EJHG Open

European Journal of Human Genetics (2017) 25, 477–484
Official journal of The European Society of Human Genetics
www.nature.com/ejhg

ARTICLE

Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom

Himanshu Chheda[1,11], Priit Palta[1,2,11], Matti Pirinen[1], Shane McCarthy[3], Klaudia Walter[3], Seppo Koskinen[4], Veikko Salomaa[4], Mark Daly[5,6,7], Richard Durbin[3], Aarno Palotie[1,5,6,8], Tero Aittokallio[1,9] and Samuli Ripatti*[,1,3,10] for the Sequencing Initiative Suomi (SISu) Project

Isolated populations with enrichment of variants due to recent population bottlenecks provide a powerful resource for identifying disease-associated genetic variants and genes. As a model of an isolate population, we sequenced the genomes of 1463 Finnish individuals as part of the Sequencing Initiative Suomi (SISu) Project. We compared the genomic profiles of the 1463 Finns to a sample of 1463 British individuals that were sequenced in parallel as part of the UK10K Project. Whereas there were no major differences in the allele frequency of common variants, a significant depletion of variants in the rare frequency spectrum was observed in Finns when comparing the two populations. On the other hand, we observed >2.1 million variants that were twice as frequent among Finns compared with Britons and 800 000 variants that were more than 10 times more frequent in Finns. Furthermore, in Finns we observed a relative proportional enrichment of variants in the minor allele frequency range between 2 and 5% ($P<2.2\times10^{-16}$). When stratified by their functional annotations, loss-of-function variants showed the highest proportional enrichment in Finns ($P=0.0291$). In the non-coding part of the genome, variants in conserved regions ($P=0.002$) and promoters ($P=0.01$) were also significantly enriched in the Finnish samples. These functional categories represent the highest *a priori* power for downstream association studies of rare variants using population isolates.
European Journal of Human Genetics (2017) **25,** 477–484; doi:10.1038/ejhg.2016.205; published online 1 February 2017

INTRODUCTION

Population isolates have not only provided insights into population diversity and history, but are also an exciting opportunity to identify rare and low-frequency variants associated with complex diseases.[1–4] Regardless of whether looking across the whole genome or focusing on genetic variation in the coding regions, these studies have consistently observed the highest enrichment in the variation that predictably disrupts protein coding genes.

Within coding regions, variant alleles that have high penetrance whilst predisposing to disease are likely to be deleterious and therefore kept at low frequencies by purifying selection in larger outbred populations.[5–7] Isolated populations resulting from recent bottlenecks have a substantial reduction in rare neutral variation and also many functional and even deleterious variants present at relatively higher frequencies because of increased drift and reduced selective pressure. Hence, recent isolates can be used to study causal variants that are rare in other populations in association with complex diseases.[1–4]

Finland is a well-known example of an isolated population where multiple historical bottlenecks resulting from consecutive founder effects have shaped the gene pool of current-day Finns.[8] Previous studies suggest the latest historical migration into Finland ~4000 years ago.[9] Owing to lack of evidence of major migratory movements, it has been suggested that there were small but significant migrating groups of people.

Settlements resulting from the latter migratory movements mainly occurred along the south–east cost of Finland. Further, due to geopolitical reasons there have been additional major migratory movements within Finland in the 16th century in the eastern and northern parts of Finland. These settlements, initially founded by a small number of people, have grown in size over time leading to secondary population bottlenecks. An extreme example of the latter is Kuusamo, a county in the northeast part of Finland.[10,11] Historical records show that in 1718, there were 165 houses consisting of 615 individuals belonging to 39 families. Rapid population growth leading to a present day population of >15 000 individuals further increased the allelic drift in this sub-isolate. Consequences of these historical events have led to reduced genetic variation and higher overall linkage disequilibrium levels in Finland as compared with the outbred populations.[10,11]

During the last 1000 years, the Finnish population size has grown more than two orders of magnitude – from around 50 000 individuals to more than 5 million individuals. Furthermore, the most rapid

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland; [2]Estonian Genome Center, University of Tartu, Tartu, Estonia; [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK; [4]Department of Health, National Institute for Health and Welfare, Helsinki, Finland; [5]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; [6]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA; [7]Department of Genetics, Harvard Medical School, Boston, MA, USA; [8]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; [9]Department of Mathematics and Statistics, University of Turku, Turku, Finland; [10]Public Health, Clinicum, University of Helsinki, Helsinki, Finland
*Correspondence: Professor S Ripatti, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Tukholmankatu 8, Helsinki 00290, Finland. Tel: +358 40 567 0826; Fax: +358 20 610 8272; E-mail: samuli.ripatti@helsinki.fi or samuli.ripatti@fimm.fi
[11]These authors contributed equally to this work.
Received 15 June 2016; revised 5 December 2016; accepted 14 December 2016; published online 1 February 2017

growth has happened during the last 10 generations (~250 years), with population size growing from 500 000 to 5.4 million individuals. Combined with the historical bottleneck effect, these events have caused a massive departure from population genetic equilibrium whilst 'shifting' the proportion and frequency of many initially rare variants.

Such deviations have led to an increase in the prevalence of some monogenic Mendelian disorders in Finland as compared with the other parts of the world and are referred to as the Finnish disease heritage[12] (FDH). Pronounced effects of the bottleneck have also been observed for complex diseases and disorders. For instance, schizophrenia is prevalent almost three times in northeastern sub-isolates as compared with rest of Finland.[13] Similarly, protective effects of enriched variants have also been observed as exemplified by variants in the LPA gene that protect against risk of cardiovascular diseases.[1]

However, the dynamics and properties of this genetic 'enrichment' are poorly understood on the genome scale, particularly outside the protein coding regions. We set out to provide a more comprehensive view of this enrichment and the other bottleneck effects in Finland by comparing whole-genome sequencing data in Finnish and British samples. In this study, we show how the historical bottlenecks have affected the genetic landscape of Finns and the frequency profile of variants across the entire genome. Whole-genome sequencing data gave us a unique opportunity to determine the enrichment of variants across both coding as well as non-coding regions of the human genome.

