bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Title: The sequences of 150,119 genomes in the UK biobank

Authors: Bjarni V. Halldorsson[1,2], Hannes P. Eggertsson[1], Kristjan H.S. Moore[1], Hannes Hauswedell[1], Ogmundur Eiriksson[1], Magnus O. Ulfarsson[1,3], Gunnar Palsson[1], Marteinn T. Hardarson[1], Asmundur Oddsson[1], Brynjar O. Jensson[1], Snaedis Kristmundsdottir[1], Brynja D. Sigurpalsdottir[1], Olafur A. Stefansson[1], Doruk Beyter[1], Guillaume Holley[1], Vinicius Tragante[1], Arnaldur Gylfason[1], Pall I. Olason[1], Florian Zink[1], Margret Asgeirsdottir[1], Sverrir T. Sverrisson[1], Brynjar Sigurdsson[1], Sigurjon A. Gudjonsson[1], Gunnar T. Sigurdsson[1], Gisli H. Halldorsson[1], Gardar Sveinbjornsson[1], Kristjan Norland[1], Unnur Styrkarsdottir[1], Droplaug N. Magnusdottir[1], Steinunn Snorradottir[1], Kari Kristinsson[1], Emilia Sobech[1], Gudmar Thorleifsson[1], Frosti Jonsson[1], Pall Melsted[1,3], Ingileif Jonsdottir[1,4], Thorunn Rafnar[1], Hilma Holm[1], Hreinn Stefansson[1], Jona Saemundsdottir[1], Daniel F. Gudbjartsson[1,3], Olafur T. Magnusson[1], Gisli Masson[1], Unnur Thorsteinsdottir[1,4], Agnar Helgason[1,5], Hakon Jonsson[1], Patrick Sulem[1], Kari Stefansson[1]

Affiliations:

1 deCODE genetics / Amgen Inc., Sturlugata 8, Reykjavik, Iceland

2 School of Technology, Reykjavik University, Reykjavik, Iceland

3 School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland

4 Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland

5 Department of Anthropology, University of Iceland, Reykjavik, Iceland

*Correspondance to: Bjarni V. Halldorsson, deCODE genetics / Amgen Inc., Sturlugata 8, 102 Reykjavik, Iceland. bjarnih@decode.is, Phone: 354-5701808, fax 354-5701901

Kari Stefansson, deCODE genetics / Amgen Inc., Sturlugata 8, 102 Reykjavik, Iceland. kstefans@decode.is, Phone:354-5701900, fax 354-5701901.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

We describe the analysis of whole genome sequencing (WGS) of 150,119 individuals from the UK biobank (UKB). This yielded a set of high quality variants, including 585,040,410 SNPs, representing 7.0% of all possible human SNPs, and 58,707,036 indels. The large set of variants allows us to characterize selection based on sequence variation within a population through a Depletion Rank (DR) score for windows along the genome. DR analysis shows that coding exons represent a small fraction of regions in the genome subject to strong sequence conservation. We define three cohorts within the UKB, a large British Irish cohort (XBI) and smaller African (XAF) and South Asian (XSA) cohorts. A haplotype reference panel is provided that allows reliable imputation of most variants carried by three or more sequenced individuals. We identified 895,055 structural variants and 2,536,688 microsatellites, groups of variants typically excluded from large scale WGS studies. Using this formidable new resource, we provide several noteworthy examples of trait associations with rare variants with large effects not found previously through studies based on exome sequencing and/or imputation.

Introduction

The study of how diversity in the sequence of the human genome affects human diversity depends on reliable characterization of human sequence and phenotypic diversity. Over the past decade insights into this relationship have been obtained from whole exome (WES) and WGS of large cohorts with rich phenotypic data[1,2].

The UK biobank (UKB)[3] documents phenotypic variation across 500,000 largely healthy subjects[4] across the United Kingdom. The UKB WGS consortium is sequencing the whole genomes of all the participants to an average depth of at least 23.5x. Here, we report on the first data release consisting of a vast set of sequence variants, including Single Nucleotide Polymorphisms (SNPs), short insertions/deletions (indels), microsatellites and structural variants (SVs), based on WGS of 150,119 individuals. All variant calls were performed jointly across individuals, allowing for consistent comparison of results. The resulting dataset provides an unparalleled opportunity to study sequence diversity in humans and its impact on phenotype variation.

Previous studies of the UKB have produced genomewide SNP array data[5] and WES data[6,7]. While SNP arrays typically only capture a small fraction of common variants in the genome, when combined with a reference panel of WGS individuals[8], a much larger set of variants in these individuals can be surveyed through imputation. Imputation however misses variants private to the individuals typed only on SNP arrays and provides unreliable results for variants with insufficient haplotype sharing between carriers in the reference and imputation sets. Poorly imputed variants are typically rare, highly mutable or in genomic regions with complicated haplotype structure, often due to structural variation.

WES is mainly limited to regions known to be translated and consequently reveals only a small proportion (2-3%) of sequence variation in the human genome. It is relatively straightforward to assign function to variants inside protein coding regions, but there is abundant evidence that variants outside of coding exons are also functionally important[9–11],

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

explaining a large fraction of the heritability of traits[12,13]. In particular, numerous variants are known to impact disease and other traits through their effects on non-coding genes or RNA[14] and protein[15,16] expression.

Large scale sequencing efforts have typically focused on identifying SNPs and short indels. While these are the most abundant types of variants in the human genome, other types, including structural variants (SVs) and microsatellites, affect a greater number of bps and consequently are more likely to have a functional impact[17,18]. Even the SVs that overlap exons are difficult to ascertain with WES due to the much greater variability in the depth of sequence coverage in WES studies than in WGS due to the capture step of targeted sequencing. Microsatellites, polymorphic tandem repeats of 1 to 6 bps, are also commonly not examined in large scale sequence analysis studies. These variants have a higher mutation rate than SNPs and indels[19], can affect gene expression[20] and contribute to a range of diseases[21].

Here, we highlight some of the insights gained from this vast new resource of WGS data that would be challenging or impossible to ascertain from WES and SNP array datasets. First, we show that exons account for a small fraction of the genomic regions displaying reduced diversity indicative of selection due to functional importance. Second, we describe three ancestry-based cohorts within the UKB; with 431,805, 9,633 and 9,252 individuals with British-Irish, African and South Asian ancestries, respectively. Third, using the rich UKB phenotype collection, we report novel findings from genomewide association (GWAS) – shedding light on the impact of very rare SNP, indels, microsatellites and structural variants on diseases and other traits.

Results

SNPs and indels

The whole genomes of 150,119 UKB participants were sequenced to an average coverage of 32.5x (at least 23.5x per individual, Fig. S22) using Illumina NovaSeq sequencing machines at deCODE Genetics (90,667 individuals) and the Wellcome Trust Sanger Institute (59,452 individuals). The 150,119 individuals were used in variant discovery, participants of the UKB can withdraw consent at any time, 149,960 out of 150,119 individuals could be used for subsequent analysis.

