestral allele in the ITG (Supplementary Table 17). We find that the SDS method has little power in our samples because of their small size, and failed to detect selection even at the lactose tolerance SNP in the UKO, a known strong signal of recent selection (Supplementary Fig. 22).

Discussion

Isolated populations have special characteristics that can be leveraged to increase the power of association studies, as several previous studies have shown[19,34]. Nevertheless, only a small proportion of functional variants have increased in frequency in any one isolate, so multiple isolates must be investigated to reveal the full diversity of associated variants. Here, we probed an extended allele frequency spectrum of variants potentially underpinning human complex disease through the analysis of whole-genome sequence data in multiple isolates matched to nearby non-isolated populations, capturing common, low-frequency and rare variants. We quantified different levels of isolation resulting from different demographic histories and have demonstrated that the *Isx* statistic, calculated even from SNP-chip data, reliably captures these relevant features. This study provides a systematic evaluation of the genetic characteristics of multiple European isolates and for the first time empirically demonstrates enrichment of rare functional variants across multiple isolates. With the advent of large-scale whole-genome sequencing, studies in isolates are poised to continue as major contributors to our understanding of complex disease etiology.

Methods

Dataset and variant calling: The dataset includes 3059 whole-genome low-depth sequences generated at The Wellcome Trust Sanger Institute using the Illumina Genome Analyzer II and Illumina HiSeq 2000 platforms, as well as 100 high-depth sequences from the Illumina HiSeq X Ten (Fig. 1a and Supplementary Table 1). Informed consent was obtained from all subjects and the study was approved by the HMDMC (Human Materials and Data Management Committee) of the Welcome Trust Sanger Institute. The multi-sample genotype calling across all of the low-coverage sequencing data from the current study, as well as 2353 from the 1000 Genomes Project Phase 3 release, and 3781 from UK10K (a total of 9375) was performed with the defined site selection criteria (Supplementary Note). Genotype likelihoods were calculated with samtools/bcftools (0.2.0-rc9) and then genotypes were called and phased using Beagle v4 (r1274)[35]. We assessed the performance of the genotype calling from the low coverage data using the available genotype chip data for a subset of the cohorts consisting of 4665 individuals, and calculated the discordance rates on chromosome 20 separately for the categories REF-REF, REF-ALT and ALT-ALT.

The sample sizes are very different across these collections, and we used three different standard-sized subsets of the samples for different analyses: (1) the whole dataset; (2) the sample-size-matched dataset, obtained either by randomly selecting samples from general population to match the isolated population (for example, we randomly select 377 from FIG to match FIK), or by randomly selecting a subset of the isolated population to match the general population (for example, we randomly select 108 IVB to match the general population ITG); (3) the minimum-sample-size dataset of 36 individuals per population. By doing this, we maximize the use of the data for different analyses, and we specify which dataset is used for each analysis. The sequencing depth is also different across different populations, within a 2.5-fold range (apart from GRG, in which variants were called differently, details in Supplementary Notes), and we allowed for these differences when interpreting the results.

Variant counts: We first re-annotated all variants using the Variant Effect Predictor (VEP) annotation from Ensembl 76 with the "- pick" option, which gives one annotation per variant. We then performed variant counting at both the population and individual level, stratifying by functional categories and frequency bins. These counts were either plotted in figures or summarized as median values in tables. We carried out these analyses using both the sample-size-matched dataset and the minimum-sample-size dataset.

