nature genetics

Large-scale whole-genome sequencing of the Icelandic population

Daniel F Gudbjartsson[1,2,21], Hannes Helgason[1,2,21], Sigurjon A Gudjonsson[1], Florian Zink[1], Asmundur Oddson[1], Arnaldur Gylfason[1], Soren Besenbacher[3], Gisli Magnusson[1], Bjarni V Halldorsson[1,4], Eirikur Hjartarson[1], Gunnar Th Sigurdsson[1], Simon N Stacey[1], Michael L Frigge[1], Hilma Holm[1,5], Jona Saemundsdottir[1], Hafdis Th Helgadottir[1], Hrefna Johannsdottir[1], Gunnlaugur Sigfusson[6], Gudmundur Thorgeirsson[7,8], Jon Th Sverrisson[9], Solveig Gretarsdottir[1], G Bragi Walters[1], Thorunn Rafnar[1], Bjarni Thjodleifsson[7], Einar S Bjornsson[8,10], Sigurdur Olafsson[8,10], Hildur Thorarinsdottir[10], Thora Steingrimsdottir[8,11], Thora S Gudmundsdottir[11], Asgeir Theodors[10], Jon G Jonasson[8,12,13], Asgeir Sigurdsson[1], Gyda Bjornsdottir[1], Jon J Jonsson[14,15], Olafur Thorarensen[16], Petur Ludvigsson[16], Hakon Gudbjartsson[1,2], Gudmundur I Eyjolfsson[17], Olof Sigurdardottir[18], Isleifur Olafsson[19], David O Arnar[7,8], Olafur Th Magnusson[1], Augustine Kong[1,2], Gisli Masson[1], Unnur Thorsteinsdottir[1,8], Agnar Helgason[1,20], Patrick Sulem[1] & Kari Stefansson[1,8]

Here we describe the insights gained from sequencing the whole genomes of 2,636 Icelanders to a median depth of 20×. We found 20 million SNPs and 1.5 million insertions-deletions (indels). We describe the density and frequency spectra of sequence variants in relation to their functional annotation, gene position, pathway and conservation score. We demonstrate an excess of homozygosity and rare protein-coding variants in Iceland. We imputed these variants into 104,220 individuals down to a minor allele frequency of 0.1% and found a recessive frameshift mutation in *MYL4* that causes early-onset atrial fibrillation, several mutations in *ABCB4* that increase risk of liver diseases and an intronic variant in *GNAS* associating with increased thyroid-stimulating hormone levels when maternally inherited. These data provide a study design that can be used to determine how variation in the sequence of the human genome gives rise to human diversity.

The advent of high-throughput genotyping and sequencing has revolutionized the ability to investigate how diversity in the sequence of the human genome affects human diversity[1]. Large-scale genotyping of common variants led to an avalanche of discoveries of variants associating with common and complex diseases[2]. Now studies based on whole-genome and exome sequencing are beginning to yield rare variants associating with common diseases[3–12]. They also provide unprecedented information about human sequence diversity and insights into the structure and history of human populations[13–16]. Several large-scale sequencing projects are ongoing or in the planning stages, foremost among them the 1000 Genomes Project[16] and the Exome Sequencing Project (ESP)[13,14], which have already provided valuable information about human genome diversity and tools to use in genetic discovery.

Our efforts at studying the human genome and its impact on diseases and other traits have focused on the Icelandic population. Genetic studies of the Icelandic population benefit from a genealogy of the nation reaching centuries back in time, a founder effect and broad access to nationwide healthcare information. The transition from genome-wide association studies (GWAS) based on common SNPs on microarrays to those based on a vast number of rare variants identified by whole-genome and exome sequencing presents new opportunities and challenges.

Here we describe the insights gained from sequencing the whole genomes of 2,636 Icelanders. First, we describe the density and

[1]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]Bioinformatics Research Centre, Aarhus University, C.F. Mollers Alle, Aarhus, Denmark. [4]Institute of Biomedical and Neural Engineering, Reykjavík University, Reykjavík, Iceland. [5]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, Minnesota, USA. [6]Children's Hospital, Landspitali University Hospital, Reykjavik, Iceland. [7]Department of Medicine, Landspitali University Hospital, Reykjavik, Iceland. [8]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [9]Department of Internal Medicine, Akureyri Hospital, Akureyri, Iceland. [10]Department of Internal Medicine, Division of Gastroenterology and Hepatology, Landspitali University Hospital, Reykjavik, Iceland. [11]Department of Obstetrics and Gynecology, Landspitali University Hospital, Reykjavik, Iceland. [12]Department of Pathology, Landspitali University Hospital, Reykjavik, Iceland. [13]Icelandic Cancer Registry, Reykjavik, Iceland. [14]Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [15]Department of Genetics and Molecular Medicine, Landspitali University Hospital, Reykjavik, Iceland. [16]Department of Pediatrics, Section of Child Neurology, The Children's Hospital of Reykjavik, Landspitali University Hospital, Reykjavik, Iceland. [17]Icelandic Medical Center (Laeknasetrid), Laboratory in Mjodd (RAM), Reykjavik, Iceland. [18]Department of Clinical Biochemistry, Akureyri Hospital, Akureyri, Iceland. [19]Department of Clinical Biochemistry, Landspitali University Hospital, Reykjavik, Iceland. [20]Department of Anthropology, University of Iceland, Reykjavik, Iceland. [21]These authors contributed equally to this work. Correspondence should be addressed to D.F.G. (daniel.gudbjartsson@decode.is) or K.S. (kari.stefansson@decode.is).

Received 17 February 2014; accepted 13 February 2015; published online 25 March 2015; doi:10.1038/ng.3247

frequency spectra of sequence variants in relation to their annotation. Second, we examine the geographical variation in sequence diversity in Iceland. Third, we show how variants down to a frequency of 0.1% can be imputed into the genomes of individuals who are only genotyped on microarray platforms and how the phenotypes of first- and second-degree relatives can be incorporated into analysis using the genealogy. Finally, we provide three examples of how rare variants in these data can be mined for associations with an extensive set of phenotypes and one example of how these data can be used to analyze clinical problems.

RESULTS

Sequencing, genotype calling and annotation

We have sequenced the whole genomes of 2,636 Icelanders using Illumina technology to a mean depth of at least 10× (median of 20×), including 909 sequenced to a mean depth of at least 30× (**Supplementary Fig. 1** and **Supplementary Tables 1** and **2**). A coverage of at least 1× was achieved for 2.72 Gb and of 10× or greater was achieved for 2.70 Gb for individuals with at least 10× coverage, and a coverage of at least 30× was achieved for 2.35 Gb for individuals with at least 30× coverage. Autosomal SNPs and indels, up to a length of 60 bp, were identified, and their genotypes were called for all samples simultaneously using the Genome Analysis Toolkit (GATK version 2.3.9; **Supplementary Fig. 2**)[17]. We used information about haplotype sharing, taking advantage of the fact that all the sequenced individuals had also been chip typed and undergone long-range phasing (**Supplementary Fig. 3**)[18].

Whole-genome sequencing probes most of the genome and is expected to cover most variant sites, although the data have limitations such as poor coverage of the first exons of genes and GC-rich regions in general, limited resolution of structural polymorphisms (because the read length yielded by the sequencing technology is only 101–120 nt of paired-end reads) and low coverage of problematic regions such as the centromeres.