Sequence reads were mapped to human reference genome GRCh38[22] using BWA[23]. SNPs and short indels were jointly called over all individuals using both GraphTyper[24] and GATK HaplotypeCaller[25], resulting in 655,928,639 and 710,913,648 variants, respectively. We used several approaches to compare the accuracy of the two variant callers, including comparison to curated datasets[26] (Table S4,Fig. S15), transmission of alleles in trios (Table S8,Table S11), comparison of imputation accuracy (Table S5) and comparison to WES data (Table S12). As GraphTyper provided the more accurate genotype calls the GraphTyper genotypes were used for all subsequent analyses of short variants, although further insights might be gained from exploring these call sets jointly. To contain the number of false positives, GraphTyper employs a logistic regression model that assigns each variant a score

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(AAscore) predicting the probability that it is a true positive. We focus on the 643,747,446 (98.14%) high quality GraphTyper variants, indicated by an AAscore above 0.5, hereafter referred to as GraphTyperHQ.

We find that 4.1% of the 149,960 individuals carry an actionable genotype according to ACMG[27] v3.0 (73 genes). Using WES[28] and ACMG v2.0 (59 genes), 2.0% were reported to carry an actionable genotype, when restricting our analysis to ACMG v2.0 and same criteria we find 2.5% based on WGS.

The number of variants identified per individual is 40 times larger than the number of variants identified through the WES studies of the same UKB individuals (Table 1, Methods). Although referred to as "whole exome sequencing" we find that WES primarily captures coding exons and misses most variant in exons that are transcribed but not translated, missing 72.2% and 89.4%, of the 5' and 3' untranslated region (UTR) variants, respectively. Even inside of coding exons currently curated by Encode[9], we estimate that 10.7% of variants are missed by WES (Table 1). Conversely, almost all variants identified by WES are captured by WGS (Table 1).

Identification of functionally important regions

The number of SNPs discovered in our study corresponds to an average of one every 4.8 basepairs (bp), in the regions of the genome that are mappable for short sequence reads. This amounts to detection of 7.0% of all theoretically possible SNP variants for these regions. We observe 81.5% of all possible autosomal CpG>TpG variants, 11.8% of other transitions and only 4.0% of transversions (Table S1). Restricting the analysis to 17,902,255 autosomal CpG dinucleotides methylated in the germline[10], we observe transition variants at 89.1% of all methylated CpGs. Due to this saturation of mutations (Fig. 4d), the ratio of transitions to transversions (1.66) is lower than found in smaller WGS sets[1] and de-novo mutation (DNM) studies[29].

The vast majority of all variants identified are rare (Table S9), 46.0% and 40.6% of all SNPs and short indels, respectively, are singletons (carried by a single sequenced individual), and 96.6% and 91.7% have frequency below 0.1%. Due to the scale of the UKB WGS data, an observation of the same allele in unrelated individuals does not always imply identity by descent. A clear indication of this is that only (14%) of the highly saturated CpG>TpG variants are singletons, in contrast to 47% for other SNP variants. These recurrence phenomena have been described in other sample sets using sharing of rare variants between different subsets[2,11]. We used a DNM set from 2,976 trios in Iceland[29] to assess recurrence directly, variants present in both that set and the UKB must be derived from at least two mutational events. Out of the 194,687 Icelandic DNMs we find 53,859 (27.7%) in the UKB set providing a direct observation of sequence variants that are derived from multiple mutational events. As expected, we find that CpG>TpG mutations are the most enriched mutation class in the overlap (Fig. 4d), due to their high mutation rate[30].

The rate and pattern of variants in the genome is informative about the mutation and selection processes that have shaped the genome[31]. The number of sequence variants in the exome has been used to rank genes according to their intolerance to loss-of-function (LoF)

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and missense variation[11,32]. The focus has been on the exome due to the availability of WES datasets and the relatively straightforward functional interpretation of coding variants. Inter species conservation[33] have been used to characterize selection beyond the exome, using the extensive accumulation of mutations over millions of years, but these methods fail to identify sequence conservation specific to humans. Sequence variation within human populations[34,35] has been used to characterize human specific conservation, however, this requires many human genomes to make accurate inference as a much smaller number of mutations separate two humans than two species.

In large cohorts most theoretically possible CpG>TpG variants at methylated CpGs have been observed to occur in coding exons and their absence has been used as a sign of negative selection[11,36]. In line with previous reports[11] we see less saturation of stop-gain CpG>TpG variants than those that are synonymous (Fig. 4a). Synonymous mutations are often assumed to be unaffected by selection (neutral)[36] however we find that synonymous CpG>TpG mutations are less saturated (85.7%) than those that are intergenic (89.9%).

We used sequence variant counts in the UKB to seek conserved regions in 500bp windows across the genome. More specifically, we tabulated the number of variants in each window and compared this number to an expected number given the nucleotide composition of the window. We then assigned a rank (Depletion Rank, DR) from 0 (most depletion) to 100 (least depletion) for each 500bp window. As expected, coding exons have low DR (mean DR = 28.3), however, a large number of non-coding regions show lower DR, including non-coding regulatory elements Among the 1% of regions with lowest DR, 14.1% are coding and 85.9% are non-coding, with an overrepresentation of splice, UTR, gene upstream and downstream regions (Fig. 4e). After removing coding exons, among the 1% of regions with lowest and highest DR score we see a 3.4 and 0.4-fold overrepresentation of GWAS variants, respectively (Table 2). Regions under strong negative selection are also expected to have a greater fraction of rare variants (FRV) than the rest of the genome[35]. As most variants are carried only by a few individuals, we define FRV as the fraction of variants carried by at most 4 WGS individuals and find FRV of 74.7% in windows with DR less than 5 compared to FRV of 69.6% in regions with DR above 95 (Fig. 4f). In particular, this also holds true when we limit to only non-coding regions (74.4% vs 69.7%).

We find that there is a correlation between DR and interspecies scores as measured by GERP[33] ($r^2 = 0.0049$, $p = 0.00052$, Fig. 1d). Interestingly, in the windows with 1% lowest DR, 48.4% of windows do not show sequence conservation between species (GERP < 0), indicating that DR is informative about human specific selection. Overall, our results indicate that DR can be used to measure negative selection across the entire genome and as such provides a valuable resource for identifying non-coding sequence of functional importance.

Multiple cohorts within UKB

Most GWAS[37–39] on the UKB set have been based on a prescribed[5] Caucasian subset of 409,559 participants who self-identify as White British. To better leverage the value of genotypes of UKB participants for GWAS, we defined three cohorts encompassing 450,690 individuals (Table S2), based on genetic clustering of microarray genotypes informed by self-described ethnicity and supervised ancestry inference (Methods). The largest cohort, XBI,

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

contains 431,805 individuals who, include 99.6% of the aforementioned 409,559 prescribed Caucasian set, along with around 23,900 additional individuals previously excluded because they did not identify as "White British" (thereof 13,000 who identified as "White Irish"). A principal component analysis (PCA) of the 132,000 XBI individuals with WGS data (Methods), based on 4.6 million loci, reveals an extraordinarily fine-scaled differentiation by geography in the British–Irish Isles gene pool (Fig. S3).

We defined two other cohorts based on our inference of ancestry derived from Africa (XAF, N=9,633) and South Asia (XSA, N=9,252) (Fig. 1). The 37,598 UKB individuals who do not belong to XBI, XAF or XSA were assigned to the cohort OTH (others). The WGS data of the XAF cohort represents one of the most comprehensive surveys of African sequence variation to date, with reported birthplaces of its members covering 31 of the 44 countries on mainland sub-Saharan Africa (Fig. S7). Due to the greater genetic diversity of African populations, and resultant differences in patterns of linkage disequilibrium, the XAF cohort may prove valuable for fine-mapping of association signals that are linked to multiple strongly correlated variants in XBI or other non-African GWA studies.