Population-genetic analyses: We used the whole dataset for the analyses in this section, unless otherwise specified. Principal component analyses (PCAs) were performed separately with common variants or rare variants using EIGENSTRAT v.501[36]. Shared ancestry between the populations studied here was evaluated using ADMIXTURE v1.22[22]. The relationships between the populations studied here, combined with worldwide populations from the HGDP-CEPH panel[37], were also examined using ancestry graph analyses implemented in TreeMix v.1.12[23]. We also used formal test of f3-satitisitcs[25] to investigate population mixture in the history of the populations studied here, as well as worldwide populations from the HGDP-CEPH panel. Rare f_2 variants (with only two copies of the alternative allele in the minimum-sample-size dataset) and moderately rare $f_{3\text{-}10}$ variants (3-10 copies of the alternative allele in the same dataset) are particularly informative for investigating recent human history[21]. We investigated the sharing pattern of these two types of variant by summing all f_2 variants or any random two alleles of the $f_{3\text{-}10}$ variants shared by pairs of individuals. We plotted the results as a heat map using the image[1] function from the base R package (https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/image.html). Variants were aggregated by pair of individuals using the 'count' function of the plyr package, then arranged in matrix form and colorized using 'colorRampPalette' from the colorspace package (https://cran.r-project.org/web/packages/colorspace/index.html). Runs of homozygosity (ROH), inbreeding coefficient (F) as well as the length of LD-blocks were calculated in PLINK, and finally genome-wide F_{ST} values between isolates and their general populations were calculated with the software 4P[38] using the minimum-sample-size dataset.

Demographic inferences: LD-based[39-41] demographic inference was performed in the NeON R package[27] using the minimum-sample-size dataset; the median and confidence interval were estimated using the 50th, 5th and 95th percentiles of the distribution of long-term *Ne* in each time interval. We used the multiple sequentially Markovian coalescent (MSMC) method[26] to infer demographic changes before 20,000 years ago using four individual sequences from each population. In order to account for some loss of heterozygous sites in the low-depth data, we used a slow mutation rate of 0.8 x 10^{-8} mutations per nucleotide per generation and a longer generation time of 33 years. We then estimated more recent demographic changes (from the present to ~9,000 years ago) using IBDNe[28] with the minimum-sample-size dataset. We used IBDseq[42] to detect IBD segments in sequence data from chromosome 2 in all populations. We then used IBDNe with the default parameters and a minimum IBD segment length of 2 centiMorgan (cM) units. We assumed a generation time of 29 years.

Isolation index: In order to quantify the different isolation levels of different isolates, we developed an index that combines three demographic parameters: (a) *Tdg*, (b) *Ne*, and (c) the level of private isolate ancestry (*M*). We call this estimate the Isolation index (*Isx*). It is defined as:

$$Isx = \frac{\log(Tdg(100 * M)^2)}{\log(Ne)}$$

Both *Tdg* and *Ne* were inferred from the LD-based method using the NeON R package[27]. *M* is difficult to estimate directly from SNP genotype data, so here we estimated the difference of shared ancestral components between an isolate and its general population from ADMIXTURE analysis. We ran ADMIXTURE with only one isolate and it closest general population using K=2. We then estimated the difference in the means of ancestry between the isolate and its general population. The *M* parameter was defined as Delta Ancestry.

***Rxy* analysis:** *Rxy* statistics[29] between each pair of populations (an isolate and its closest general population) for different functional categories were calculated using the matched-sample-size data for missense and LoF variants, including stop gain, splice donor and acceptor variants, using synonymous variants as controls (we did not use intragenic variants as control because of the ascertainment in the ITG which has high-depth exome sequences and low depth for the rest of the genome). We also calculated *Rxy* statistics for variants with CADD scores[43] greater than 10 and 20, using variants with CADD scores less than 5 as controls. The mean and standard deviation for each *Rxy* value were obtained from 100 bootstraps.

***DVxy* analysis:** A new statistic, *DVxy*, was developed to quantify the enrichment of low-frequency functional variants in the isolates using both the matched-sample-size and minimum-sample-size datasets. It calculates the proportion of functional variants in each isolate compared with its general population, correcting for genetic drift at the same time. We calculated *DVxy* specifically for the subset of variants with DAF 2-5% in the isolate, and at least three times lower in its closest general population, or vice-versa. We called these variants "drifted variants" (DV). *DVxy* was calculated for both coding regions and whole genomes.