METHODS

Sample selection

We sequenced the whole genomes of unrelated 1463 Finns at low coverage (~4.6×). These samples belonged to the FINRISK[14] and H2000 cohorts. The FINRISK study comprises samples of the working-age population, to study the risk factors associated with chronic diseases across Finland and is carried out every 5 years. The H2000 is a population-based national survey aimed at studying the prevalence and determinants of important health problems amongst the working-age and the aged population (http://www.terveys2000.fi/julkaisut/baseline.pdf). Amongst these, 856 individuals have low HDL and 691 individuals have been diagnosed with psychosis. Further, 371 individuals belong to a sub-isolate within Finland, the Kuusamo region. Due to known genetic differences, for the comparison between Britons and the Finns, we restricted the analyses only to those Finnish individuals that are not from Kuusamo. To study the effects of a bottleneck within a bottleneck, 371 samples from non-Kuusamo Finns were further used for comparison against 371 samples from Kuusamo. All study participants gave their written informed consent to the study of origin.

Whole-genome sequencing and variant discovery

Low read-depth whole-genome sequencing was performed at the Wellcome Trust Sanger Institute (WTSI). Joint variant calling of the raw binary sequence alignment map (BAM files) along with the UK10K samples was performed as part of the Haplotype Reference Consortium (HRC).[15] The genotypes were further refined by re-phasing using SHAPEIT3 algorithm.[16] As a part of the joint calling quality control, only those sites that have a minor allele count of at least 5 copies in the entire data set (32 611 samples) went through additional filtering. Hence we have restricted the analyses to those variants with minor allele count ≥ 5. The BAM files have been submitted to the European Genome-phenome Archive (EGA). To minimize the batch effects, we performed these analyses on only those British samples from UK10K (1463 samples from 3781 samples) that were also sequenced at the WTSI. We have only included autosomal single-nucleotide variants for these analyses.

To determine the quality of the data, we compared the Finnish whole genome sequencing data with Illumina PsychArray genotypes for 629 individuals. We performed a two-step quality control for the chip genotyped data. The calls were first made using GenCall. We excluded the samples for gender mismatch and duplicates. Additional quality control steps were performed based on zCall data. All the samples with call rate <98% and heterozygosity >3s.d. were removed. Further, we performed SNP-wise QC to exclude variants with call rates <95% and Hardy–Weinberg *P*-value $<10^{-6}$.

The filtered chip data were used for concordance analyses of the low pass whole-genome Finnish sequencing data using the GATK GenotypeConcordance module. From this comparison, we estimate that for variant sites with minor allele frequency >5% there is a non-reference sensitivity of 99.1% of variants with a low non-reference discrepancy of <0.1%. For the variant sites with minor allele frequency between 2 and 5%, we observe a non-reference sensitivity of 97.3% and non-reference discrepancy of 4.3%. For the variant sites between minor allele frequencies (MAF) 0.5–2%, the non-reference sensitivity is 93.9% and the non-reference discrepancy is 11.9%. Below MAF 0.5%, the number of variants were too low to calculate the genotype concordance.

Annotations

The various functional categories were obtained as follows:

(a) Coding sequence, promoters, untranslated region annotations were obtained from UCSC Genome Browser[17] using the Gencode v19 gene models.[18]

(b) Coding variants were further stratified using the Variant Effect Predictor[19] into loss-of-function variants, missense variants and synonymous variants. Polyphen[20] predictions were used to classify missense damaging variants.

(c) Dnase1 hypersensitivity sites (DHS) were obtained from Trynka *et al.*[21] We merged the coordinates for all cell types into one category.

(d) Conserved regions in mammals were obtained from Linblad-Toh *et al.*[22] These were post processed by Ward & Kellis.[23]

(e) FANTOM5 enhancer coordinates were obtained from Andersson *et al.*[24]) Super-enhancers were obtained from Hnisz *et al.*[25] The genomic coordinates were merged over all cell types.

(g) Transcription factor binding sites were obtained from Encode project.[26]

Enrichment analysis

We calculated the enrichment for each category beyond the baseline enrichment observed (enrichment calculated using all variants in Finns and Britons), assuming the following model.

Consider a category of variants in which we have observed F variants in the first population (eg, Finnish) and B variants in the second population (eg, British) and let M = F+B. Let s be the proportion of variants from the first population, and u the ratio of the numbers of variants in the first population to that in the second population. According to the binomial distribution, our point estimate for s is $\hat{s} = F/M$ and has variance approximately $\hat{s}(1-\hat{s})/M$. It follows that a point estimate for u is $\hat{u} = F/B$, and the variance of $\log(\hat{u})$ is $1/(M\hat{s}(1-\hat{s}))$ by the Delta method. This allows us to estimate 95% confidence intervals for $\hat{u}$.

Suppose that we are comparing two categories of variants, and have observed F_1 and B_1 variants in category 1 and F_2 and B_2 variants in category 2. To test whether u_1 is different from u_2, we compute $\log(\hat{u}_1) - \log(\hat{u}_2)$. Under the null hypothesis of no difference, this statistic has mean 0 and variance approximately $(1/M_1+1/M_2)/(\hat{s}(1-\hat{s}))$, where $\hat{s} = (F_1+F_2)/(M_1+M_2)$, which we use to derive a *P*-value. Note that the standard proportion test between s_1 and s_2 gives essentially the same *P*-value.

We calculate the statistical power gained for the enriched variants. For quantitative traits, the standard linear model for genotype-phenotype association test statistic follows a chi-squared distribution with one degree of freedom and non-centrality parameter (NCP) of $2Nf(1-f)b^2$, where *N* is the sample size, *f* is MAF and *b* is the (additive) effect size of the minor allele measured on the scale of the phenotype. For case–control analysis, the corresponding NCP is $2Nf(1-f)r(1-r)b^2$, where *N* is the total sample size (cases+controls), *r* is the proportion of cases among all samples and b is the additive effect of the minor allele on the log-odds of the disease.[27] Both NCPs are derived assuming that the variant explains only a little of the phenotypic variation at the population level,

which is a reasonable assumption when the minor allele is rare and/or the allelic effect size is small.