We crossed GraphTyperHQ variants with exon annotations and found that on average around one in thirty is homozygous for rare (minor allele frequency, MAF < 1%) LoF mutations in the homozygous state and the median number of heterozygous rare LoF is 24 per individual. We detect rare LoF in 19,105 genes and a total of 2,017 genes were found to harbor rare LoFs in the homozygote state (n individuals = 5,102). A marked difference in the number of homozygous LoFs carriers was found between the cohorts, with XSA having the largest fraction of homozygous LoF carriers (Fig. S9b). A notable feature of the XSA cohort is elevated genomic inbreeding due to endogamy[40], particularly among self-identified Pakistanis[41] (Fig. S9a).

On average, each individual carried alternative alleles for 3,410,510 SNPs and indels (Fig. 1d), per haploid genome. XAF individuals carry more alternative alleles (Fig. 1d), primarily due to ancestry-based differences from human reference genome[22]. Indeed, a greater number of variants are generally found in individuals born outside of Europe (Fig. S10). The average number of singletons per individual varies considerably by ancestry (Fig. 1d). Thus, individuals from the XBI, XAF and XSA cohorts have an average of 1,330, 9623 and 8340 singleton variants, respectively. In XBI, singleton counts (Fig. 3a) indicate that expected variants discovered per marginal British–Irish genome is still substantial, but varies geographically, averaging around 1,000 in Northern England and 2,000 South-Eastern England. This pattern is largely explained by denser sampling of some regions (Fig. 3b, c) rather than regional ancestry differences.

Imputation

We were able to reliably impute variants into the entire UKB sample set down to very low frequency (Fig. 1e, Methods). We imputed phased genotypes which permit analysis that depend on phase such as identification of compound LoF heterozygotes. A single reference panel was used to impute all individuals in UKB, but results are presented separately for the three cohorts (Table S14). This reference panel can be used for accurate imputation in individuals from the UK and many other populations. In the XBI cohort, 98.5% of variants

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

with frequency above 0.1% and 65.8% of variants in the frequency category of 0.001-0.002% (representing 3-5 WGS carriers) could be reliably imputed (Fig. 1e). Variants were also imputed with high accuracy in XAF and XSA (Fig. 1e), where 97.5% and 94.9% of variants in frequencies 1-5% and 56.6% and 48.9% of variants carried by 3-5 sequenced individuals could be imputed, respectively. It is thus likely that the UKB reference panel provides the best available option for imputing genotypes into population samples from Africa and South Asia.

We found a number of clinically important variants that can now be imputed from the dataset. These include rs63750205 (NM_000518.5(HBB):c.*110_*111del) in the 3' UTR of HBB, a variant that has been annotated in ClinVar[42] as likely pathogenic for beta Thalassemia. rs63750205-TTA has 0.005% frequency (freq) in the imputed XBI cohort (imputation information (imp info) 0.98) and is associated with lower mean corpuscular volume by 2.88 s.d. (95% CI 2.43-3.33, $p = 1.5 \cdot 10^{-36}$).

In the XSA cohort we found rs563555492-G, a previously reported[43] missense variant in *PIEZO1* (freq = 3.65% XSA, 0.046% XAF, 0.0022% XBI) that associates with higher haemoglobin concentration, effect 0.36 s.d. (95% CI 0.28-0.44, $p = 8.1 \cdot 10^{-19}$). The variant can be imputed into the XSA population with imp info of 0.99.

In the XAF cohort we found the stop gain variant rs28362286-C (p.Cys679Ter) in *PCSK9* (freq = 0.93% XAF, 0.00016% XBI, 0.0070% XSA) which is imputed in the XAF cohort with imp info 0.93. The variant lowers non-HDL cholesterol by 0.92 s.d. (95% CI 0.75-1.09, $p = 2.3 \cdot 10^{-26}$). We found a single homozygous carrier of this variant, which has 2.5 s.d. lower non-HDL cholesterol than the population mean, is 61 years old and appears to be healthy.

SNP and indel associations not present in WES data

We highlight three examples of associations of SNP and indel variants associated with traits in the XBI cohort that could not be easily identified in WES or SNP array data.

The first example is an association in the XBI cohort between a rare variant rs117919628-A (freq = 0.32%; imp info = 0.90) in the promoter region of GHRH, encoding the growth hormone releasing hormone close to one of its TSS (Transcription start site) and less height (effect = -0.32 s.d. (95% CI 0.27-0.36), $p = 1.6 \cdot 10^{-39}$). GHRH is a neuropeptide secreted by the hypothalamus to stimulate the synthesis of the growth hormone (GH). We note that the effect (-0.32 s.d. or -3cm) of rs117919628 is greater than any variants reported in large height GWAS (~1200 associated variants)[44,45]. In addition to reducing height, rs117919628-A is associated with lower IGF-1 serum levels (Insulin-growth factor 1, effect = -0.36 s.d. (95% CI 0.32-0.40), $p = 3.2 \cdot 10^{-58}$), a hormone which production is stimulated by GH and mediates the effect of GH on childhood growth, a further support for *GHRH* as the gene mediating the effects of rs117919628-A. Due to its location around 50 bp upstream of the *GHRH* 5'UTR, this variant is not targeted by the UKB WES, and neither is the only strongly correlated variant rs372043631 (intronic). The height associations of these two variants have not been reported, presumably because they are absent from all versions of the 1,000 genomes[46] and in imputations based on the haplotype reference consortium/UK 10K[47] (HRC/UK10K) these two variants have low imp info (0.54) and would thus fail quality checks. In *GHRH*, we also

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

observe a very rare frameshift deletion rs763014119-C (Phe7Leufster2; freq = 0.0092%) associated with reduced height and IGF-1 levels (height effect = -0.63 s.d (95% CI 0.36-0.89), $p = 4.6 \cdot 10^{-6}$; IGF-1 effect = -0.74 s.d. (95% CI 0.49-0.99), $p = 4.9 \cdot 10^{-9}$). This variant is not correlated with the promoter variant rs117919628 (no individuals carry the minor allele of both variants).

Our second example is rs939016030-A a rare 3' UTR essential splice acceptor variant in the gene encoding tachykinin 3 (*TAC3*; freq = 0.033%; c.*2-1G>T in NM_001178054.1 and NM_013251.3). The XBI cohort has 89 WGS carriers and 281 in the imputation set. This variant is not found in WES of the UKB[47] and highly correlated with two other variants, one intronic and one intergenic (rs34711498, rs368268673) also not found by WES. These 3 variants were absent from the HRC/UK10K[48] imputation, and are only present in Europeans, with highest frequency in the UK according to Gnomad[11]. The minor allele of this 3'UTR essential splice variant rs939016030-A is associated with later age of menarche, with an effect of 0.57 s.d. (95% CI 0.41-0.74) or 11 months ($p = 1.0 \cdot 10^{-11}$). Rare coding variants in *TAC3* and its receptor *TACR3* are reported to cause hypogonadotropic hypogonadism[49] under an autosomal recessive inheritance. However, in the UKB, the association of the 3´UTR splice acceptor variant, is only driven by heterozygotes (~ 1 in 1500 individuals) since no homozygotes were detected in the cohort.