The effects of the sequence variants on the 19,135 protein-coding genes in the RefSeq database[19] were annotated using the Variant Effect Predictor (VEP)[20]. VEP predicts the consequence of each sequence variant on all neighboring RefSeq genes on the basis of a set of 35 consequence terms defined by the Sequence Ontology (SO) (**Supplementary Table 3**)[21]. We grouped sequence variants into four categories, in order of decreasing severity: (i) loss of function, including stop-gain or -loss variants, frameshift indels, splice donor or acceptor variants and initiator codon variants, (ii) moderate impact, including missense variants, in-frame indels and splice-region variants, (iii) low impact, including synonymous variants and 3′- and 5′-UTR variants; and (iv) other, including deep intronic and intergenic variants. As a measure of the quality of our SNP data, our transition/transversion (Ts/Tv) ratios were similar to previously reported ones[13,22] (**Supplementary Table 4**).

Variant counts by impact

We identified a total of 19,689,642 SNPs and 1,441,572 indels in the 2,636 sequenced Icelanders. The frequency distribution of variants based on functional annotation is shown in **Table 1** and **Supplementary Figure 4**. Although only 7% of the sequence variants were indels, indels were over-represented in the loss-of-function category, owing to their tendency to shift the reading frame when they occur in exons[23,24]. Thus, indels represented 41% of the 6,795 loss-of-function variants but only 1.9% of the 125,542 moderate-impact variants (**Table 1**). Most sequence variants are expected to be rare because of the incessant production of genetic diversity. We found that the fraction of variants with a minor allele frequency (MAF) below 0.1% (corresponding to five or fewer copies of the minor allele in our data) was 61.6%, 46.4%, 37.5% and 36.0% in the loss-of-function, moderate-impact, low-impact and other categories, respectively (**Table 1**).

We examined the length and number of indels by genomic location (**Fig. 1** and **Supplementary Table 5**). Approximately twice as many deletions as insertions were called, which may be because insertions are more difficult to call than deletions. Deletions were longer than insertions. In protein-coding regions, we observed a deficit of deletions and insertions that were not multiples of three. This deficit was stronger among insertions, and the deficit was apparent across the range of indel lengths. The difference between the length distributions of coding and noncoding indels is likely the result of negative selection against frameshift indels[20,21]. One consequence of this negative selection is longer indels in protein-coding regions than outside them, primarily because 1- to 2-bp indels were most common and are not multiples of three. In addition, the deficit of insertions relative to deletions was greater in protein-coding regions than outside them because a higher proportion of insertions were not multiples of three. Curiously, in noncoding regions, there was an excess of indels that were multiples of four, which we speculate is more likely due to the mechanism of mutation than to selection.

In clinical sequencing or the study of mendelian traits, the goal is to find a single causative genotype. Every individual is either heterozygous or homozygous for the minor allele at a large number of positions in the genome that have to be accounted for in the search for causative variants (**Table 2**). For mendelian diseases, the causative variants are most often loss-of-function or moderate-impact mutations[25]. The population prevalence of mendelian traits constrains the frequency of highly penetrant mutations. Consequently, it is crucial to filter candidate genotypes on the basis of reliable estimates of

Table 1 Number of indels and SNPs by impact group and MAF

		Loss of function	Moderate impact	Low impact	Other	
Type	MAF	Frameshift indel, splice acceptor or donor, stop gain or loss, initiator codon	In-frame indel, missense, splice region	Synonymous, stop retained, 3′ or 5′ UTR	Intronic, intergenic	Total
SNP	≥0.5%	602 (0.0070%)	36,282 (0.42%)	108,850 (1.3%)	8,445,855 (98.3%)	8,591,589
	0.1–0.5%	915 (0.023%)	29,659 (0.76%)	59,076 (1.5%)	3,836,528 (97.7%)	3,926,178
	<0.1%	2,462 (0.034%)	57,209 (0.80%)	101,751 (1.4%)	7,010,453 (97.7%)	7,171,875
	All	3,979 (0.020%)	123,150 (0.63%)	269,677 (1.4%)	19,292,836 (98.0%)	19,689,642
Indel	≥0.5%	418 (0.0609%)	797 (0.12%)	8,352 (1.2%)	676,820 (98.6%)	686,387
	0.1–0.5%	677 (0.229%)	568 (0.19%)	4,380 (1.5%)	290,492 (98.1%)	296,117
	<0.1%	1,721 (0.375%)	1,027 (0.22%)	6,757 (1.5%)	449,563 (97.9%)	459,068
	All	2,816 (0.195%)	2,392 (0.17%)	19,489 (1.4%)	1,416,875 (98.3%)	1,441,572

Percentages are for the proportion of variants that fall in each class within the row.

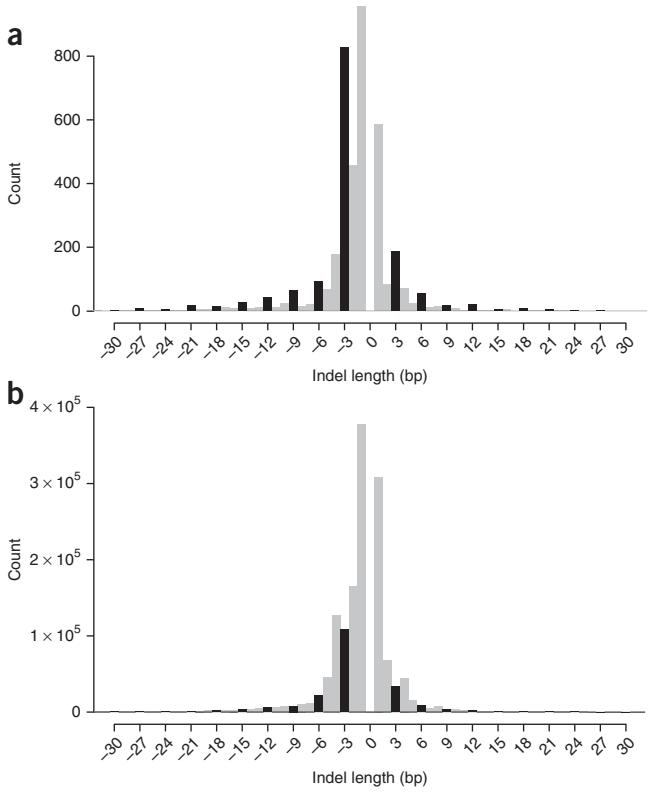

Figure 1 Distribution of indel lengths inside and outside protein-coding regions. (**a**) The 4,001 indels inside protein-coding regions. (**b**) The 1,437,571 indels outside protein-coding regions. Insertions have a positive length, and deletions have a negative length. Indels that are not a multiple of three are colored gray. Indels that are a multiple of three are colored black.

frequency (**Table 2**). The distribution of frequencies for loss-of-function variants illustrates this point nicely. Our sequenced individuals carried, on average, 149 loss-of-function variants, of which 1.4 were only seen in 1 or 2 of the 2,636 sequenced individuals (MAF < 0.04%) and are thus likely candidates for dominant determinants of rare traits. In the context of rare recessive traits, we note that only 1 in 12 individuals were homozygous for a loss-of-function variant with MAF <2%, a threshold that under Hardy-Weinberg equilibrium would correspond to 1 in 2,500 individuals being homozygous.