Thus, for both quantitative and binary traits, the sample size N_2 required in population 2 for the same power to detect an association as in population 1 is $N_2 = N_1\ f_1(1-f_1)/(f_2(1-f_2))$ assuming equal effect size and case proportion across the studies.

RESULTS

Overall frequency distribution of genome-wide level variation

As a part of the SISu project, we sequenced the genomes of 1463 Finnish samples at low read-depth (average 4.6×) sampled across Finland. We compared these profiles to a sample of 1463 British individuals sequenced at average depth 7× as a part of the UK10K Consortium.[28] We restricted the analyses to 1463 individuals to minimize the artefacts arising from comparing data from different sequencing centers. Further, to reduce potential batch effects, these data sets were jointly processed as part of the Haplotype Reference Consortium.[15] After stringent quality control steps, we compared the MAFs of 10 457 802 and 11 172 232 single-nucleotide variants (SNVs) identified with minor allele count 5 or greater in 1463 Finns and in the same number of Britons, respectively (Table 1).

As a direct result of the bottleneck effect, we observed that Finns have significantly fewer rare variants (MAF < 0.5%) compared with Britons (Figure 1). On the other hand, in Finns, we determined proportionally small but significant enrichment of low-frequency variants (MAF range between 2 and 5%, binomial $P < 2.2 \times 10^{-16}$). The latter is also a direct effect of the historical bottleneck, followed by population growth. And as expected, we observed no differences in the number of common (MAF > 5%) variants (Figure 1).

For each frequency range, we also calculated the percentage of variants shared between both population samples. As anticipated, the number of variants observed as rare in Britons (MAF < 0.5%) and also found as polymorphic in Finns was considerably lower than the opposite: only 54.7% of variants with MAF < 0.5% in Britons were polymorphic in Finns while 72% of variants with MAF < 0.5% in Finns were also polymorphic in Britons (Figure 2). However, for the MAF range of 0.5–5% the opposite was true: a lower proportion of variants seen in Finns were also polymorphic in Britons (eg, for 0.5–2% range, 84.9 and 94.3% of variants are shared, respectively). For common variants (MAF > 5%), essentially all (99.9%) were observed to be shared in both directions (Table 1 and Figure 2).

Enrichment of variants across functional categories

We also calculated the relative enrichment of Finnish SNVs across various functional categories shown to be relevant in different phenotypic traits including disease.[29] For each of these categories, we compared its distribution profile with that of the 'expected' whole genome baseline distribution in Finns (Figure 1a). Although there were several small deviations from the expected baseline in almost all functional categories, the greatest differences were consistently observed in the MAF range of 2–5% (Supplementary Figures 1–8).

In accordance with the latter observation, we compared the enrichment of different functional categories for MAF range 2–5% (Figure 3a). Across studied functional categories, the coding regions showed the highest enrichment in Finns (Figure 3a). More specifically, we observed >1.3-fold enrichment of loss-of-function ($P = 0.0291$) variants and >1.1-fold enrichment of missense ($P = 0.0197$) variants (Figure 3b), similarly as was demonstrated previously in Finns by exome sequencing.[1] Furthermore, we observed consistent enrichment of rare and low-frequency (MAF ≤ 5%) missense damaging variants (Figure 3b).

As observed for the low-frequency variants (MAF 2–5%) in the coding regions, we found enrichment in the non-coding regions as well. In the non-coding parts of the genome, the promoter regions showed the largest enrichment compared with the expected baseline ($P = 0.012$, Supplementary Figure 3), followed by the conserved non-coding regions of the human genome ($P = 0.01$, Figure 3c). Although the Fantom5 enhancer regions showed proportional enrichment, it was not significant compared with the expected baseline (Figure 3a; Supplementary Figure 4). The other functional categories followed the baseline enrichment (Figure 3a). We also observed that, although enriched when compared with the Britons, the DHS and the super-enhancer elements are only marginally depleted beyond the expected bottleneck effects ($P_{DHS} = 0.04$ and $P_{super\text{-}enhancers} = 0.007$, Supplementary Figures 5 and 7).

MAF-enrichment of variants and effect on statistical power

We observed that 20.16% of all variants in Finns have minor allele frequencies elevated at least twofold. Furthermore, 1.36% of these variants were enriched ≥ 50 fold. For the proportionally enriched functional categories, we calculated the number of variants with elevated frequencies in Finns as compared to Britons (Table 2) and observed even higher MAF-enrichment for many of these categories. Missense damaging variants showed the highest enrichment with 37.98% variants showing minor allele frequencies at least twice as high as observed in the Britons. 29.71% of the loss-of-function variants showed at least twofold MAF-enrichment compared with the British sample.

We also performed the same analyses in Britons. Across all categories, for variants that are enriched at the most 10-fold, the Britons consistently show much higher number of variants across all enriched functional categories. However, in the loss-of-function and missense damaging categories beyond 10-fold enrichment, the Finns have larger proportion of variants enriched. Interestingly, beyond 50-fold enrichment, the Finns have a relatively larger proportion of variants enriched across all functional categories (Table 2).