The third example is a rare variant (rs1383914144-A; freq = 0.40%) near the centromere of chromosome 1 (start of 1q) that is associated with lower uric acid levels (effect = -0.43 s.d. (95% CI 0.40-0.46) or -0.58 mg/dL (95% CI 0.54-0.62), $p = 8.1 \cdot 10^{-170}$) and protection against gout (OR = 0.36 (95% CI 0.28-0.46), $p = 4.2 \cdot 10^{-15}$). A second variant rs1189542743, 4Mb downstream at the end of 1p is strongly correlated ($r^2 = 0.68$) and yields a similar association to uric acid. Neither variant is targeted by UKB WES nor imputed by the HRC/UK10K and no association was reported in this region in the uric acid GWAS[50]. The effect of rs1383914144-A on uric acid is larger than for any variant reported in the latest GWAS meta-analysis of this trait.

Structural variants play an important role in human genetics

We identified structural variants (SVs) in each individual using Manta[51] and combined these with variants from a long read study[52] and the assemblies of seven individuals[53]. We genotyped the resulting 895,055 SVs (Fig. 2) with GraphTyper[53], of which we considered 637,321 reliable.

On average we identified 7,963 reliable SVs per individual, 4,185 deletions and 3,778 insertion (Fig. 1d). These numbers are comparable to the 7,439 SVs per individual found by Gnomad-SV[54], another short read study, but considerably smaller than the 22,636 high quality SVs found in a long read sequencing study[52], particularly due to an underrepresentation of insertions and SVs in repetitive regions. SVs show a similar frequency distribution (Fig. 2) as SNPs and indels and a similar distribution of variants across cohorts (Fig. 1d).

We present four examples of structural variants, not easily found in WES data, associated with human traits. First, a 14,154 bp deletion that deletes the first exon in *PCSK9*, previously

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

discovered using long read sequencing in the Icelandic population, shown to be rare (0.037%) and to associate with lowere non-HDL levels[52]. We found thirty two WGS carriers in the XBI cohort (freq 0.012%) and 72 carriers in the XBI imputed set (freq 0.0087%) who had 1.22 s.d. (95% CI 0.90-1.55) lower non-HDL cholesterol levels than non-carriers ($p = 1.2 \cdot 10^{-13}$).

Our second examples is a 4,160 bp deletion, (freq = 0.037% in XBI), that removes the promoter region from 4,300 to 140 bp upstream of the *ALB* gene that encodes Albumin. Not surprisingly, carriers of this deletion have markedly lower serum albumin levels (effect 1.50 s.d. (95% CI 1.35-1.62) $p = 9.5 \cdot 10^{-118}$). The variant is also associated with traits correlated with albumin levels; carriers had lower calcium and cholesterol levels: 0.62 s.d. (95% CI 0.50-0.75, $p = 2.9 \cdot 10^{-22}$) and 0.45 s.d. (95% CI 0.30-0.59, $p = 1.1 \cdot 10^{-9}$), respectively.

Our third example is a 16,411 bp deletion (freq = 0.0090% in XBI) that removes the last two exons (4 and 5) of *GCSH*, that encodes Glycine cleavage system H protein. Carriers of this deletion have markedly higher Glycine levels in the UKB metabolomics dataset (effect 1.45 s.d. (95% CI 1.01-1.86), $p = 1.2 \cdot 10^{-10}$).

The final example is a rare (freq 0.892% in XBI) 754bp deletion overlapping exon 6 of *NMRK2*, encoding nicotinamide riboside kinase 2 that removes 72 bp from the transcribed RNA that corresponds to a 24 amino acid inframe deletion in the translated protein. Carriers of this deletion have a 0.22 s.d. (95% CI 0.18-0.27) earlier age at menopause ($p = 1.1 \cdot 10^{-26}$). Nearby is the variant rs147068659, reported to associate with this trait[55], with an effect 0.20 s.d. (95% CI 0.16-0.24) earlier age at menopause ($p = 2.0 \cdot 10^{-20}$) in the XBI cohort. The deletion and rs147068659 are correlated ($r^2 = 0.67$), after conditional analysis the deletion remains significant ($p = 6.4 \cdot 10^{-8}$) whereas rs147068659 does not ($p = 0.39$), indicating the deletion is causal. *NMRK2* is primarily expressed in heart and muscle tissue[56]; in our dataset of right atrium heart tissue, one individual out of a set of 169 RNA sequenced individuals is a carrier of this deletion. As expected we observe decreased expression of exon 6 in this individual (Fig. S4) and an increase in the fraction of transcript fragments skipping exon 6 (Fig. S5).

Microsatellites are commonly overlooked

We identified 14,321,152 alleles at 2,536,688 microsatellite loci using popSTR[57] in the 150,119 WGS individuals, who carry on average of 810,606 non-reference microsatellite alleles. The number of non-reference alleles carried per individual shows a similar distribution across the UKB cohorts as other variant types characterized in this study (Fig. 1d). Microsatellites are among the most rapidly mutating variants in the human genome and a source of genetic variation that is usually overlooked in GWAS. Repeat expansions are known to associate with a number of phenotypes, including Fragile X syndrome[58]. We are able to impute microsatellites down to a very low frequency (Fig. S2) in all three cohorts, providing one of the first large scale datasets of imputed microsatellites.

We genotyped a microsatellite within the *CACNA1A* gene that encodes voltage-gated calcium channel subunit alpha 1A. Individuals who have twenty or more repeats of this microsatellite generally suffer from lifelong conditions that affect the brain, including

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Familial hemiplegic migraine (FHM1), Epilepsy, Episodic Ataxia Type 2 (EA2) and Spinocerebellar ataxia type 6 (SCA6)[59–62]. Carriers in the XBI cohort of 22 copies of the microsattelite repeat were at greater risk for hereditary ataxia (freq = 0.0071%, OR = 304, p = $1.1 \cdot 10^{-31}$).

In the XBI cohort we also confirm an association between a microsatellite within the 3' UTR of *DMPK*, encoding DM1 protein kinase, and myotonic dystrophy. Expression of *DMPK* is negatively correlated with the number of repeats of the microsatellite[63]. The risk of myotonic dystrophy increases with copy number of the repeats, rising rapidly with the number of repeats carried by an individual up to an odds ratio of 161 for individuals carrying 39 or more repeats (Table S17, Fig. S14).

Variants that are not imputed

Although the vast majority of WGS variants can be imputed to the larger set of SNP array genotyped individuals it is interesting to examine the variants that are not imputed. A subset of these variants are in regions where there are no nearby variants present in the SNP array data and regions where there is disagreement between the GRCh38[22] and CHM13[64] assemblies. Lifting variants over to the CHM13 assembly may allow us to impute a subset of these variants. The failure of those variants to impute on GRCh38 can presumably be attributed to a misassembly on GRCh38. In addition, we identify a number of variants that are most likely recurrently somatic, such as the gain of function mutations in JAK2[65–67] and CALR[67] know to be associated with myeloproliferative disorders, including polycythaemia vera and essential thrombocythemia.