Assessing selection through variant density and frequency

The density and frequency of variants in genes and genomic regions provide some information about the nature and strength of selection acting on them[26,27]. Specifically, as the strength of negative selection increases, we expect a greater fraction of rare variants (FRV), defined here as variants with derived allele frequency (DAF) < 0.5%, and a lower density of variants, that is, fewer variants per unit of sequence length. In neutral regions of the genome, the FRV and density of variants are a function of the demographic history of the population being studied. We note that, although both measures are sensitive to sample size, increasing with the number of individuals sequenced, they are informative about the relative magnitude of negative selection acting on regions of the genome when evaluated in the same set of individuals. Moreover, both measures are affected by the sequencing method employed and the quality filters used when calling sequence variants. The coverage of our sequence data was not uniform across the genome, which affected the number of sequence variants detected, in particular for very rare variants. In the analyses of FRV and variant density, we therefore only considered positions with an average coverage between 15× and 30× (corresponding to a total of 2.47 Gb and 97% of variants). Furthermore, we restricted the calculation of FRV to variants for which ancestral status could be inferred from the Ensembl Compara ancestral sequences for *Homo sapiens* using multiple-sequence alignments of six primates[28] (corresponding to 87% of variants), as a putative ancestral allele is required for estimation of the DAF.

A detailed breakdown of FRV by the variant annotation categories for SNPs and indels relative to the genome-wide averages of 61.5% and 57.5%, respectively, is shown in **Figure 2a**. Overall, our results are consistent with previous reports[26], such that variants with a greater impact on gene products tended to have a higher FRV. The FRV was lower for indels than for SNPs for all annotations except loss of function. This may be partly accounted for by the fact that rare indels are more likely to be missed than rare SNPs because indels are more difficult to call. Loss-of-function variants were by far the rarest, followed by moderate-impact variants. SNPs were much denser than indels, but the relative densities of both types of variants varied substantially across regions (**Fig. 2b**). We found the lowest density of variants in splicing and coding regions. In addition, indels were denser in UTRs and upstream or downstream regions than in intergenic regions, whereas SNPs were most dense in intergenic regions.

Table 2 Mean genotype counts per individual by frequency and variant impact

	Homozygous minor		Heterozygous	
Impact	MAF < 2%	All	MAF < 0.04%	All
Loss of function	0.083	21	1.4	128
Moderate impact	2.20	1,138	17.9	5,996
Low impact	5.79	4,455	32.1	21,984
Other	404.08	381,041	2,220	1,829,991

The MAF threshold of 0.04% for heterozygotes corresponds to the minor allele only being carried in one or two sequenced individuals. The MAF threshold of 2% for homozygotes corresponds to 1 in 2,500 individuals being homozygotes under Hardy-Weinberg equilibrium.

Variants are not evenly distributed within genes. In particular, a greater number of loss-of-function variants have been observed toward the 3′ and 5′ ends of coding sequences[29]. The FRV and variant density are shown as a function of variant position within the gene and impact on gene products (loss of function, moderate impact and low impact) in **Figure 3**. In multi-exon genes, we observed a greater FRV for loss-of-function and moderate-impact variants in middle exons than in first and last exons (**Fig. 3b**). This indication of stronger negative selection on middle exons is supported by the finding of a lower density of these variants in middle exons (5.8 variants/kb) than in first and last exons (both 6.3 variants/kb) (**Fig. 3c**,**d**). Less negative selection pressure on the first exon could be due to some genes having alternative first exons[30]. Also, misidentification of the ATG start codon would lead to erroneous annotation of sequence variants in the first exon as being of moderate impact or loss of function rather than as 5′-UTR variants. There is a reason to expect a relaxation of negative selection pressure on loss-of-function variants in the last exon as these variants are less likely to lead to nonsense-mediated decay than those occurring in preceding exons[31].

A large fraction of intronless genes encode olfactory receptors (17%) that have been reported to harbor more loss-of-function mutations than any other class of genes[29]. In our data, the FRVs for all three categories of variants in olfactory genes were below the genome-wide average

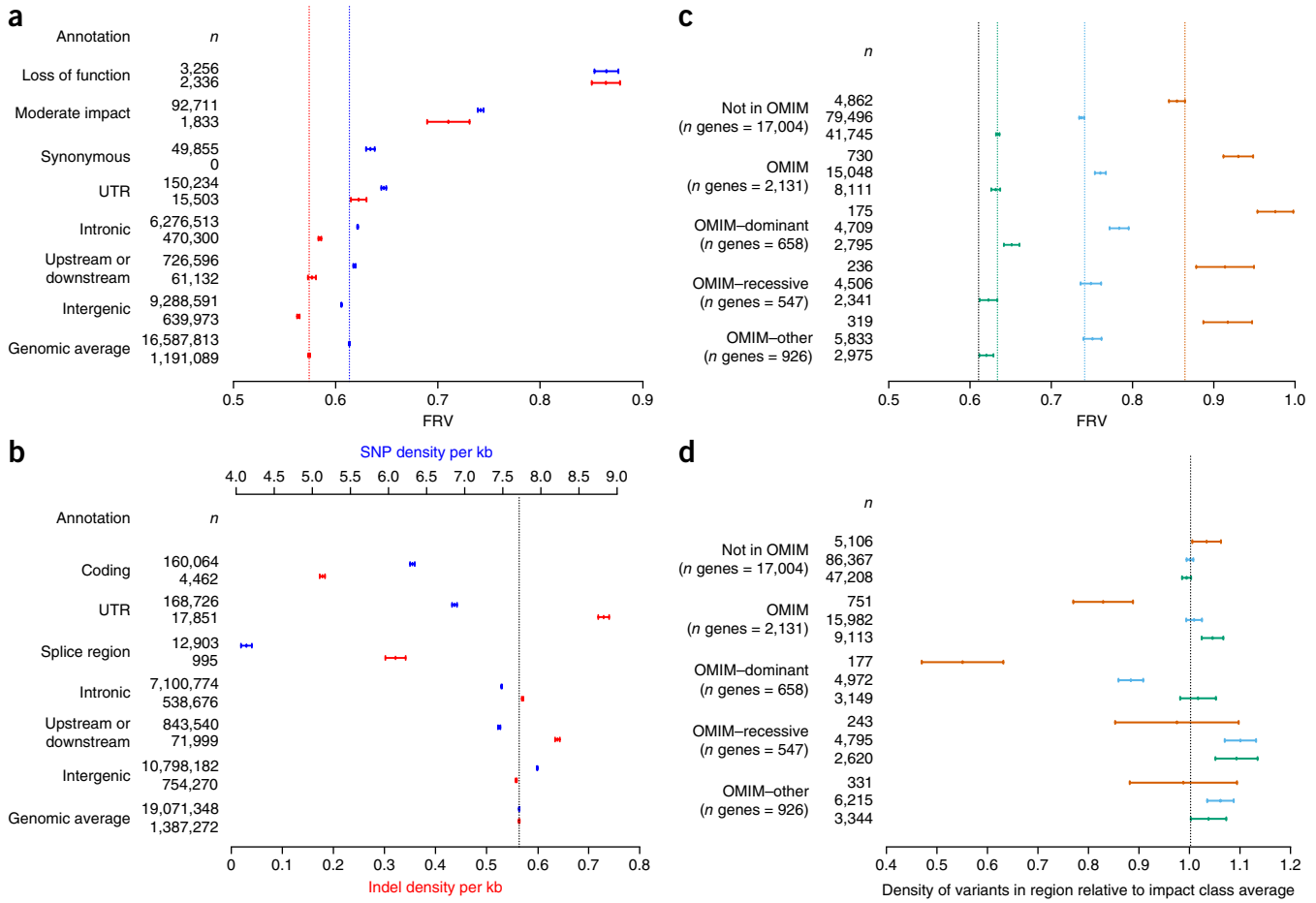

Figure 2 FRV and variant density by impact class and OMIM disease-related gene classification. (**a**) FRV by annotation. (**b**) Variant density. SNPs are shown in blue, and indels are shown in red. (**c**) FRV by OMIM disease gene classification and impact class. (**d**) Variant density relative to the impact class average by OMIM disease-related gene classification and impact class. Loss-of-function, moderate-impact and low-impact variants are shown in red, blue and green, respectively. The line segments indicate the 95% confidence interval around each observed FRV or variant density. The dotted lines indicate the genomic average FRV or variant density.