We extended these analyses to compare the enrichment in known GWAS loci[30] and Clinvar variants.[31] Similar to the above results, the Britons have a higher proportion of GWAS for variants that are enriched <50-fold (Table 2). However, the Finnish have a larger proportion of known associated loci for more than 50-fold enriched variants. Further, in the Finnish sequencing data, we observed 16

Table 1 Summary of SNVs studied in Finnish and British samples

Minor allele count/frequency	*No of SNVs in Finns*	*No of SNVs in Britons*	*% variants present in Finns also shared by Britons*	*% variants in Britons also shared by Finns*
5 copies–0.5%	1 629 869	2 313 870	72.6	59.9
0.5–2%	2 020 773	2 119 423	84.9	94.3
2–5%	1 388 186	1 325 135	95.2	99.3
>5%	5 418 973	5 413 803	99.99	99.99

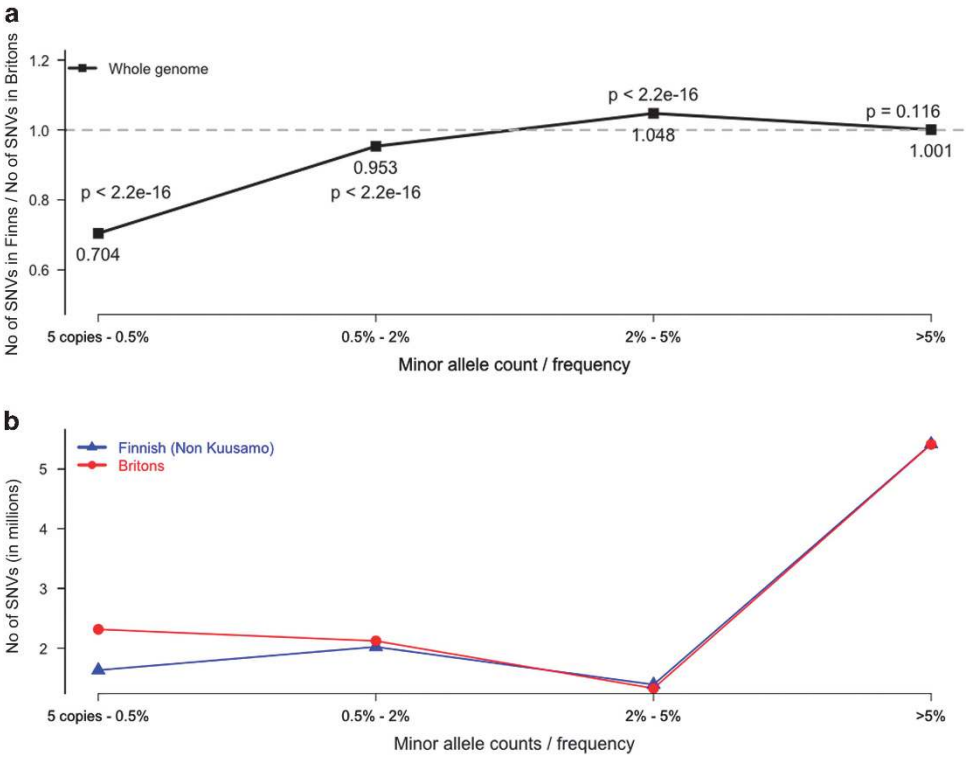

Figure 1 (**a**) Allele frequency spectrum of variants across the whole genome in Finns compared with the Britons. The black line represents the ratio of the number of variants observed in Finns to those in Britons. (**b**) The number of variants seen in each population across the genome in different MAF bins. The lines in blue and red represent the number of variants for each bin observed in Finns and Britons, respectively.

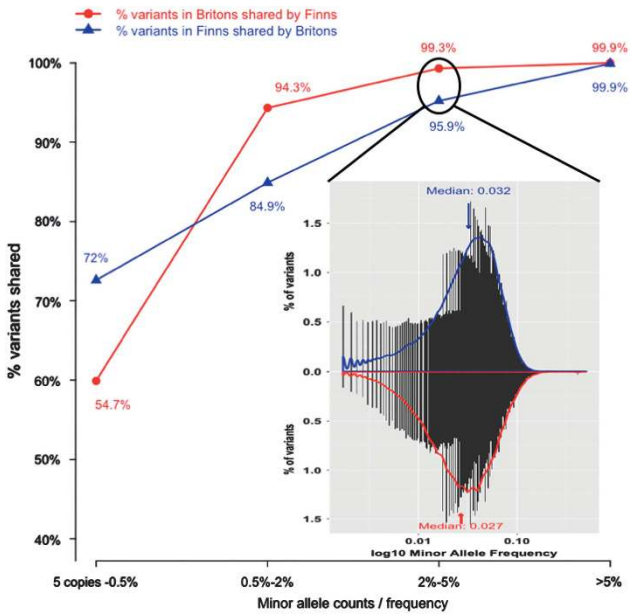

Figure 2 Variants shared between the two populations. The percentage of variants that are shared between the Finns and the Britons across different allele frequency bins. The histograms represent the allele frequencies of the shared variants in the other population for the MAF bin 2–5%.

variants associated with FDH. Most of these were enriched at least fivefold. As a specific example, a variant in the AGA gene (c.488G>C; $MAF_{Finns} = 0.0096$; $MAF_{Britons} = 0$) is enriched 28-fold in Finns and is associated with aspartylglucosaminuria (OMIM #208400).

This enrichment of minor allele frequencies for certain variants boosts the statistical power to detect possible associations with traits and diseases. To quantify this gain in power, we calculated the number of samples required to detect association with high probability in Finns as compared with the Britons (Figure 4a). For variants that are twice as common among Finns to Britons, only half the number of individuals would be required to detect the associations in Finnish samples. Further, for variants with minor allele frequencies enriched 5x and 10x times, only 20 and 10% of the samples respectively are required to detect associations. These analyses indicate the gain in power for studying association analyses in isolated populations such as the Finnish population.

To elucidate the power gained for variants enriched 10-fold in Finns, we have simulated an additive genetic model for both quantitative trait association and case–control association analyses (Figures 4b and c). For variants with MAF 0.1% in Britons, Finns have 80% statistical power to detect associations at genome-wide significance ($\alpha = 5 \times 10^{-8}$) with beta regression coefficients or 'beta' of ~1 s.d. (Figure 4b). Similarly, for the case–control scenario, Finns have 80% statistical power to detect association with odds ratio of ~2.5 with 5000 cases and 5000 controls (Figure 4c).