For coding variants, we defined missense or missense plus LoF variants as functional variants. We counted the number of functional DVs and neutral (intergenic) DVs in each isolate (population *x*) and the corresponding general population (population *y*). The ratio between the fraction of DV variants from the isolated population (corrected by the count of intergenic variants) and the corresponding fraction of DV variants from its general population was defined as the *DVxy* statistic. If *DVxy* is equal to 1, there is no enrichment for the functional DVs in the isolate; less than 1 indicates depletion, and greater than 1 indicates enrichment.

$$DVxy\_coding = \frac{\%DVx\ missense}{\%DVx\ intergenic} \Bigg/ \frac{\%DVy\ missense}{\%DVy\ intergenic}$$

For the whole genome, we used different CADD score cut-offs and bins. We calculated a DV statistic by stratifying the variants according to their CADD scores (0-5, neutral variants; 5-10, mildly deleterious; 10-20, deleterious; and greater than 20, highly deleterious) for each isolate and its closest general population. We finally calculated a ratio of the fraction of DV variants (from each class) between the isolate and its general population, and vice-versa. The following formula shows the *DVxy-wg* calculation for variants with CADD score between *i* (isolate) and *j* (general population).

$$DVxy_{CADD(ij)} = \%\ DVx\ (CADD\ i-j) \Big/ \%DVy(CADD\ i-j)$$

The 95% confidence interval for each calculation was obtained by randomly sampling data from 20 chromosomes 100 times.

***SVxy* analysis:** We further investigated the relaxation of purifying selection in the isolated populations using singleton variants. Here, we also used the minimum-sample-size dataset. Another new statistic, *SVxy*, was developed to measure the ratio of missense vs synonymous singletons per gene in each population, as well as the ratio of the sum of singletons in all genes which have at least one singleton in the pair of the populations (one isolate and one general population). We counted the number of missense singletons and synonymous singletons per gene in each population, and *SVgene* was calculated as:

SVgene = (SV missense count +1)/ (SV synonymous count +1)

SVgene >1 indicates relaxation of purifying selection; *SVgene* = 1 indicates neutrality; and *SVgene* <1 indicates purifying selection.

We then divided the gene list into essential genes[30] and non-essential genes (the rest), and calculated a statistic, G_{SV}, for each population, defined as:

G_{SV} = percentage of essential genes with *SVgene* >1/percentage of non-essential genes with *SVgene* >1

We finally calculated a statistic, *SVxy*, which is the ratio of *SVpop* of each isolate to *SVpop* of its general population. *SVpop* for each isolate and its general population was calculated using all genes which have at least one singleton in the pair of the populations and defined as *SVpop* = Σ (SV missense counts)/Σ(SV synonymous counts).

We used the same annotation as in the variant counts. We calculated a confidence interval for each estimate using bootstrapping of 80% of the genes 100 times.

Correlation analyses: We calculated pair-wise correlation coefficients between the *Isx* values, population-genetic measurements ROH, F, F_{ST}, and number and length of LD blocks, as well as the newly-developed statistics *DVxy* and *SVxy* using the Pearson correlation in R.

Positive selection analyses: We calculated genome-wide pairwise derived allele frequency differences (deltaDAF) for each pair of populations (an isolate and its general population) as described previously[31] using the matched-sample-size dataset. We also carried out PCAdapt analyses[32] for each pair of populations using the whole dataset. Both analyses look for high derived allele frequency variants in the isolates, and will not be affected by sample size. Finally, we ran the singleton density score (SDS) method[33] using the whole UKO and UKG datasets, which have the largest sample sizes for both isolate and its general population, and thus the greatest power for this method.

Data availability:

Amalgamated genotype calls across all populations studied are available through the European Genome/Phenome Archive (EGAD00001002014) with Data Access Agreement described in the Supplementary Information.