of 61%, and the normalized density of variants was unusually high (9.4 variants/kb compared to the genomic average of 8.3 variants/kb). The excess of common loss-of-function and moderate-impact variants in olfactory genes indicates not just an unusually low level of purifying selection but also the effect of positive selection on variants that affect olfactory perception[32,33]. The non-olfactory intronless genes had drastically different FRV and density patterns that were similar to those of the first and last exons of multi-exon genes.

The Online Mendelian Inheritance in Man (OMIM) database is a catalog of human genes and mutations with an emphasis on rare traits and highly penetrant mutations[25]. The FRVs and variant densities for three classes of variants (loss of function, moderate impact and low impact) in OMIM genes that have been linked to disease are shown in **Figure 2**. As expected, the FRV was always highest for loss-of-function variants and lowest for low-impact variants (**Fig. 2c**). However, the FRV was greater for loss-of-function and moderate-impact variants in the 2,131 OMIM disease-related genes than in the 17,004 other RefSeq genes. Moreover, the density of loss-of-function variants was substantially lower in the OMIM disease-related genes than in other RefSeq genes (**Fig. 2d**). A further division of OMIM disease-related genes into subgroups on the basis of the mode of inheritance of the traits they influence[27] showed that the 658 dominant-mode OMIM genes harbored the greatest FRVs for all 3 classes of variants and a lower density of moderate-impact and loss-of-function variants. These results are consistent with stronger negative selection acting on sequence variants in OMIM genes and in particular on variants that affect traits through a dominant mode of inheritance.

The conservation of sequence among mammalian species reflects the magnitude of purifying selection since their divergence[34]. The Genomic Evolutionary Rate Profiling (GERP) score is a measure of the sequence conservation between mammalian species. A positive GERP score indicates that a site may be under purifying selection or subject to a below-average mutation rate, whereas a negative score may indicate weaker purifying selection or an above-average mutation rate[34]. The relationship between GERP scores estimated using 33 mammalian species (excluding humans) and the FRV and density of SNPs in our data by annotation category is shown in **Figure 4a**,**b**. For SNPs with negative GERP scores, we saw no correlation between the score and the FRV. In contrast, for SNPs with positive GERP scores, we saw a marked positive correlation between the score and the FRV. Synonymous SNPs were the exception among coding SNPs, in that they had very similar FRVs regardless of GERP score. The difference in the range of GERP scores was striking; nearly all splice donor and acceptor site variants had positive GERP scores, whereas synonymous and splice-region variants could have very negative scores. The relationship between GERP scores and SNP density was much simpler: the

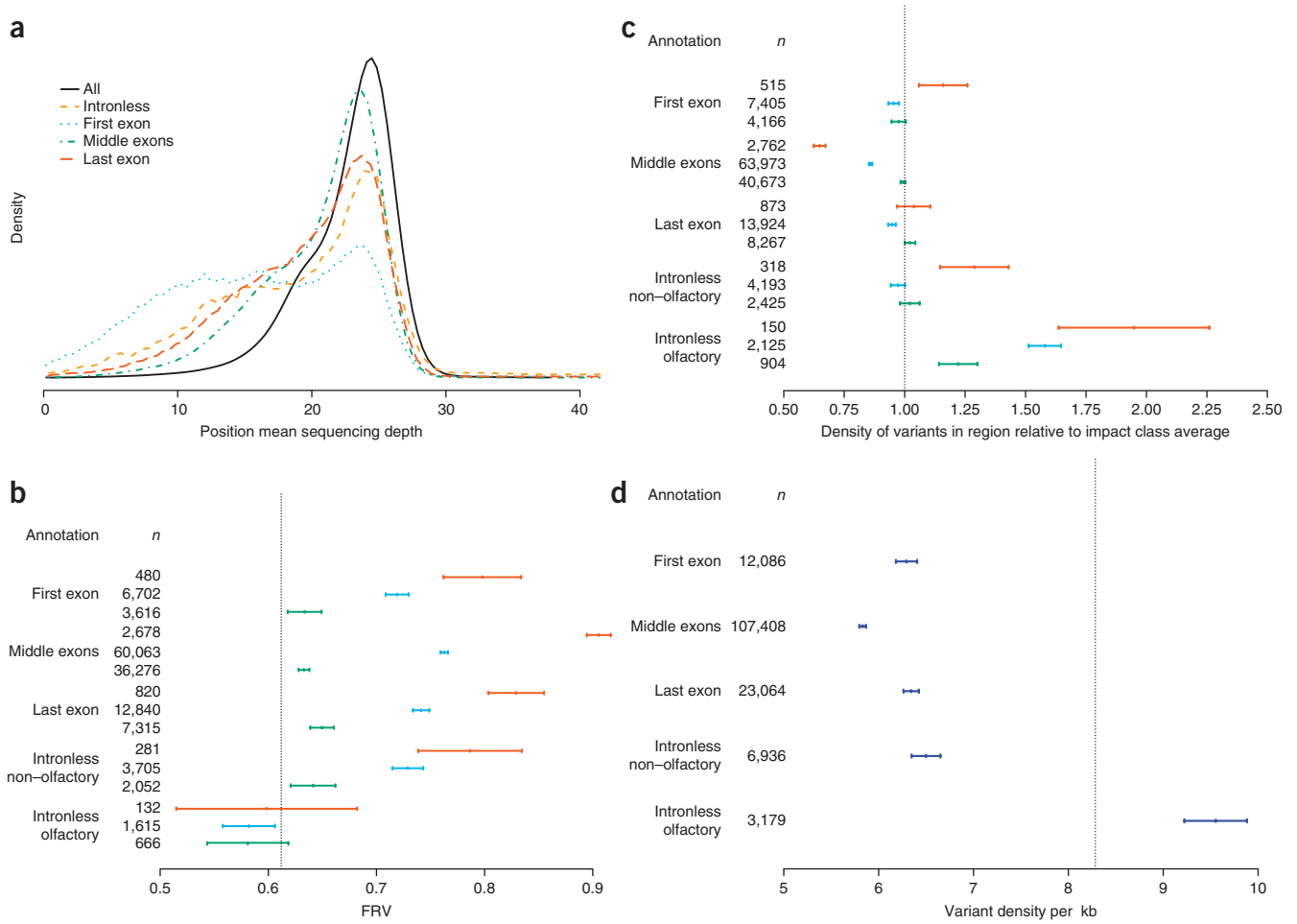

Figure 3 Sequencing coverage, FRV and variant density by exon rank. (**a**) Distribution of the mean coverage by position for the whole genome, intronless genes, and the first, middle and last exons of multi-exon genes among the 2,636 whole genome–sequenced Icelanders. (**b**) FRV by exon rank and impact class. (**c**) Variant density relative to the impact class average by exon rank and impact class. (**d**) Variant density by exon rank. Loss-of-function, moderate-impact and low-impact variants are shown in red, blue and green, respectively, in **b** and **c**. The line segments indicate the 95% confidence interval around each observed FRV or variant density. The dotted lines indicate the genomic average FRV or variant density.

higher the GERP score, the lower the density. Across the range of its values, GERP score was a good predictor of SNP density and a better predictor overall of density than sequence annotation. Comparing the sequences of humans and chimpanzees suggests that the mutation rates are similar in the two species on a local scale[35]. The strong correlation between GERP score and SNP density in our data, even for negative GERP scores, shows that GERP scores predict mutation rates in humans. In turn, the GERP score itself must be strongly correlated with mammalian mutation rates.