The increase in statistical power can be further exemplified by the missense variant *PCSK9*-R46L that is known to be associated with low density lipoprotein (rs11591147; $MAF_{Finns} = 0.03862$; $MAF_{Britons} = 0.02016$; $\beta = -0.47$). This variant is enriched 1.92 times in Finns. For this variant, we achieve 80% statistical power to detect an association at genome-wide levels of significance with 2415 Finns. However, with the same sample size in the Britons, we have only 19% power to detect the association. Similarly, for the splice site variant in the LPA gene (c.4974-2A>G), which is a protective variant against coronary heart disease ($MAF_{Finns} = 0.03213$; $MAF_{Britons} = 0.003076$;

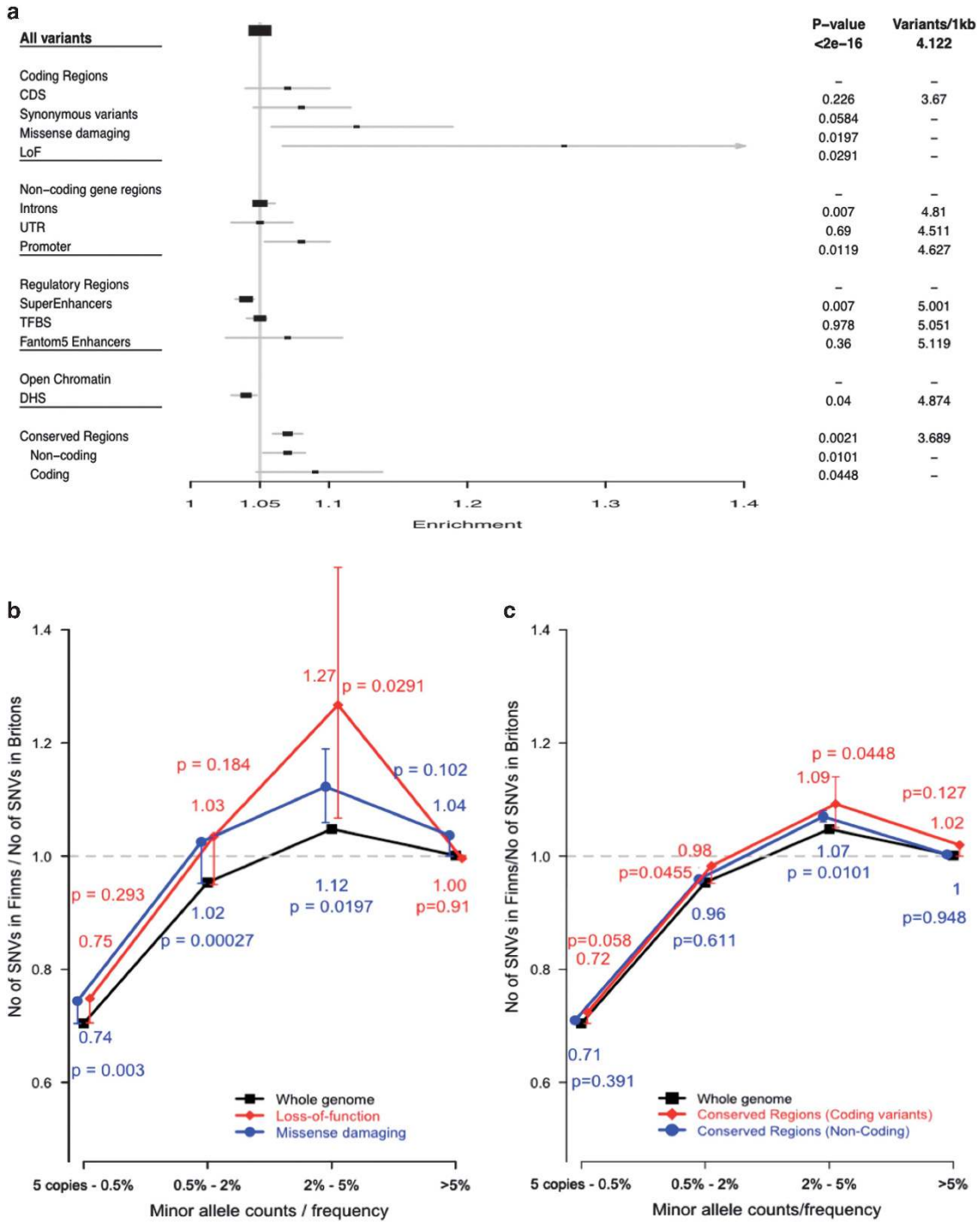

Figure 3 Enrichment of variants across various categories. (**a**) Forest plot showing the enrichment across various functional categories for the variants in the minor allele frequency range 2–5%, where we observe consistent enrichment across most categories. The sizes of the boxes correspond to the size of each category and the black horizontal lines represent the 95% confidence intervals. Proportional enrichment is calculated compared with Britons. (**b**) Proportional enrichment of LoFs in Finns compared with Britons. The red line represents the ratio of the number of LoF variants in Finns compared to Britons. The black line shows the baseline enrichment observed across the whole genome. (**c**) Proportional enrichment of the number of variants in the conserved regions in Finns compared with Britons. The red line represents the variants common between conserved regions and the coding regions. The blue line represents the variants in the conserved regions but not in the coding regions. The black line shows the baseline enrichment observed across the whole genome.