References

1. Zeggini, E. Using genetically isolated populations to understand the genomic basis of disease. *Genome Med* **6**, 83 10.1186/s13073-014-0083-5 (2014).
2. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief Funct Genomics* **13**, 371-377 10.1093/bfgp/elu022 (2014).
3. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-464 10.1073/pnas.1322563111 (2014).
4. Pollin, T.I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702-1705 10.1126/science.1161524 (2008).
5. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* **44**, 1326-1329 10.1038/ng.2437 (2012).
6. Tachmazidou, I. *et al.* A rare functional cardioprotective *APOC3* variant has risen in frequency in distinct population isolates. *Nat Commun* **4**, 2872 10.1038/ncomms3872 (2013).
7. Huyghe, J.R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197-201 10.1038/ng.2507 (2013).
8. Li, A.H. *et al.* Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat Genet* **47**, 640-642 10.1038/ng.3270 (2015).
9. Lim, E.T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* **10**, e1004494 10.1371/journal.pgen.1004494 (2014).
10. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190-193 10.1038/nature13425 (2014).
11. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**, 1272-1281 10.1038/ng.3368 (2015).
12. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-298 10.1038/ng.2882 (2014).
13. Panoutsopoulou, K. *et al.* Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun* **5**, 5345 10.1038/ncomms6345 (2014).
14. Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* **21**, 659-665 10.1038/ejhg.2012.229 (2013).
15. Colonna, V. *et al.* Small effective population size and genetic homogeneity in the Val Borbera isolate. *Eur J Hum Genet* **21**, 89-94 10.1038/ejhg.2012.113 (2013).

16. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* **40**, 437-442 10.1038/ng.106 (2008).

17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 10.1038/nature15393 (2015).

18. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 10.1038/nature14962 (2015).

19. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-444 10.1038/ng.3247 (2015).

20. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am J Hum Genet* **83**, 359-372 10.1016/j.ajhg.2008.08.007 (2008).

21. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528 10.1371/journal.pgen.1004528 (2014).

22. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 10.1101/gr.094052.109 (2009).

23. Pickrell, J.K. & Pritchard, J.K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 10.1371/journal.pgen.1002967 (2012).

24. O'Connor, T.D. *et al.* Rare variation facilitates inferences of fine-scale population structure in humans. *Mol Biol Evol* **32**, 653-660 10.1093/molbev/msu326 (2015).

25. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489-494 10.1038/nature08365 (2009).

26. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925 10.1038/ng.3015 (2014).

27. Mezzavilla, M. & Ghirotto, S. *Neon*: An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPs. *J Comput Sci Syst Biol* **8**, 37-44 10.4172/jcsb.1000168 (2015).

28. Browning, S.R. & Browning, B.L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet* **97**, 404-418 10.1016/j.ajhg.2015.07.012 (2015).

29. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet* **47**, 126-131 10.1038/ng.3186 (2015).

30. Blomen, V.A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-1096 10.1126/science.aac7557 (2015).

31. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol* **15**, R88 10.1186/gb-2014-15-6-r88 (2014).

32. Duforet-Frebourg, N., Bazin, E. & Blum, M.B.G. Genome scans for detecting local adaptation using a Bayesian factor model. *Mol Biol Evol* **31**, 2483-2495 10.1093/molbev/msu182 (2014).

33. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 10.1126/science.aag0776 (2016).

34. Zoledziewska, M., Sidore, C. & Chiang, C.W. Height-reducing variants and selection for short stature in Sardinia. *Nat Genet* **47**, 1352-1356 10.1038/ng.3403 (2015).

35. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 10.1086/521987 (2007).

36. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 10.1038/ng1847 (2006).

37. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-1104 10.1126/science.1153717 (2008).

38. Benazzo, A., Panziera, A. & Bertorelle, G. 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol Evol* **5**, 172-175 10.1002/ece3.1261 (2014).

39. Hill, W.G. Estimation of effective population size from data on linkage disequilibrium. *Genetical Res.* **38**, 209-216 (1981).

40. Hayes, B.J., Visscher, P.M., McPartlan, H.C. & Goddard, M.E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Res* **13**, 635-643 10.1101/gr.387103 (2003).

41. Tenesa, A. *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17**, 520-526 10.1101/gr.6023607 (2007).

42. Browning, B.L. & Browning, S.R. Detecting identity by descent and estimating genotype error rates in sequence data. *Am J Hum Genet* **93**, 840-851 10.1016/j.ajhg.2013.09.014 (2013).

43. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 10.1038/ng.2892 (2014).