The Gene Ontology (GO) project provides a classification of genes by molecular function, cellular component and biological process[36]. Using the subset of GO terms from the PANTHER classification system[37], we examined the FRV and density within each of 307 GO categories containing more than 50 genes (**Fig. 4c**). Of the 19,135 RefSeq genes, 17,427 had a GO classification. Genes with GO classifications had greater FRV scores (FRV = 71.3%) but a lower density of variants (6.5 variants/kb) than those with no GO classification (FRV = 68.0%, 6.8 variants/kb). We removed olfactory genes from other GO classes because of their extremely high variant density and low FRVs[29]. Differences between GO classes were greater for variant density than for FRV. We combined the FRV and variant density to search for outlying GO classes with either high diversity (high density and low FRV) or low diversity (low density and high FRV). Interestingly, many of the high-diversity GO classes were linked to the communication of cells with their environment, such as sensory, extracellular matrix, cell adhesion and defense response to bacteria. In addition to the already discussed outlier of olfactory receptor genes (9.4 variants/kb, FRV = 59.1%), other sensory GO classes with relatively high diversity included cilium (involved in chemosensation, thermosensation and mechanosensation) (7.3 variants/kb, FRV = 71.4%), visual perception (7.6 variants/kb, FRV = 72.5%) and sensory perception of chemical stimulus (7.7 variants/kb, FRV = 63.7%), which are primarily taste receptors after our exclusion of olfactory receptors. In contrast, the low-diversity GO classes were involved in functions of the nucleus and the cytosol, including, for example, genes that tend to be linked to DNA replication and repair, RNA transcription, the nucleolus, cell cycle regulation, organelle organization and biogenesis.

Chromatin profiling has been used to estimate regulatory activity on the basis of Encyclopedia of DNA Elements (ENCODE) data[38]. The variant densities and FRVs for 13 different chromatin states, averaged over 9 cell types[38], are shown in **Figure 4d**. All 13 classes fell within the line segment between the values corresponding to intergenic regions and UTRs. The three promoter and three transcription chromatin states had similar FRVs and variant densities as UTRs. The FRVs and variant

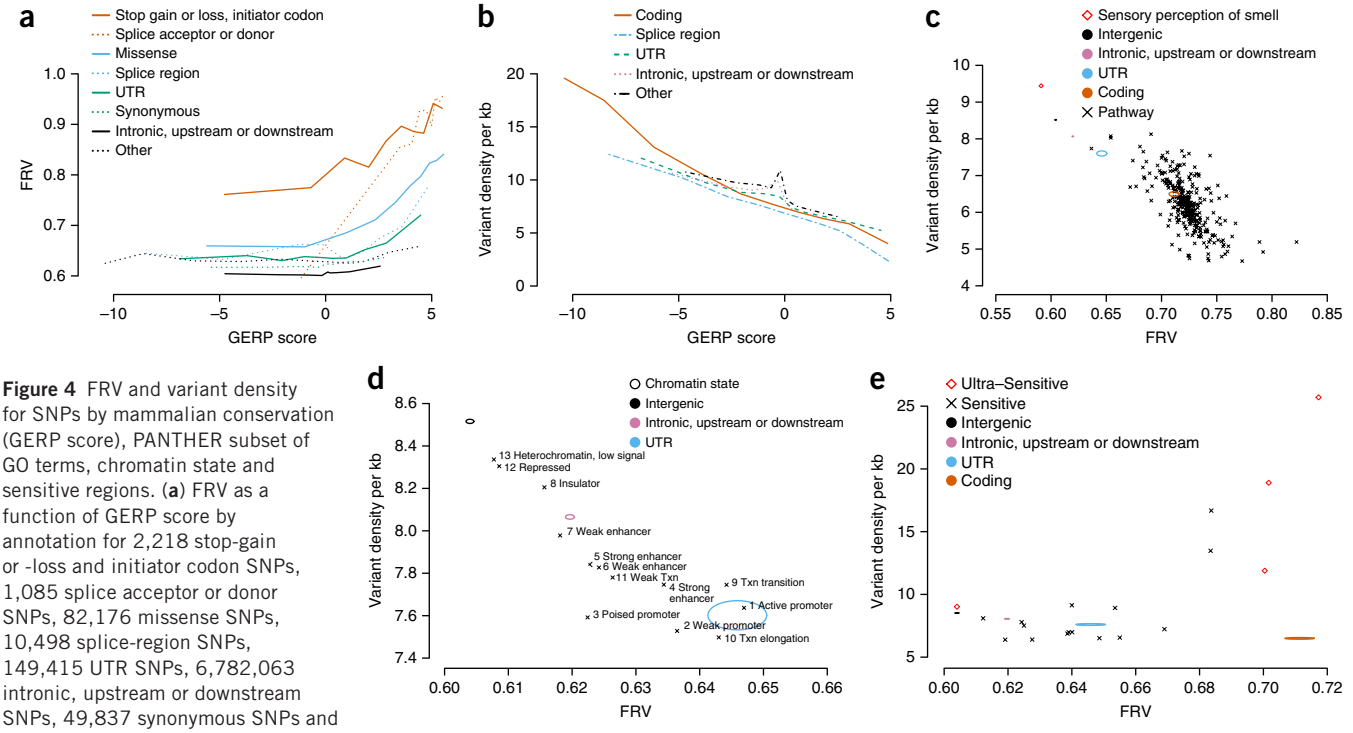

Figure 4 FRV and variant density for SNPs by mammalian conservation (GERP score), PANTHER subset of GO terms, chromatin state and sensitive regions. (**a**) FRV as a function of GERP score by annotation for 2,218 stop-gain or -loss and initiator codon SNPs, 1,085 splice acceptor or donor SNPs, 82,176 missense SNPs, 10,498 splice-region SNPs, 149,415 UTR SNPs, 6,782,063 intronic, upstream or downstream SNPs, 49,837 synonymous SNPs and 8,663,577 other SNPs. (**b**) Variant density as a function of GERP score by annotation for coding regions spanning 25.4 Mb, splice regions spanning 2.5 Mb, UTR regions spanning 24.2 Mb, intronic, upstream or downstream regions spanning 984.4 Mb and other regions spanning 1,242.7 Mb. (**c**) A red diamond represents the sensory perception of smell class, which contains the olfactory receptors, and black X's represent the remaining 307 GO classes with more than 50 gene members, after the removal of the olfactory receptor genes. (**d**) The 13 chromatin states are indicated by black X's. Txn, transcription. (**e**) The 15 sensitive and 4 ultra-sensitive regions are indicated by black X's and red diamonds, respectively. For reference, the 95% confidence ellipses for annotated regions are also shown.

densities of the four enhancer chromatin states fell between those for the intronic, upstream or downstream regions and UTRs.