Discussion

The dataset provided by sequencing the whole genomes of 150 thousand UKB participants is unparalleled in its size and provides the most extensive characterization of the sequence diversity in the germline genomes of a single population to date. The UK population is diverse in its genetic ancestry and includes individuals born in countries all over the globe. Our African and South Asian ancestry cohorts each nuber over 9,000 individuals, representing some of the largest available WGS sets of these ancestries and which are likely to have an impact both clinically and in further characterizing the relationship between sequence and traits.

We have characterized an extensive set of sequence variants in the WGS individuals, providing two sets of SNP and indel data, as well as microsatellite and SV data, variant classes that are frequently not interrogated in GWAS. We give examples of how these variants play a role in the relationship between sequence and phenotypic variation. The number of SNPs and indels are 40-fold greater than from WES of the same individuals. Even within annotated coding exons WES misses 10.7% of variants. WES misses most of the remainder of the genome, including functionally important UTR, promoter regions and exons yet to be annotated. The importance of these regions is exemplified by the discovery of rare non-coding sequence variants with larger effects on height and menarche than any variants described in GWAS to date.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The DR score presented here is an important resource in identifying which regions are functionally important. Although coding exons are clearly under strong purifying selection, as represented by a low DR score, they represent only a small fraction of the regions with low DR score. Clinical geneticists, typically focusing on protein exons, have only been able to identify the genetic cause in fewer than half of clinical cases studied. Currently, 98.4% of variants annotated as pathogenic in the ClinVar[42] database are within coding exons. Greater attention needs to be given to other regions of the genome, particularly those with low DR score, where non-coding exons (UTRs), enhancer and promoter regions are overrepresented.

There are still some sequence variants that are not found with short read WGS, in particular regions more easily reached by long read sequencing[52], including VNTRs, repetitive regions and regions that have only recently been captured by human genome assemblies[64]. Improved assembly[64,68], sequencing and representation of the genome and its variation will have important implications for advancing our understanding of the relationship between sequence variation and human diseases and other traits.

A study of WES from 455K individuals in UK biobank recently reported several examples of associations, including those from gene burden analysis[69]. According to the authors, a majority were either in part previously reported, or could not be replicated. It is noteworthy that none of the associations reported here were found in that comprehensive survey of UKB exome variation. Near complete sequence of the human genome has been known for over twenty years. Genome scientists have yet to assign function to a large fraction of this sequence and geneticists have had only partial success in understanding the relationship between diversity in the sequence of the human genome and phenotypic diversity. The large scale sequencing described here, as well as the continued effort in sequencing the entire UKB, promises to vastly increase our understanding of the function and impact of the non-coding genome. When combined with the extensive characterization of phenotypic diversity in the UKB, these data should greatly improve our understanding of the relationship between human genome variation and phenotype diversity.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Author Contributions

Paper was written by BVH and KS with input from HPE, KHSM, OE, DFG, OTM, GM, UT, AH, HJ and PS. KHSM and AH defined cohorts. OE and HJ identified functionally important regions. FJ and UT were responsible for laboratory operations. DNA sequencing was performed by DNM, SS, KK and OTM. Sample isolation was performed by ES and JS. GM was responsible for the sequence analysis pipeline, developed by BVH, AG, PIO, MA, STS, FZ and SAG, and run by GTS. BVH, HPE, HHa, GP, SK, GH and SAG developed analysis tools. Association analysis was performed by BVH, MOU, AO, BOJ, SK, BDS, DB, VT, US and PS. Phenotypes were defined by MOU, VT, GT, IJ, TR, HHo, HS and PS. SNP and SV genotyping was performed by HPE, PIO and BS. Microsatellite genotyping was performed by SK. Data analysis was performed by BVH, HPE, HHa, GP, AO, OAS, GS and KN. RNA sequence data was analyzed by GHH, supervised by PM. DFG supervised association, data and Depletion Rank analysis. Figures were drawn by MTH and KHSM. Study was supervised by BVH and KS. All authors agreed to the final version of the manuscript.

Data availability

WGS and genotype data can be accessed via the UKB research analysis platform (RAP).

Code availability

BamQC, https://github.com/DecodeGenetics/BamQC.
GraphTyper, https://github.com/DecodeGenetics/graphtyper.
GATK resource bundle, gs://genomics-public-data/resources/broad/hg38/v0.
Svimmer, https://github.com/DecodeGenetics/svimmer.
popSTR, https://github.com/DecodeGenetics/popSTR.
Dipcall, https://github.com/lh3/dipcall.
RTG Tools, https://github.com/RealTimeGenomics/rtg-tools.
bcl2fastq, https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html.
Samtools, http://www.htslib.org/.
Samblaster https://github.com/GregoryFaust/samblaster.

Ethics declaration

All authors are employees of deCODE genetics/Amgen.

Acknowledgements

We thank the participants of the UKB. The sequencing of 450,000 WGS individuals from the UKB, including the 150,119 described here has been funded by the UKB WGS consortium consisting of UK Government's research and innovation agency, UK Research and Innovation (UKRI), through the Industrial Strategy Challenge Fund, The Wellcome Trust and the pharmaceutical companies Amgen, AstraZeneca, GlaxoSmithKline and Johnson & Johnson. DNA sequenced was performed at the Welcome Trust Sanger Institute and deCODE genetics.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435 (2015).
2. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nat. 2021 5907845* **590**, 290–299 (2021).
3. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med.* **12**, e1001779 (2015).
4. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nat. 2018 5627726* **562**, 203–209 (2018).
6. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nat. 2020 5867831* **586**, 749–756 (2020).
7. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet. 2021 537* **53**, 942–948 (2021).
8. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
9. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
10. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nat. 2020 5837818* **583**, 699–710 (2020).
11. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
12. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
13. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
14. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet. 2016 485* **48**, 481–487 (2016).
15. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet. 2020 5210* **52**, 1122–1131 (2020).
16. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73 (2018).
17. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125 (2013).
18. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75 (2015).
19. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44**, 1161 (2012).
20. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet. 2015 481* **48**, 22–29 (2015).
21. Gatchel, J. R. & Zoghbi, H. Y. Diseases of Unstable Repeat Expansion: Mechanisms and

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Common Principles. *Nat. Rev. Genet. 2005 610* **6**, 743–755 (2005).
22. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
23. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
24. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
25. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
26. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).
27. Miller, D. T. *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med. 2021 238* **23**, 1381–1390 (2021).
28. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nat. 2020 5867831* **586**, 749–756 (2020).
29. Halldorsson, B. V. *et al.* Human genetics: Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science (80-.).* **363**, (2019).
30. Jónsson, H. *et al.* Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. data* **4**, 170115 (2017).
31. Seplyarskiy, V. B. *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science (80-.).* **373**, 1030–1035 (2021).
32. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet. 2018 511* **51**, 88–95 (2018).
33. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
34. J, di I. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
35. Dukler, N., Mughal, M. R., Ramani, R., Huang, Y.-F. & Siepel, A. Extreme purifying selection against point mutations in the human genome. *bioRxiv* 2021.08.23.457339 (2021). doi:10.1101/2021.08.23.457339
36. Agarwal, I. & Przeworski, M. Mutation saturation for fitness effects at human CpG sites. *bioRxiv* 2021.06.02.446661 (2021). doi:10.1101/2021.06.02.446661
37. Deaton, A. M. *et al.* Gene-level analysis of rare variants in 363,977 whole exome sequences identifies an association of GIGYF1 loss of function with type 2 diabetes. *medRxiv* 2021.01.19.21250105 (2021). doi:10.1101/2021.01.19.21250105
38. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet. 2021 532* **53**, 185–194 (2021).
39. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. **597**, (2021).
40. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet. 2017 499* **49**, 1403–1407 (2017).
41. Arciero, E. *et al.* Fine-scale population structure and demographic history of British Pakistanis. *bioRxiv* 2020.09.02.279190 (2020). doi:10.1101/2020.09.02.279190
42. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