Author contributions

Y.X., C.T.-S., R.D. and E.Z.: design and supervision of the project. G.D., P.G., A.P., S.R., N.S., D.T. and J.F.W.: population liaison, sampling and DNA provision. N.S., JF.W. and R.D: comments and approval of the manuscript on behalf of the population consortia. Y.X., M.M. and M.H.: statistical method development. M.M., M.H., S.M., V.N., A.G., Q.A., V.C., L.S., C.F., G.R., H.C. and P.R. and J.A.: population-genetic analyses, statistical analyses and data interpretation. Y.C. and A.M.: bioinformatics support. S.M., N.S. and R.D: data processing and QC. Y.X., M.M., M.H., S.M., V.C., C.T-.S. and E.Z.: manuscript drafting. All authors: approval of the final version of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

We thank all study participants for making this work possible. Our work was supported by the Wellcome Trust (098051). We also thank the UK10K Consortium and SISu Consortium for making their data available to this study; detailed acknowledgements for these contributions are included in the Supplementary Information.

Figure legends

Fig. 1. General characteristics and demographic history of isolated and matched general populations. a. Geographical locations of samples. The base map was plotted in R using the mapdata package and circles were added using Photoshop. b. PCA using common variants. c. PCA using low-frequency variants. d. Sharing of rare variants within and between populations. Upper left triangle: f_2 variants; lower right triangle $f_{3\text{-}f10}$ variants. e. Effective population size (Ne) inferred from IBDNe for UKO and UKG during the past 9 KY. f. The lowest Ne inferred by IBDNe for all populations for the past 3KY, plotted as a function of the time at which it occurred.

Fig. 2. Isolation index (*Isx*) and its correlation with other genetic measures. a. Information summarized in *Isx*. b. Example of the correlation between *Isx* and other statistics, here *DVxy-coding*. c. Summary of the correlations between *Isx* and other population-genetic statistics. All the correlation coefficients are high and statistically significant.

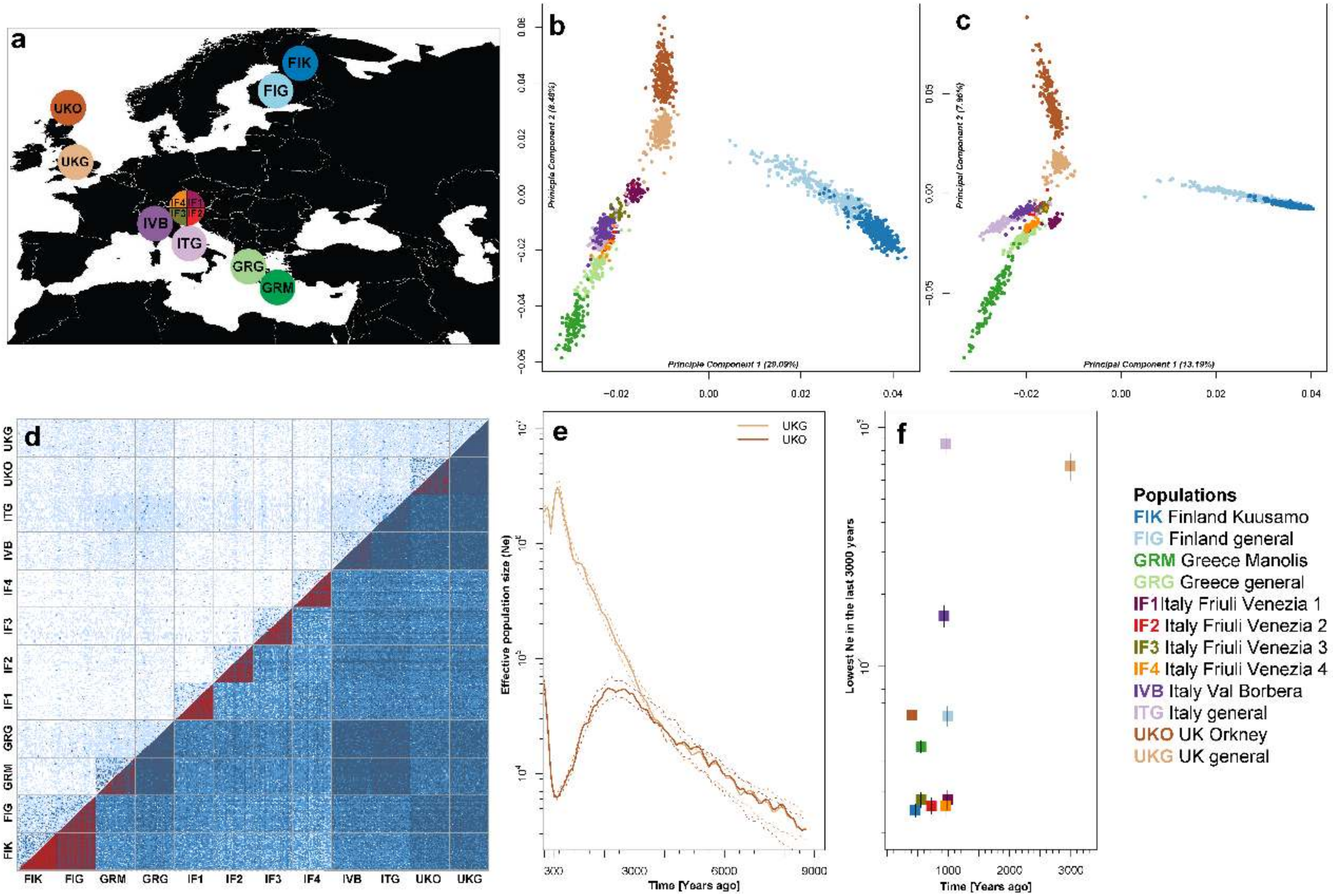
Fig. 3. Purifying selection in the isolates and general populations. a. *Rxy*-missense statistic in each isolate, showing no evidence for increased genetic load in the isolates. The mean and standard deviation for each Rxy value from 100 bootstraps are shown. b. *DVxy-wg* (*DVxy*-whole genome) statistic in isolates and general populations, stratified by CADD score, showing enrichment of highly-functional low-frequency variants. c. *DVxy-coding* statistic in isolates and general populations, showing enrichment of low-frequency missense variants in isolates. d. SVxy-missense statistic in each isolate, showing relaxation of purifying selection in isolates in singletons. The standard errors for both *DVxy* and *SVxy* were calculated by randomly sampling data from 20 chromosomes 100 times. All of these analyses are based on the minimum-sample-size dataset (36 individuals from each population).