OR = 0.84), 36 200 cases and 50 000 controls is required to achieve 80% power at genome-wide significance level in Finns. Using the same number of cases and controls in the Britons, there is 0.05% power to detect the association. Furthermore, the gain in statistical power can help to detect enriched genetic variants with modest effects associated with diseases that are present at a higher prevalence in Finland, as

Table 2 MAF-enrichment of variants in Finns and Britons

Enrichment	*Genome-wide*	*LoF*	*Missense-damaging*	*Conserved regions-coding*	*Conserved regions-non-coding*	*Promoters*	*Functional variants*
Finns							
2–5x	722 587 (6.9%)	177 (7.8%)	1704 (9.9%)	3109 (8.7%)	22 823 (7.7%)	9179 (7.2%)	421;220;1
5–10x	561 243 (5.4%)	225 (9.9%)	2083 (12.2%)	3013 (8.4%)	19 385 (6.5%)	7415 (5.8%)	73;100;5
10–50x	682 275 (6.5%)	233 (10.3%)	2360 (13.8%)	3643 (10.2%)	23 207 (7.8%)	9129 (7.2%)	83;125;10
≥50x	142 354 (1.4%)	38 (1.7%)	366 (2.1%)	655 (1.8%)	4584 (1.5%)	1888 (1.5%)	29;23;0
Britons							
2–5x	935 392 (8.4%)	203 (8.6%)	1839 (10%)	3483 (9.1%)	28 862 (9%)	11 242 (8.2%)	725;305;0
5–10x	1 319 454 (11.8%)	405 (17.1%)	3991 (21.7%)	6336 (16.5%)	44 903 (14%)	17 498 (12.8%)	135;375;0
10–50x	893 522 (8%)	235 (9.9%)	2447 (13.3%)	4005 (10.4%)	29 368 (9.2%)	11 557 (8.4%)	94;331;0
≥50x	25 552 (0.2%)	12 (0.5%)	58 (0.3%)	125 (0.3%)	818 (0.3%)	382 (0.3%)	18;4;0

This table summarizes the number of variants that are MAF-enriched in both populations across the whole genome and the categories in which there is a significant enrichment. The percentages refer to the proportion of enriched variants in that particular category. The first column describes the fold-change of minor allele frequency between the two population samples. The last column describes the number of variants enriched in the GWAS Catalogue, ClinVar and FDH mutations, respectively.

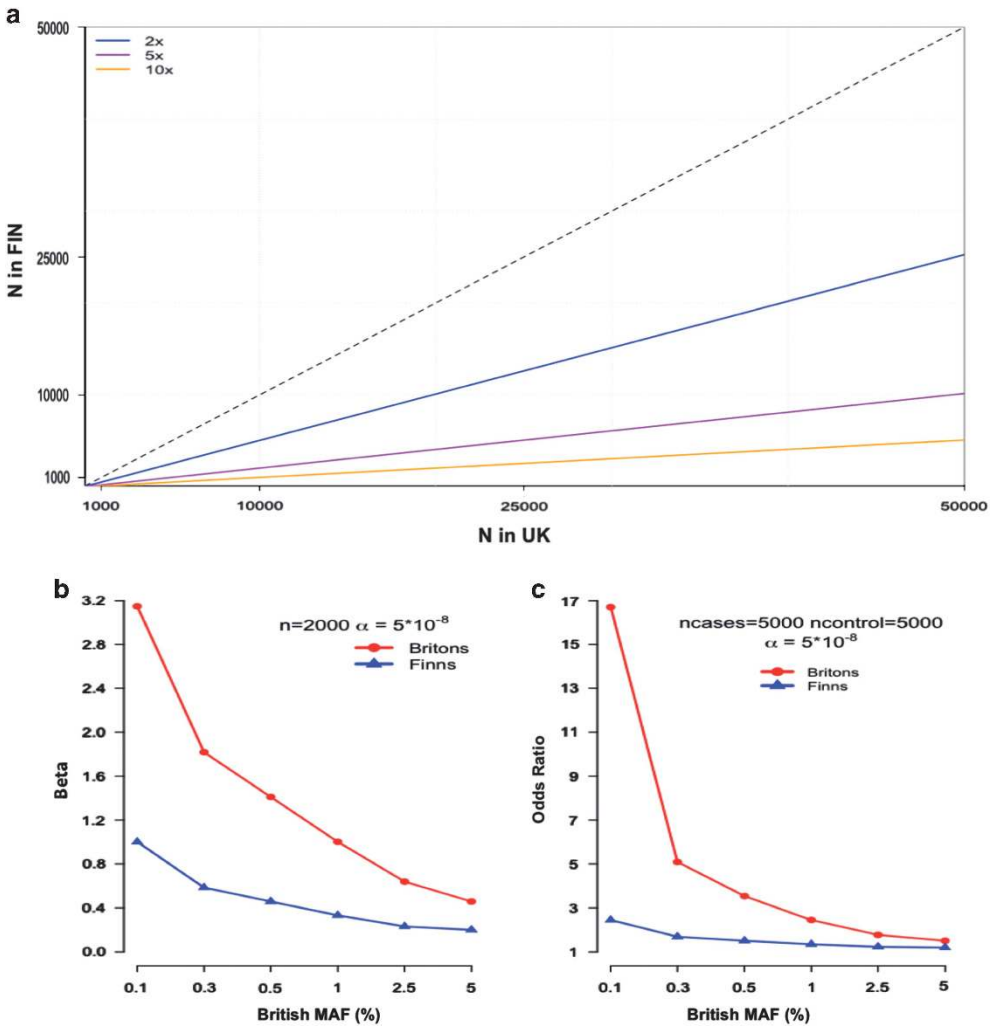

Figure 4 Statistical power gained due to enrichment of a variant in Finns. (**a**) Plot showing the number of Finnish samples required to detect association if a variant is enriched twofold (blue), fivefold (purple) and 10-fold (yellow) in Finns as compared with Britons. (**b**) Regression coefficient (beta) desired to achieve a statistical power of 80% at genome-wide significance level as a function of minor allele frequency for a quantitative trait for variants enriched 10-fold in Finns. The red line indicates the betas in Britons and the blue line indicates the betas in Finns. (**c**) Odds ratio desired to achieve a statistical power of 80% at genome-wide significance level as a function of minor allele frequency for a case–control analysis for variants enriched 10-fold in Finns. The red line indicates the odds ratio in Britons and the blue line indicates the odds ratio in Finns.

exemplified by a variant located in the intron of the *RADIL* gene (c.536-18508 T > A) and associated with intracranial aneurysms (rs150927513; MAF_{Finns}: 0.0591; $MAF_{Britons}$ = 0.0021; RR = 1.59).[32]

Sub-isolate of an isolated population

Amongst the sequenced Finnish individuals, 371 belonged to the Kuusamo sub-isolate within Finland. When comparing the SNV frequency profiles of these individuals against the same number of randomly selected non-Kuusamo Finns, we observed a significant reduction in the number of rare variants (Supplementary Figure 9). Although there was no overall enrichment of low-frequency variants, when looking at the variants stratified by their functional categories, we found a significant enrichment of LoF variants in the MAF 0.5–2% frequency range ($P = 0.0272$; Supplementary Figure 10).