43. Sun, Q. *et al.* Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. *J. Hum. Genet. 2021* 1–7 (2021). doi:10.1038/s10038-021-00968-0
44. L, Y. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
45. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nat. 2017 5427640* **542**, 186–190 (2017).
46. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
47. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet. 2016 4810* **48**, 1279–1283 (2016).
48. Chou, W.-C. *et al.* A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Sci. Reports 2016 61* **6**, 1–9 (2016).
49. Topaloglu, A. K. *et al.* TAC3 and TACR3 mutations in familial hypogonadotropic hypogonadism reveal a key role for Neurokinin B in the central control of reproduction. *Nat. Genet. 2008 413* **41**, 354–358 (2008).
50. Tin, A. *et al.* Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet.* **51**, 1459 (2019).
51. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
52. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet. 2021 536* **53**, 779–786 (2021).
53. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **To Appear**, (2019).
54. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
55. Ruth, K. S. *et al.* Genetic insights into biological mechanisms governing human ovarian ageing. *Nat. 2021 5967872* **596**, 393–397 (2021).
56. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).
57. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B. & Halldórsson, B. V. popSTR: population-scale detection of STR variants. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btw568
58. Verkerk, a J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
59. Ophoff, R. A. *et al.* Familial Hemiplegic Migraine and Episodic Ataxia Type-2 Are Caused by Mutations in the Ca2+ Channel Gene CACNL1A4. *Cell* **87**, 543–552 (1996).
60. Kordasiewicz, H. B., Thompson, R. M., Clark, H. B. & Gomez, C. M. C-termini of P/Q-type Ca 2+ channel α1A subunits translocate to nuclei and promote polyglutamine-mediated toxicity. *Hum. Mol. Genet.* **15**, 1587–1599 (2006).
61. Luo, X. *et al.* Clinically severe CACNA1A alleles affect synaptic function and neurodegeneration differentially. *PLOS Genet.* **13**, e1006905 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

62. Tian, X. *et al.* A Voltage-Gated Calcium Channel Regulates Lysosomal Fusion with Endosomes and Autophagosomes and Is Required for Neuronal Homeostasis. *PLOS Biol.* **13**, e1002103 (2015).
63. Furling, D., Lemieux, D., Taneja, K. & Puymirat, J. Decreased levels of myotonic dystrophy protein kinase (DMPK) and delayed differentiation in human myotonic dystrophy myoblasts. *Neuromuscul. Disord.* **11**, 728–735 (2001).
64. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021). doi:10.1101/2021.05.26.445798
65. Kralovics, R. *et al.* A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders. *http://dx.doi.org/10.1056/NEJMoa051113* **352**, 1779–1790 (2009).
66. James, C. *et al.* A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nat. 2005 4347037* **434**, 1144–1148 (2005).
67. Klampfl, T. *et al.* Somatic Mutations of Calreticulin in Myeloproliferative Neoplasms. *http://dx.doi.org/10.1056/NEJMoa1311347* **369**, 2379–2390 (2013).
68. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nat. 2020 5857823* **585**, 79–84 (2020).
69. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nat. 2021* 1–10 (2021). doi:10.1038/s41586-021-04103-z
70. JD, S. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
71. Jun, G., Flickinger, M., Hetrick, K., … J. R.-T. A. J. of & 2012, undefined. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Elsevier*
72. Eggertsson, H. P. & Halldorsson, B. V. read_haps: using read haplotypes to detect same species contamination in DNA sequences. *Bioinformatics* **37**, 2215–2217 (2021).
73. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
74. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
75. Cleary, J. G. *et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. doi:10.1101/023754
76. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
77. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods 2018 158* **15**, 595–597 (2018).
78. LV, W. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet. Respir. Med.* **3**, 769–781 (2015).
79. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics 2017 181* **18**, 1–7 (2017).
80. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068 (2008).
81. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165**, 2213–2233 (2003).
82. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation

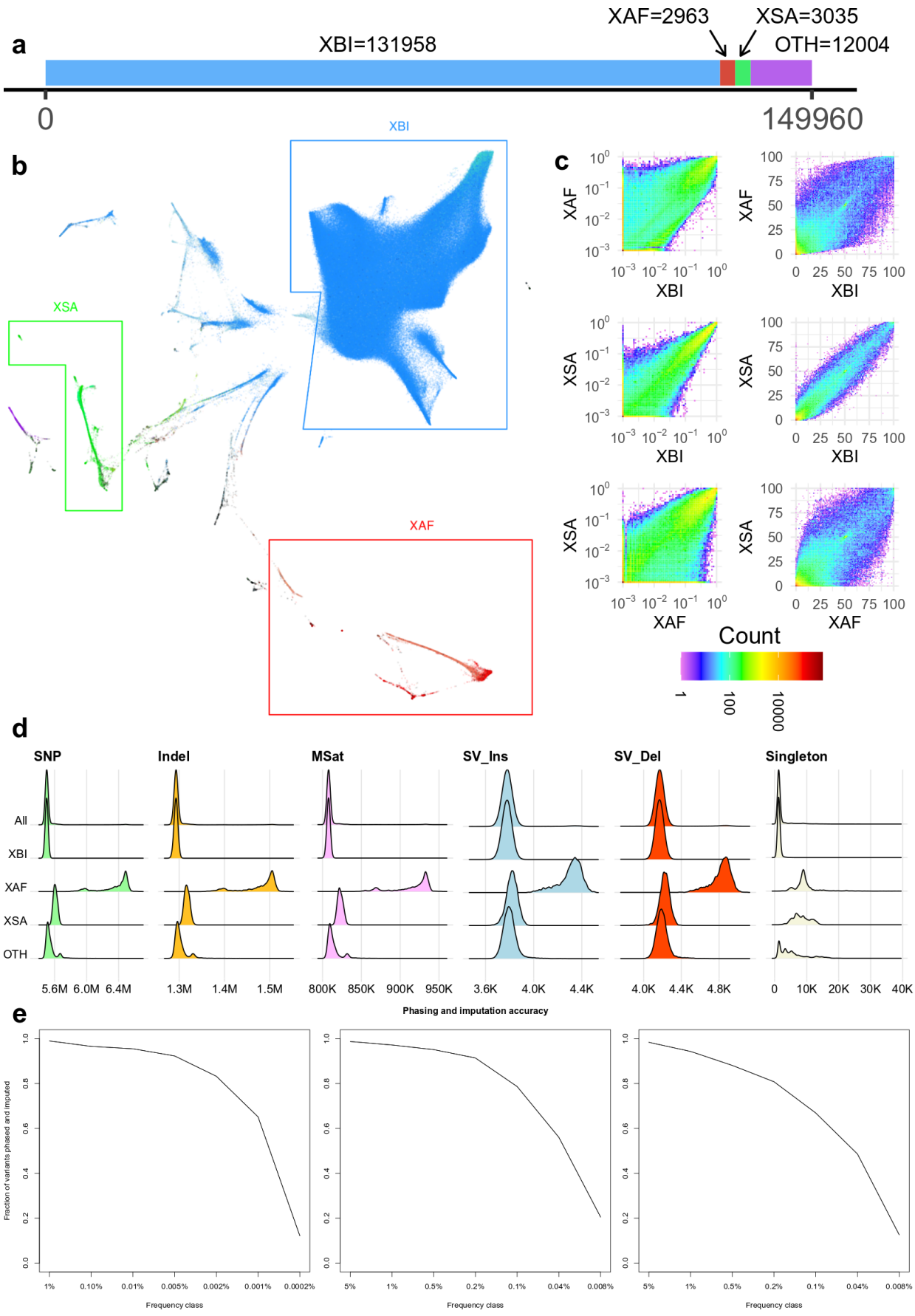
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5**, e1000529 (2009).

83. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. **26**, 2069–2070 (2010).
84. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol. 2016 171* **17**, 1–14 (2016).
85. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
86. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat. Genet.* **48**, 314–317 (2016).
87. Thorolfsdottir, R. B. *et al.* Coding variants in RPL3L and MYZAP increase risk of atrial fibrillation. *Commun. Biol. 2018 11* **1**, 1–9 (2018).
88. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet. 2021 538* **53**, 1260–1269 (2021).
89. Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C. & Gravel, S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genet.* **15**, e1008432 (2019).
90. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
91. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
92. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
93. Purcell, S. M. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, (2007).
94. Clark, D. W. *et al.* Associations of autozygosity with a broad range of human phenotypes. *Nat. Commun. 2019 101* **10**, 1–17 (2019).
95. Kunert-Graf, J., Sakhanenko, N. & Galas, D. Allele Frequency Mismatches and Apparent Mismappings in UK Biobank SNP Data. *bioRxiv* 2020.08.03.235150 (2020). doi:10.1101/2020.08.03.235150
96. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
97. Pebesma, E. Simple features for R: Standardized support for spatial vector data. *R J.* **10**, 439–446 (2018).
98. Applied Spatial Data Analysis with R. *Appl. Spat. Data Anal. with R* (2008). doi:10.1007/978-0-387-78171-6
99. Gräler, B., Pebesma, E. & Heuvelink, G. Spatio-temporal interpolation using gstat. *R J.* **8**, 204–218 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figures



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fig. 1 SNP indel and cohort characteristics a) The number of WGS samples analyzed for phenotypes in our study. b) UMAP plot generated from the first 40 principal components of all UKB participants. c) Joint frequency spectrum of variants on chr20 between all pairs of populations. d) Number of SNPs, Indels, microsatellites, SV insertions, SV deletions and singleton SNPs carried per individual in the overall set and partitioned by population. e) Imputation accuracy in the three populations, XBI, XAF and XSA. A variant was consider imputed if Leave one out r2 of phasing was greater than 0.5 and imputation info was greater than 0.8. x-axis splits variants into frequency classes based on the number of carriers in the sequence dataset, with the number representing the minimum number of carriers in the frequency class. Variants are split by variant type..

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

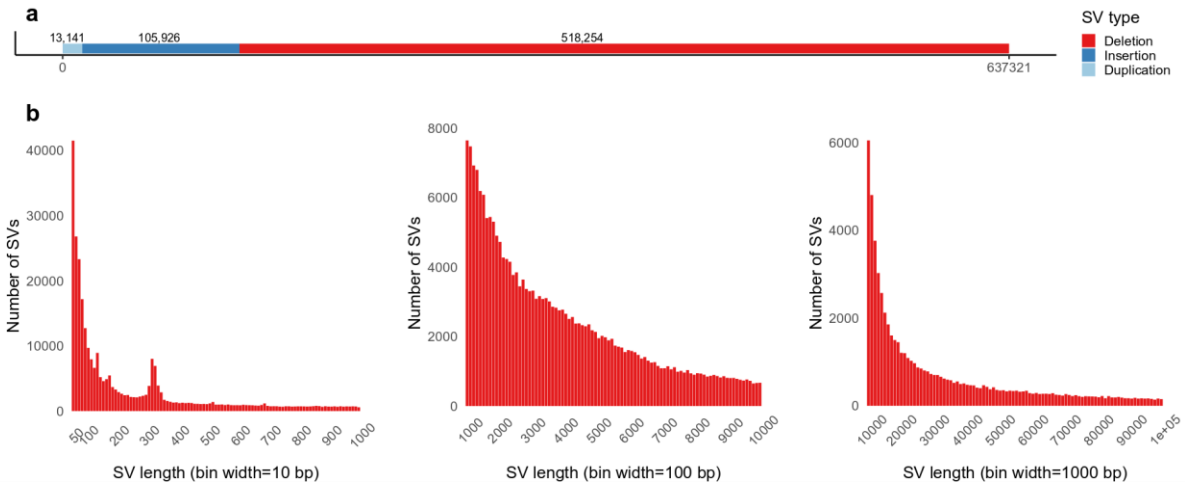


Fig. 2 Structural variants. a) Number of SVs discovered in the dataset by variant type. b) Length distribution of SVs, from 50-1,000 bp, 1,000-10,000bp and 10,000-100,000bp.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

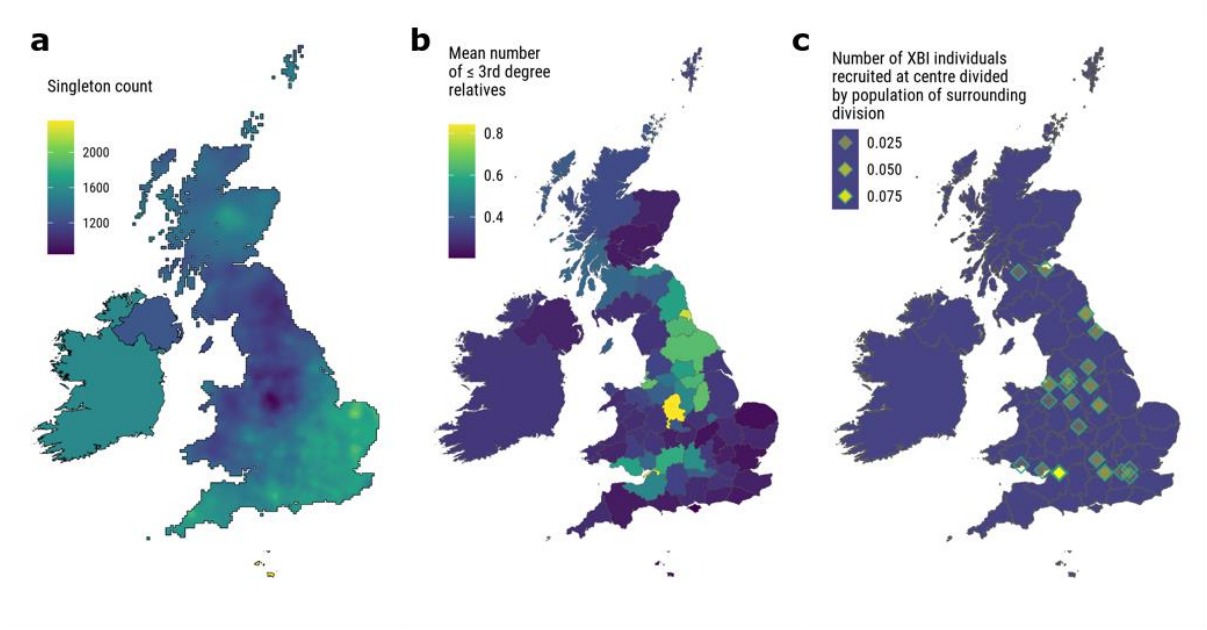


Fig. 3 Characteristic of XBI cohort across Great Britain and Ireland a) Number of singletons carried by individuals in the XBI cohort as a function of place of birth. b) Mean number of 3rd degree relatives by administrative division c) Location of UKB assessment centers and estimated fraction of surrounding population recruited to the UKB.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