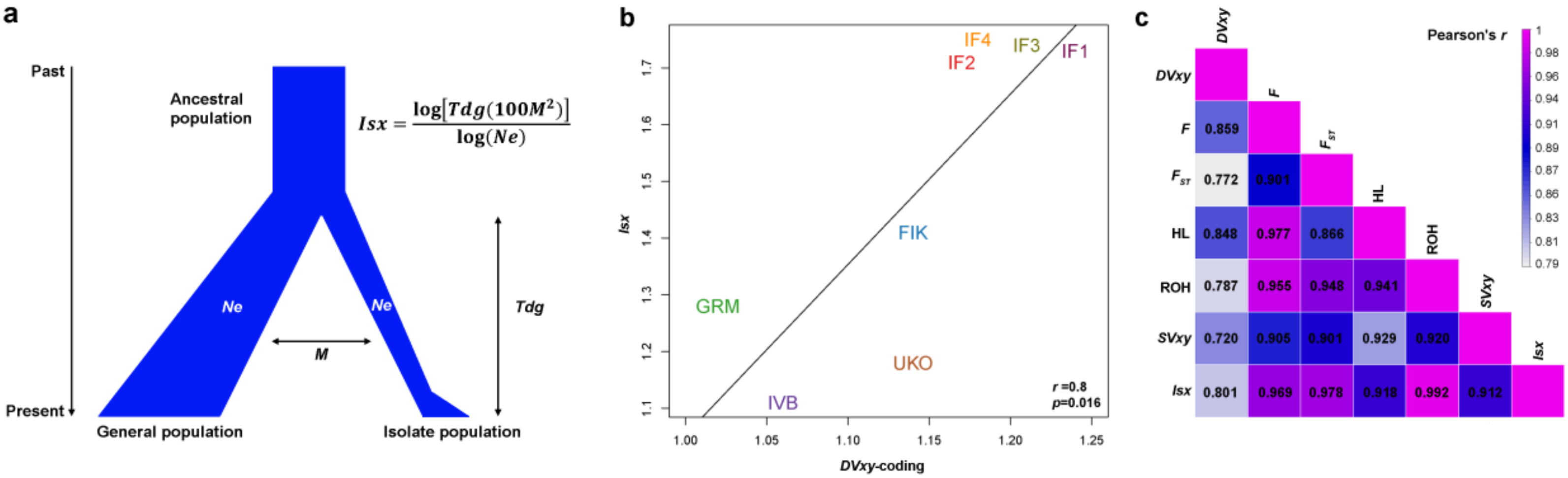
Tables:

Table 1. Summary of variants discovered in this study

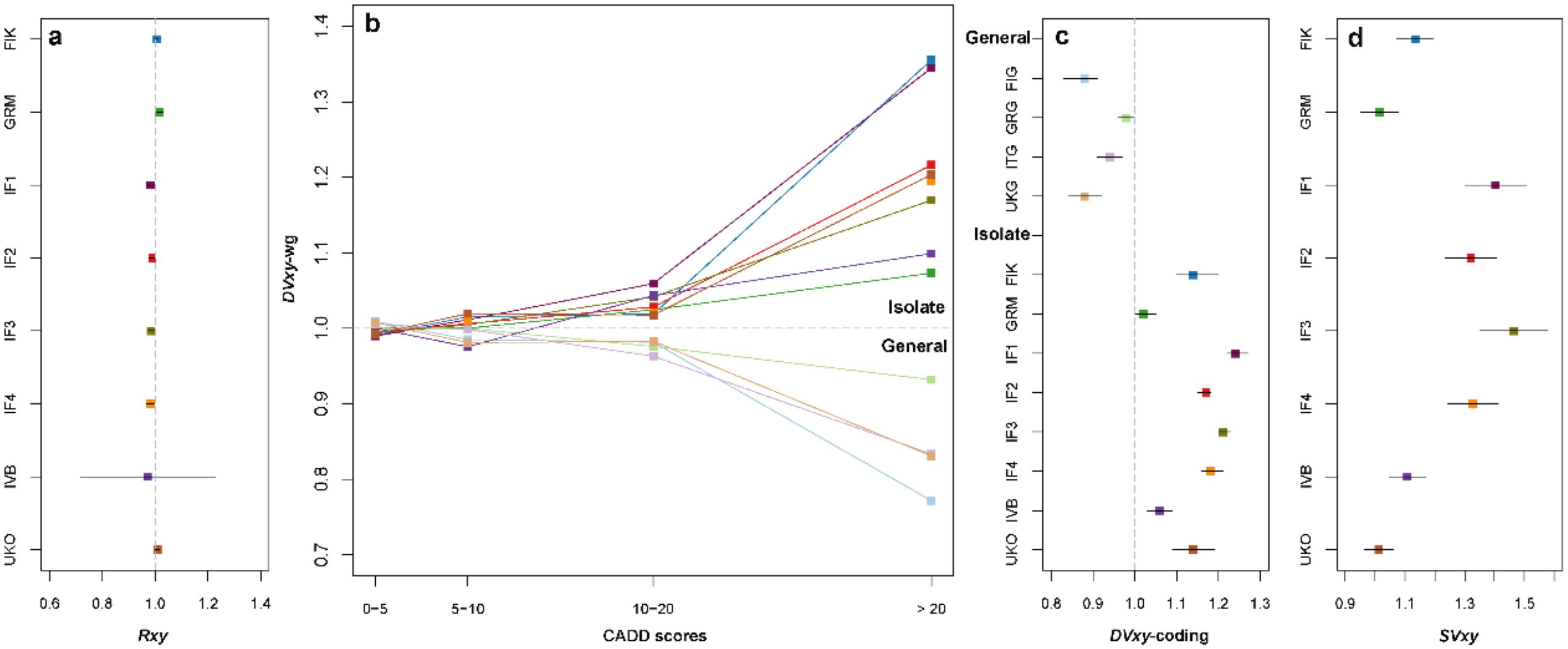
POP	n	average depth	MAF ≤2%		MAF >2-≤5%		MAF >5%		Novel common SNPs in isolate*	Novel common SNPs in isolate**
			total	novel %	total	novel %	total	novel %		
FIK	377	4x	4,066,373	10.90	1,553,076	1.20	6,025,077	0.70	190,527	70,579
FIG	1564	6x	6,548,833	11.80	1,540,915	0.80	6,053,704	0.70	n.a.	n.a.
GRM	249	4x	5,129,513	7.20	1,447,981	1.10	6,111,923	0.80	513,272	49,884
GRG* **	99	10-30x	3,757,110	n.a.	1,321,955	n.a.	5,842,537	n.a.	n.a.	n.a.
IF1	60	4-10x	1,456,881	1.30	1,420,929	1.30	5,890,714	0.80	320,191	119,157
IF2	45	4-10x	1,063,098	1.30	1,554,145	1.00	6,001,568	0.80	273,694	94,496
IF3	47	4-10x	961,059	1.30	1,455,284	1.10	6,068,304	0.80	299,603	107,281
IF4	36	4-10x	1,030,673	1.30	1,124,789	1.10	6,001,625	0.80	308,356	122,254
IVB	222	6x	4,857,767	1.60	1,396,799	0.80	6,112,476	0.80	188,972	30,284
UKO	397	4x	5,963,416	11.70	1,471,782	0.80	6,047,383	0.80	193,300	36,512
Total	3096		12,218,797	10.50	5,503,179	0.70	8,301,524	0.30		