DISCUSSION

Studying relatively recently bottlenecked and isolated populations or sub-isolates provides an excellent opportunity to discover disease-associated genes, as some of the underlying (and initially rare) variants can reach much higher frequency after the population bottleneck. We studied this bottleneck effect and subsequent enrichment of variants in Finnish samples by comparing them to outbred British samples. We demonstrated how the historical bottlenecks have affected the genetic landscape of Finns and the frequency profile of variants across the entire genome.

As expected, we observed no major differences in the common variant frequency spectrum – as most variants with MAF > 5% probably segregated already tens of thousands of years ago, they are known to be relatively equally distributed in populations that separated more recently.[33,34] On the other hand, there was a significant depletion of variants in the rare frequency spectrum in Finns. Also, as an additional hallmark of the population bottleneck, a significant enrichment of low-frequency variants was observed (Figure 1). For most functional variants we observe an enrichment beyond the expected baseline showing that bottleneck population have a higher likelihood of accumulating deleterious and disease-associated mutations. To test the robustness of enrichment of low-frequency variants, we changed the minor allele frequency bins for the whole-genome analysis. We observed consistency in the enrichment of low-frequency variants (MAF 1–5%; Supplementary Figure 11). This phenomenon also explains the high prevalence of several monogenic Mendelian disorders, so-called 'FDH', caused by genetic disease variants found at much higher frequencies in Finland than in the rest of the Europe.[12]

We observed that within the frequency range of MAF 0.5–2%, only a subset (84.9%) of the variants in Finnish samples is also seen in the British samples (Figure 2). For the common variants, in contrast, most variants (99.9%) were shared between the two populations (Figure 2). These findings are similar to the patterns observed in the Icelandic[3] and the Sardinian populations.[2] Finns also show a similar enrichment of LoF variants and missense variants as seen in the Icelandic populations. However, the enrichment observed in the Icelandic population was found in the lower minor allele frequency range as opposed to the Finnish sample, possibly due to the differences in the historical bottleneck 'width', time since bottleneck (the Icelandic bottleneck was more recent than the Finnish), and the subsequent population growth rate. As such, this enrichment can provide a boost in statistical power when studying health-related phenotype traits affected by these enriched variants.

Other studies have recently demonstrated that functional categories such as conserved regions and Fantom5 enhancers contribute disproportionately more to the heritability of complex diseases, suggesting that in addition to coding regions also regulatory regions are enriched for trait and disease-associated variation.[29,35,36] Here, we used 12 functional annotations to determine if variants in any of these categories are enriched beyond the baseline distribution of variants (and bottleneck effect) in Finns. We observed an enrichment across most functional categories in the low-frequency bin (MAF 2–5%). As reported previously,[1] we observed a significant enrichment of low-frequency LoF and missense variants in Finns (Figure 3b). In addition to the enrichment of coding variants, however, also non-coding conserved regions and non-coding genic regions such as intron and promoter regions showed enrichment beyond the baseline bottleneck effect (Figure 3c). This enrichment likely appears due to selection against these variants in non-coding conserved regions and non-coding genic regions in outbred European populations. Furthermore, we see depletion for the super-enhancer regions and the DHS elements. This suggests that functionally, super-enhancers may be actually less active than regular enhancers, as was also proposed previously.[29,37]

Previous studies have shown the utility of bottleneck populations to identify variants with elevated frequencies associated with diseases and phenotypes.[1,2,13] Our findings show that across the genome, ~20% of all variants present in Finns have enrichment at least twice as observed in the Britons (Table 2). The percentage of variants with at least 2× enrichment further increases for loss-of-function variants and missense damaging variants (29.71 and 37.98% respectively). Our power calculation simulations show that by testing for associations with these variants, the number of samples required to achieve significant detections are much lower (Figure 4a).

This power gain gives advantages particularly in identifying (i) rare variants with small/moderate effects, (ii) diseases that are not very common and large collection of cases-controls cannot be collected and (iii) investigation of quantitative phenotypes not measured in existing large biobanks. Examples of these include variants in *AGA*, *PCSK9*, *LPA* and *RADIL* genes. Sequencing studies combined with imputation of these enriched variants in large-scale Finnish population-based cohorts with rich phenotype data and leveraging on the national health registries data from Finland will likely have great potential to help identify similar novel genetic associations for complex disorders.

Although we tried to eliminate all possible sources of biases and other technical limitations by jointly processing our data sets, our results might be somewhat limited in the very rare variant spectrum. FINRISK and Health2000 cohorts have collected samples from all over mainland Finland. In this study, however, the samples have been geographically randomly selected. As low-coverage whole-genome sequencing data are sub-optimal for detecting variants observed only in a few individuals, rare variants observed in Britons were likely to be called more confidently compared with similar variants in Finns. In addition, the British data set had slightly higher coverage than the Finnish data (4.6x *vs* 7x), which may have had some effect on calling of the rare and low-frequency variants in Finns. Such technical limitations and differences may have led to under-estimation of our main findings (except the depletion of rare variants in Finns). Our comparison was limited against British samples and autosomal SNVs only, and future studies should therefore carry out comparisons against a panel of jointly processed heterogeneous population samples, including all types of variants (also from sex chromosomes). When comparing the Kuusamo sub-isolate sample to the Finnish non-Kuusamo individuals, we found that only LoF variants (that also showed the largest enrichment between Finns and Britons) appear

significantly enriched. This is possibly due to the small sample size of the Kuusamo subset.