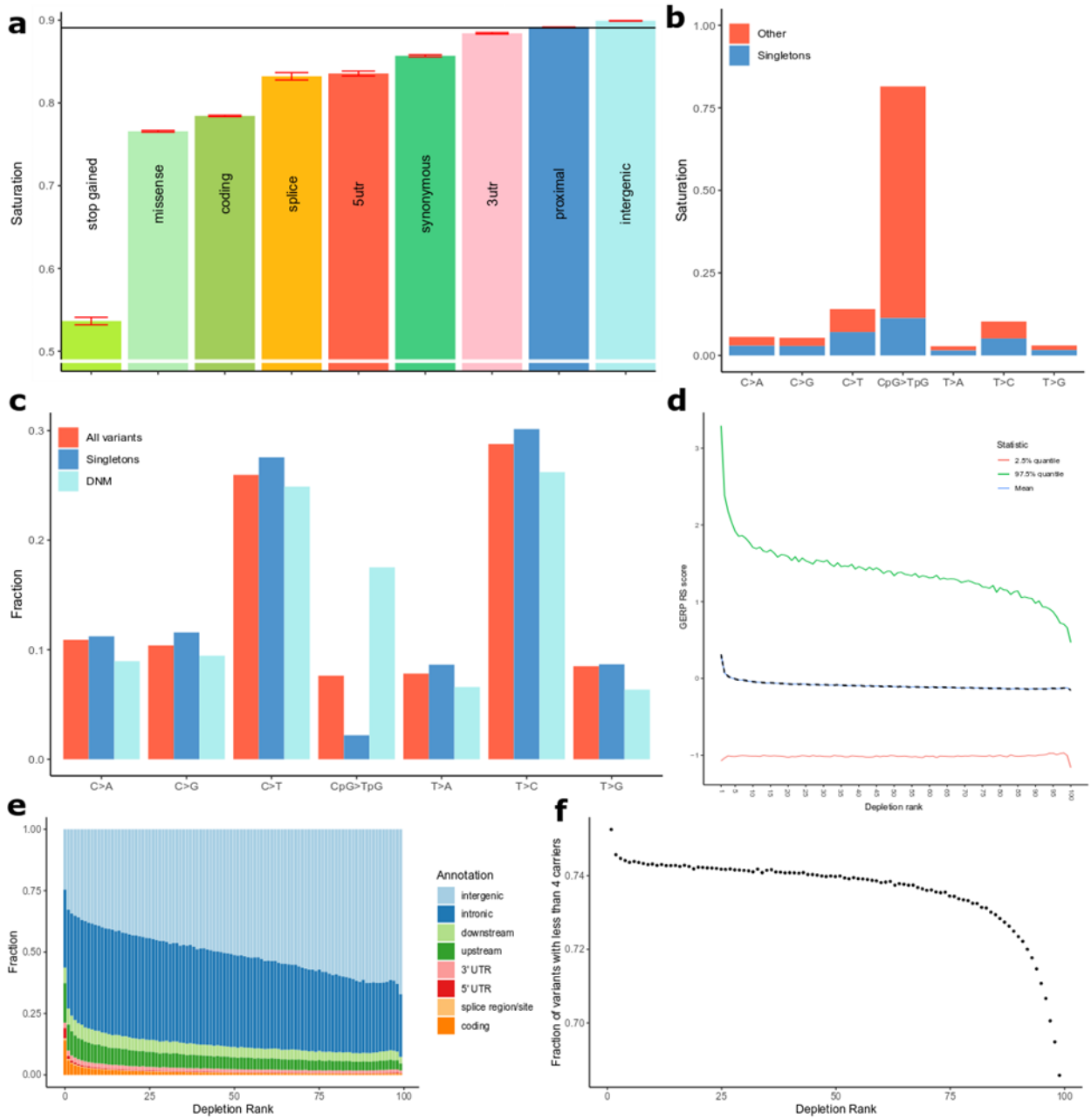


Fig. 4 Functionally important regions a) Saturation levels of transitions at methylated CpG sites across genomic annotations and predicted consequence categories. The horizontal line is the average across all methylated CpG-sites. b) Saturation levels of mutations in each class, split into singleton variants (blue) and more common variants (red). c) Fraction of SNP's in each mutation class, for all SNP's in our dataset, singletons in our dataset, and in an Icelandic set of de novo mutations (DNMs) respectively. d) Average GERP score in 500bp windows as a function of Depletion Rank, blue line represents average GERP score, red and green line 95%-th percentile e) Fraction of regions falling into functional annotation classes, as defined by Ensembl gene map, as a function of Depletion Rank. f) Fraction of rare (with 4 or fewer carriers) variants (FRV) as a function of Depletion Rank.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tables

	WGS	WES	WGS ∩ WES	WES \ WGS	Present WES	Missing WES	Present WGS	Missing WGS
coding	6,380,795	5,781,829	5,686,934	94,895	89.29%	10.71%	98.53%	1.47%
splice	445,499	397,226	388,961	8,265	87.54%	12.46%	98.18%	1.82%
5utr	2,125,413	590,484	572,996	17,488	27.56%	72.44%	99.18%	0.82%
3utr	7,214,427	764,864	743,790	21,074	10.57%	89.43%	99.71%	0.29%
proximal	249,702,570	6,189,465	5,952,145	237,320	2.48%	97.52%	99.91%	0.09%
intergenic	292,259,782	91,836	83,360	8,476	0.03%	99.97%	100.00%	0.00%

Table 1 Overlap of WES and WGS data. Results are computed for the 109,618 samples present in both datasets and is limited to those variants that are present in at least one individual in either dataset. Numbers refer to number of variants found in dataset. WGS refers to the GraphTyperHQ dataset and WES refers to a set of 200k WES sequenced indivdiduals[70]. Missing and present percentages are computed from the number of variants in the union of the two datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.16.468246; this version posted November 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DR of non-coding regions	Enrichment	95%CI	*P*-value
a)			
DR 1%	3.44	1.96-5.18	0.0004
DR 99%	0.36	0.16-0.61	<0.0002
b)			
DR 5%	2.29	1.72-2.93	<0.0002
DR 95%	0.46	0.32-0.62	<0.0002

Table 2 : Over- and underrepresentation of GWAS variants in low and high DR regions. Windows overlapping coding exons were removed. Lower DR scores indicate greater sequence conservation.