Recently, ENCODE data[39] were combined with frequency data from the 1000 Genomes Project[16] to search for sensitive and ultra-sensitive noncoding regions that might have an impact on the transcription of genes[26]. Consistent with this classification, we observed a higher FRV among the sensitive and ultra-sensitive regions than for other noncoding regions (**Fig. 4e**). Interestingly, variant density in the ultra-sensitive regions was much higher than in any other annotated region of the genome. Thus, ultra-sensitive regions, by having a high FRV and high density of variants, deviate from the expected behavior of regions under purifying selection, such as coding regions, which have a relatively high FRV and a low density of variants.

Imputation of variants

Imputation of untyped variants into the mix of typed variants is now routine in human genetics[40]. These imputations are almost always based on local linkage disequilibrium (LD) and work well for common variants, but they are not reliable for low-frequency variants and rarely work for rare variants. The long-range phasing of 104,220 Icelanders genotyped for 676,913 autosomal SNPs using Illumina chips (**Supplementary Tables 6** and **7**) increases imputation accuracy and speed by removing uncertainty in phasing. Of variants with a MAF over 0.1%, 99.5% were imputed with information over 0.8. The concordance for 28,204 chip-typed SNPs, which were not part of the long-range phasing set, was high (98.4% of SNPs with DAF >1% were imputed accurately ($r^2 > 0.9$); **Supplementary Fig. 5**). Results are shown for 13 rare sequence variants, previously known to associate with diseases and traits in the Icelandic population on the basis of this imputation scheme in **Supplementary Table 8**.

Geographical ancestry

It is reasonable to assume that most often the sequence variants found in our data and other populations are genuine. There are, however, variants in our data that have not been reported outside of Iceland for one or more of the following reasons: they are not genuine, they do not exist outside of Iceland, they are yet to be identified outside of Iceland or their recording is inconsistent. A comparison of our variants to those in dbSNP[41] and ESP[14,15] showed that essentially all common SNPs found in Iceland have been recorded in dbSNP (99.8% of SNPs with DAF >2%) (**Fig. 5a**). Rare coding SNPs are recorded substantially more often in dbSNP than noncoding ones, probably because more effort has been spent on analyzing coding regions. By design, ESP contains primarily variants from coding regions. Nonetheless, 2.9% and 1.6% of common (DAF > 2%) missense and synonymous SNPs found in Iceland, respectively, were missing from ESP (**Fig. 5b**). These variants are likely genuine because they are present in dbSNP. Indels are more difficult to call than SNPs, and their position can be ambiguous. A substantial fraction of the common indels found in Iceland were not present in either dbSNP or ESP (**Fig. 5c**). Some common indels in coding regions were missing from dbSNP but were present in ESP and vice versa, demonstrating that the recording of indels in these databases is incomplete or inconsistent. Moreover, few of the rare indels found in Iceland have been recorded in dbSNP. In coding regions, SNPs and indels with DAF <0.5% had 60% and 20% chances of being present in at least one of the databases, respectively. Interestingly, rare loss-of-function and moderate-impact SNPs found in our data were

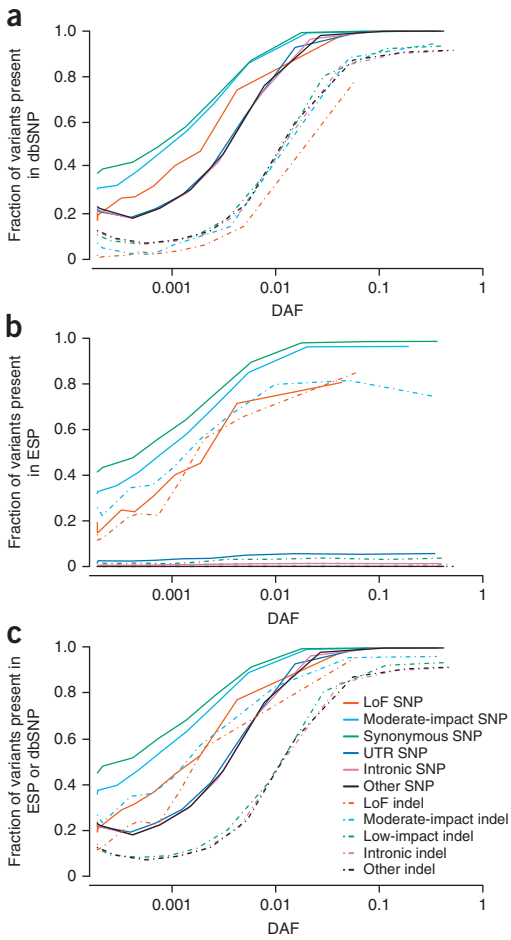

Figure 5 The fraction of SNPs and indels identified in 2,636 Icelanders present in dbSNP and ESP by consequence. The analysis was restricted to 16,587,813 SNPs and 1,191,089 indels for which the ancestral allele could be inferred. (**a**–**c**) Shown is the overlap with dbSNP only (**a**), ESP only (**b**) and the union of dbSNP and ESP (**c**) as a function of DAF by annotation and variant type. LoF, loss of function.

less likely to be present in dbSNP or ESP than synonymous SNPs of the same frequency. Assuming that there is no reporting bias against loss-of-function and moderate-impact SNPs in these databases, it follows that these variants either tend to be younger than synonymous SNPs of the same allele frequency in Iceland or that they have been subject to stronger negative selection in other populations.

We compared counts of protein-coding SNPs by frequency in our Icelandic whole-genome sequence data and the European-American portion of the ESP exome sequence data (n = 4,300) (**Fig. 6a** and **Supplementary Table 9**). We restricted the comparison to regions where both studies had good coverage and excluded repeat regions. We further restricted the analysis to SNPs with a MAF over 0.1% in each population, corresponding to seeing more than six copies of the variant in Iceland or more than nine copies of the variant in ESP, as the number of variants of lower frequency depends heavily on the sequencing and calling method used and the sample size. There were substantially more stop-gain variants in Iceland with a MAF between 0.1% and 1% (relative excess in Iceland = 56%). This difference was not apparent for more common variants. There was a similar but weaker pattern for missense SNPs with a MAF over 0.2% and a still weaker pattern for synonymous SNPs. These findings are consistent with the hypothesis that some deleterious sequence variants are able to reach a higher frequency in a small, isolated population such as Iceland's than in a large outbred population owing to the small size of the Icelandic population and the founder effect[42]. Examples of such variants are the frameshift deletion in *BRCA2* associating with breast and other cancers (MAF = 0.4% in Iceland, almost absent elsewhere)[43], the frameshift deletion in *BRIP1* that associates with ovarian cancer (MAF = 0.4% in Iceland, not seen elsewhere)[5], the stop-gain mutation in *LGR4* that associates with bone mineral density (MAF = 0.2% in Iceland, not seen elsewhere)[7] and the missense mutation in *ALDH16A1* that associates with gout (MAF = 1.7% in Iceland, MAF in other European populations between 0.05% and 0.4%)[3].