This study provides insights into the effects of a population bottleneck in various functional categories across the whole human genome. Obvious advantages of isolated populations are significantly reduced heterogeneity in genetic architecture, phenotype and the environment. The frequency of an originally rare allele that passed through the population bottleneck can be increased by several orders of magnitude (even >100-fold for some variants), after which it will decline relatively slowly (due to selective pressure). This phenomenon will therefore increase the statistical power to identify rare variants associated with complex disorders in both coding as well as non-coding regions of the human genome in isolated populations.[1,13]

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Doctoral Programme in Biomedicine-DPBM (HC), the Academy of Finland (251217, 255847 and 285380 to SR; 251704 and 286500 to AP; 295504, 269862 to TA; 257654 and 288509 to MP), the Academy of Finland Center of Excellence for Complex Disease Genetics (grant nos. 213506 and 129680), the Wellcome Trust (WT089062/Z/09/Z and WT089061/Z/09/Z to SR, and WT098051 for SM, KW and RD), EU FP7 projects ENGAGE (201413 to AP & SR), BioSHaRE (261433 to AP and SR), the Finnish Foundation for Cardiovascular Research (AP, SR and VS); the Sigrid Juselius Foundation, Biocentrum Helsinki (SR), the Nordic Information for Action eScience Center (NIASC)-a Nordic Center of Excellence financed by NordForsk (grant no. 62721 to AP, PP & SR); IUT20-60 Omics for health: an integrated approach to understand and predict human disease (PP). The authors also wish to acknowledge CSC – IT Centre for Science, Finland and the FIMM Technology Centre services for computational resources.

1 Lim ET, Wurtz P, Havulinna AS *et al*: Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* 2014; **10**: e1004494.
2 Sidore C, Busonero F, Maschio A *et al*: Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* 2015; **47**: 1272–1281.
3 Gudbjartsson DF, Helgason H, Gudjonsson SA *et al*: Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015; **47**: 435–444.
4 Casals F, Hodgkinson A, Hussin J *et al*: Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* 2013; **9**: e1003815.
5 Goldstein DB, Allen A, Keebler J *et al*: Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 2013; **14**: 460–470.
6 Kryukov GV, Pennacchio LA, Sunyaev SR: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 2007; **80**: 727–739.
7 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001; **69**: 124–137.
8 Jakkula E, Rehnstrom K, Varilo T *et al*: The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet* 2008; **83**: 787–794.
9 Kittles RA, Perola M, Peltonen L *et al*: Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet* 1998; **62**: 1171–1179.
10 Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L: Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000; **8**: 604–612.
11 Service S, DeYoung J, Karayiorgou M *et al*: Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* 2006; **38**: 556–560.
12 Peltonen L, Jalanko A, Varilo T: Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* 1999; **8**: 1913–1923.
13 Stoll G, Pietilainen OP, Linder B *et al*: Deletion of TOP3beta, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. *Nat Neurosci* 2013; **16**: 1228–1237.
14 Vartiainen E, Laatikainen T, Peltonen M *et al*: Thirty-five-year trends in cardiovascular risk factors in Finland. *Int J Epidemiol* 2010; **39**: 504–518.
15 McCarthy S, Das S, Kretzschmar W *et al*: A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016; **48**: 1279–1283.
16 O'Connell J, Sharp K, Shrine N *et al*: Haplotype estimation for biobank-scale data sets. *Nat Genet* 2016; **48**: 817–820.
17 Kent WJ, Sugnet CW, Furey TS *et al*: The human genome browser at UCSC. *Genome Res* 2002; **12**: 996–1006.
18 Harrow J, Frankish A, Gonzalez JM *et al*: GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 2012; **22**: 1760–1774.
19 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; **26**: 2069–2070.
20 Adzhubei IA, Schmidt S, Peshkin L *et al*: A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.
21 Trynka G, Sandor C, Han B *et al*: Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 2013; **45**: 124–130.
22 Lindblad-Toh K, Garber M, Zuk O *et al*: A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 2011; **478**: 476–482.
23 Ward LD, Kellis M: Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 2012; **337**: 1675–1678.
24 Andersson R, Gebhard C, Miguel-Escalada I *et al*: An atlas of active enhancers across human cell types and tissues. *Nature* 2014; **507**: 455–461.
25 Hnisz D, Abraham BJ, Lee TI *et al*: Super-enhancers in the control of cell identity and disease. *Cell* 2013; **155**: 934–947.
26 Whitfield TW, Wang J, Collins PJ *et al*: Functional analysis of transcription factor binding sites in human promoters. *Genome Biol* 2012; **13**: R50.
27 Vukcevic D, Hechter E, Spencer C, Donnelly P: Disease model distortion in association studies. *Genet Epidemiol* 2011; **35**: 278–290.
28 Consortium UK, Walter K, Min JL *et al*: The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
29 Finucane HK, Bulik-Sullivan B, Gusev A *et al*: Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015; **47**: 1228–1235.
30 Welter D, MacArthur J, Morales J *et al*: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014; **42**: D1001–D1006.
31 Landrum MJ, Lee JM, Benson M *et al*: ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016; **44**: D862–D868.
32 Kurki MI, Gaal EI, Kettunen J *et al*: High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms. *PLoS Genet* 2014; **10**: e1004134.
33 Keinan A, Clark AG: Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012; **336**: 740–743.
34 Coventry A, Bull-Otterson LM, Liu X *et al*: Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 2010; **1**: 131.
35 Trynka G, Westra HJ, Slowikowski K *et al*: Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am J Hum Genet* 2015; **97**: 139–152.
36 Consortium EP: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57–74.
37 Pott S, Lieb JD: What are super-enhancers? *Nat Genet* 2015; **47**: 8–12.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)