'Novel' variants are those not found in 1000 Genomes Project Phase 3 or UK10K project. *Variants that are common (minor allele frequency, MAF ≥5.6%, alternative allele count ≥ 4) in an isolated population but not common (MAF <1.4%, alternative allele count ≤ 1) in its closest general population. **Variants that are common (MAF ≥5.6%, alternative allele count ≥ 4) in an isolated population but not (MAF <1.4%, alternative allele count ≤ 1) in *any* of the general populations. ***Different variant calling procedure in this population.


a
UKO
UKG
FIK
FIG
IF4
IF1
IF3
IF2
IVB
ITG
GRG
GRM
b
Principle Component 2 (8.48%)
Principle Component 1 (28.09%)
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.02
0.00
0.02
0.04
c
Principal Component 2 (7.96%)
Principal Component 1 (13.19%)
0.05
0.00
-0.05
-0.02
0.00
0.02
0.04
d
UKG
UKO
ITG
IVB
IF4
IF3
IF2
IF1
GRG
GRM
FIG
FIK
FIK
FIG
GRM
GRG
IF1
IF2
IF3
IF4
IVB
ITG
UKO
UKG
e
UKG
UKO
Effective population size (Ne)
Time [Years ago]
300
3000
6000
9000
f
Lowest Ne in the last 3000 years
Time [Years ago]
1000
2000
3000
Populations
FIK Finland Kuusamo
FIG Finland general
GRM Greece Manolis
GRG Greece general
IF1Italy Friuli Venezia 1
IF2 Italy Friuli Venezia 2
IF3 Italy Friuli Venezia 3
IF4 Italy Friuli Venezia 4
IVB Italy Val Borbera
ITG Italy general
UKO UK Orkney
UKG UK general



	DVxy	*F*	F_{ST}	HL	ROH	*SVxy*	*Isx*
DVxy							
F	0.859						
F_{ST}	0.772	0.901					
HL	0.848	0.977	0.866				
ROH	0.787	0.955	0.948	0.941			
SVxy	0.720	0.905	0.901	0.929	0.920		
Isx	0.801	0.969	0.978	0.918	0.992	0.912	


a
FIK
GRM
IF1
IF2
IF3
IF4
IVB
UKO
0.6
0.8
1.0
1.2
1.4
Rxy
b
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
DVxy-wg
Isolate
General
0–5
5–10
10–20
> 20
CADD scores
General
c
FIG
GRG
ITG
UKG
Isolate
FIK
GRM
IF1
IF2
IF3
IF4
IVB
UKO
0.8
0.9
1.0
1.1
1.2
1.3
DVxy-coding
d
FIK
GRM
IF1
IF2
IF3
IF4
IVB
UKO
0.9
1.1
1.3
1.5
SVxy