We assessed the number of homozygotes for the minor allele relative to the expected number under Hardy-Weinberg equilibrium as a function of MAF (**Fig. 6b**). There was a substantial excess of homozygotes for the minor allele, with rarer variants having greater excess. The excess was greater for the offspring of parents from the same Icelandic county and lower for the offspring of parents from different counties. This excess of homozygotes for the minor allele is therefore, at least in part, driven by geographically stratified patterns of mating in Iceland[44].

Mining the sequence data for discovery and clinical purposes

We tested all sequence variants identified through our whole-genome sequencing project with an imputation information value above 0.8 for association using an additive regression model. Generalized linear regression was employed for continuous traits, and logistic regression was used for case-control analysis (n tests = 16,793,181). Sequence variants with MAF >0.3% were also tested for association under a recessive mode of inheritance (n tests = 10,327,453). The recessive test requires homozygous carriers, and rarer variants will not have a sufficient number of these for association to be detectable.

The effect of some sequence variants on phenotypes depends on the parental origin of the variants[45–48]. The Icelandic genealogy coupled with the large fraction of the population that has been chip typed allows the accurate determination of the parent of origin for the genotypes of all chip-typed individuals, regardless of age[46]. We performed parent of origin–specific association analysis of all variants identified through whole-genome sequencing that were within 100 kb of genes known or predicted to be imprinted according to the Geneimprint database[49]. This amounted to 316,241 variants spanning a total of 47.5 Mb of the human genome.

In total, we performed 27,728,712 tests corresponding to a Bonferroni significance threshold of 1.8×10^{-9}.

MYL4 and early-onset atrial fibrillation

Atrial fibrillation is the most common sustained cardiac arrhythmia of man[50]. According to the Framingham Heart Study, the risk of early-onset atrial fibrillation (diagnosed before the age of 60 years) is 3% for males and 1% for females[51]. We performed a GWAS of early-onset atrial fibrillation based on hospital discharge diagnoses (n = 1,294). Several sequence variants in LD on chromosome 17 (39.5–43.2 Mb, NCBI Build 36) associated significantly with the disease under a recessive model (**Supplementary Table 10**). One of these was a frameshift deletion in *MYL4* (c.234delC, p.Cys78Trpfs*29, MAF = 0.65%, recessive odds ratio (OR) = 110.3 and $P = 5.2 \times 10^{-10}$) (**Fig. 6c**). *MYL4* encodes myosin essential light chain that is found in embryonic muscle[52] and adult atria[53] and has not previously been associated with disease.

We identified eight homozygous carriers of c.234delC among the chip-typed Icelanders and their close relatives for whom we performed detailed review of their medical records (**Supplementary Figs. 6** and **7**).

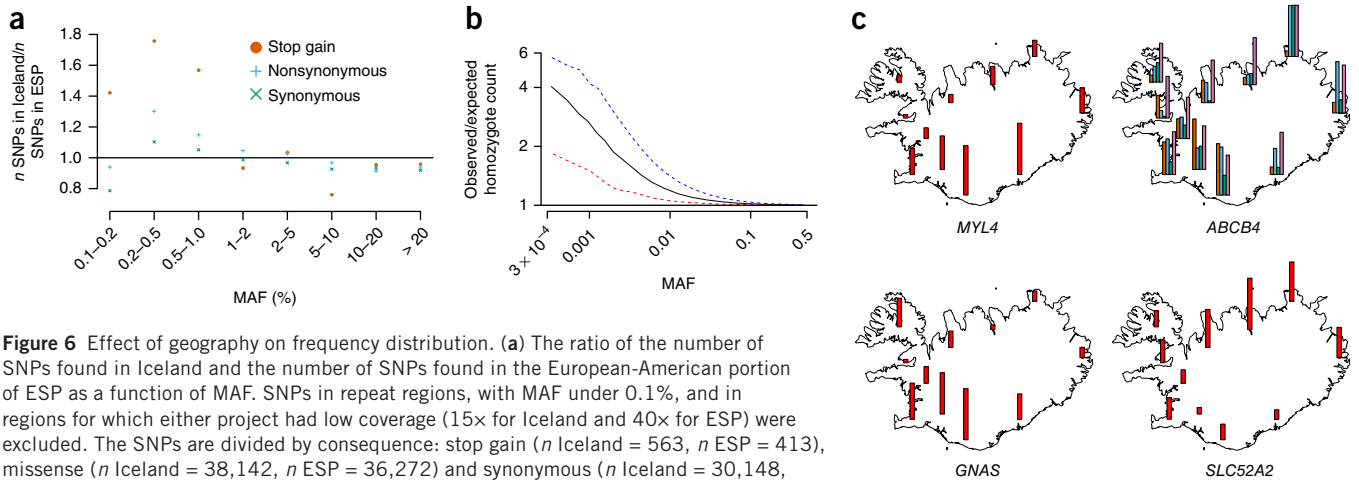

Figure 6 Effect of geography on frequency distribution. (**a**) The ratio of the number of SNPs found in Iceland and the number of SNPs found in the European-American portion of ESP as a function of MAF. SNPs in repeat regions, with MAF under 0.1%, and in regions for which either project had low coverage (15× for Iceland and 40× for ESP) were excluded. The SNPs are divided by consequence: stop gain (*n* Iceland = 563, *n* ESP = 413), missense (*n* Iceland = 38,142, *n* ESP = 36,272) and synonymous (*n* Iceland = 30,148, *n* ESP = 31,765). All counts are shown in **Supplementary Table 9**. (**b**) The ratio of observed and expected minor allele homozygote counts as a function of MAF. The black, blue and red lines represent all chip-typed Icelanders, the chip-typed offspring of parents from the same Icelandic county and the chip-typed offspring of parents coming from different Icelandic counties, respectively. The expected homozygote counts were calculated assuming Hardy-Weinberg equilibrium. (**c**) The geographical distribution of the minor alleles of the risk-conferring variants in *MYL4*, *ABCB4*, *GNAS* and *SLC52A2*, in 104,220 chip-typed Icelanders. Each bar shows the allelic frequency of the variant relative to the geographical region with the highest frequency.

Two homozygous carriers had not received a hospital discharge diagnosis of atrial fibrillation but had been diagnosed with this disease before the initiation of electronic documentation of discharge diagnoses in 1982. Also, five of the six homozygous carriers with a discharge diagnosis of atrial fibrillation had an earlier first date of diagnosis with atrial fibrillation from a detailed further review of their history than in the available discharge diagnoses. All eight homozygous carriers of the *MYL4* c.234delC frameshift deletion had thus been diagnosed with early-onset atrial fibrillation (**Supplementary Table 11**). This highlights the importance of a close collaboration with physicians for the extensive phenotyping of carriers of highly penetrant genotypes.

ABCB4 and liver diseases and function

Our GWAS of gallstone disease (n = 8,258) led us to a missense SNP, encoding p.Gly622Glu, and a frameshift insertion, encoding p.Leu445Glyfs*22, in *ABCB4* (**Supplementary Table 12**). Both variants were rare (MAF = 0.22% and 0.21%) with large effects (allelic OR = 2.74 and 3.10, $P = 7.2 \times 10^{-10}$ and 2.6×10^{-12}, respectively). Rare coding variants in *ABCB4* have been associated with progressive familial intrahepatic cholestasis through a recessive mode of inheritance[54,55] and with intrahepatic cholestasis of pregnancy and low phospholipid–associated cholelithiasis through a dominant mode of inheritance[56].

In addition to being associated with increased risk of gallstone disease, the *ABCB4* mutations were also associated with cholestasis in pregnancy, liver, gallbladder and gallways cancer, liver cirrhosis and serum levels of liver-related biomarkers, including alanine transaminase, aspartate transaminase and γ-glutamyl transpeptidase (**Supplementary Table 13**). The most significant association of all four variants was with alanine transaminase. All four *ABCB4* variants affected several liver blood tests and diseases. However, patterns of association with the variants differed both in magnitude and nature. Being able to assess all these phenotypes in the same population allows direct comparison between the variants. Sequencing the entire region around the *ABCB4* gene allowed a thorough analysis of the variants affecting the gene and makes it unlikely that additional variants with substantial effects on liver phenotypes have been missed.

Maternally inherited *GNAS* allele and TSH

Following our earlier publication[57], we performed a GWAS on 61,397 Icelanders with thyroid-stimulating hormone (TSH) measurements. In addition to observing associations under the additive model[57], we detected an association between TSH levels and the maternally inherited alleles of rs139242164[T] (maternal allele effect = 0.298 s.d., $P = 1.3 \times 10^{-12}$, MAF = 0.44%; **Supplementary Table 14**) within the first intron of the long transcript of *GNAS*.

In contrast to the increased TSH levels that are associated with maternally inherited copies of rs139242164[T], paternally inherited copies associated with decreased TSH levels (paternal allele effect = −0.105 s.d., P = 0.014). As a result, the additive model was far from reaching the significance level required for detection in the GWAS (P = 0.0017).

The main limitation of exome sequencing is that any association signal not captured by exonic variants will be missed, such as the deep intronic variants in *GNAS* described above and the association of an intronic variant in *CCND2* with type 2 diabetes mellitus[12], despite these variants having a large effect.

Clinical sequencing of *SLC52A2* in BVVL syndrome

We sequenced the genomes of two Icelandic sisters with Brown-Vialetto–Van Laere (BVVL) syndrome, a rare neurologic condition[58], and their unaffected parents to a depth of at least 34× (**Supplementary Fig. 8** and **Supplementary Table 15**).

We used a series of constraints based on annotation and allele frequency and assumed a recessive mode of inheritance to search for the most likely causative variants[59] (**Supplementary Fig. 9**), through which we were left with a single missense mutation (encoding p.Leu339Pro, MAF = 0.31%) in the riboflavin transporter gene *SLC52A2*. After these Icelandic cases of BVVL syndrome were resolved, other mutations in *SLC52A2* have been found to cause BVVL syndrome[60–62], and a variant encoding p.Leu339Pro has been described in a non-Icelandic BVVL syndrome case in a compound heterozygous state[62].

Follow-up analysis based on identifying other cases with similar clinical presentations and examining the offspring of carrier parents identified two additional BVVL syndrome cases. All four Icelandic cases suffered from a loss of vision, and three had vestibular ataxia

(P = 0.00043 and 0.0025, respectively). Our review of the literature found 77 BVVL cases, 9 with loss of vision[58,63–70].

DISCUSSION

We have sequenced the whole genomes of a large group of Icelanders, providing a comprehensive understanding of the structure of the Icelandic population and the basis to use powerful imputation to discover associations between variants in sequence and phenotypes. The density and frequency distributions of sequence variants allow us to evaluate sequence and gene annotations. There are fewer very rare variants in a small isolated population, such as the Icelandic one, than in larger outbred populations. However, the small size of the population and the founder effect also allow deleterious variants to reach higher frequencies.

Rare sequence variants tend to have a more recent origin than common ones and thus require the set of sequenced individuals to be more closely related to the set of imputed individuls. Our extended sampling of the Icelandic population allows us to carry out long-range phasing and identify the long haplotypes needed to impute recent sequence variants. Having sequenced and phased the genomes of 2,636 Icelanders thus enables us to accurately impute sequence variants down to an allelic frequency of 0.1%. The resultant large set of imputed genomes can then be used for powerful tests of association with an extensive range of traits and phenotypes. In addition to the additive model, association can be tested by parent of origin and under the recessive model. In combination, these data provide the foundation for a formidable study design that can be used to help determine how variation in the sequence of the human genome gives rise to human diversity.

URLs. Genome Analysis Toolkit (GATK) Best-Practice Variant Detection with GATK v4, for release 2.0 (accessed July 2012), https://www.broadinstitute.org/gatk/guide/best-practices; Picard version 1.55, http://broadinstitute.github.io/picard/; dbSNP (Build 137), http://www.ncbi.nlm.nih.gov/SNP/; Human Phenotype Ontology (HPO; Build 32) (accessed June 2012), http://human-phenotype-ontology.org/; NHLBI Exome Sequencing Project (ESP) (accessed October 2013), http://evs.gs.washington.edu/EVS/; Online Mendelian Inheritance in Man (OMIM) (accessed 20 January 2014), http://omim.org/; Geneimprint, http://www.geneimprint.com/site/genes-by-species.Homo+sapiens; Mouse Genome Database (MGD) at the Mouse Genome Informatics website (data retrieved October 2012), http://www.informatics.jax.org/; Ensembl Compara ancestral sequences (accessed April 2013): *Homo sapiens* (GRCh37) ancestor sequence, ftp://ftp.ensembl.org/pub/release-65/fasta/ancestral_alleles; multi-way alignments for six primates, ftp://ftp.ensembl.org/pub/release-70/emf/ensembl-compara/epo_6_primate; Icelandic Cancer Registry (ICR), http://www.krabbameinsskra.is/; bx-python (commit 5449537), https://bitbucket.org/james_taylor/bx-python.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. A Data Descriptor is available at *Scientific Data*[71]. The list of variants discovered can be found at the European Variant Archive (PRJEB8636).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all the participants in this study. This study was performed in collaboration with Illumina.

AUTHOR CONTRIBUTIONS

D.F.G., H. Helgason, S.A.G., F.Z., D.O.A., O.T.M., G. Masson, A.H., P.S. and K.S. wrote the initial draft of the manuscript. D.F.G., H. Helgason, S.A.G., F.Z., A.O., G. Magnusson, B.V.H., E.H., G.T.S., S.N.S., M.L.F., A.K., G. Masson and P.S. analyzed the data. D.F.G., H. Helgason, S.A.G., F.Z., A.G., S.B., H.G. and G. Masson created methods for analyzing the data. S.N.S., H. Holm, J.S., H.T.H., H.J. and O.T.M. performed the experiments. H. Holm, G.S., G.T., J.T.S., S.G., G.B.W., T.R., B.T., E.S.B., S.O., H.T., T.S., T.S.G., A.T., J.G.J., A.S., G.B., J.J.J., O.T., P.L., G.I.E., O.S., I.O. and D.O.A. collected the samples and information. D.F.G., D.O.A., G. Masson, U.T., A.H., P.S. and K.S. designed the study.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
2. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
3. Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat. Genet.* **43**, 1127–1130 (2011).
4. Jonsson, T. *et al.* A mutation in *APP* protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).
5. Rafnar, T. *et al.* Mutations in *BRIP1* confer high risk of ovarian cancer. *Nat. Genet.* **43**, 1104–1107 (2011).
6. Holm, H. *et al.* A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
7. Styrkarsdottir, U. *et al.* Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
8. Jonsson, T. *et al.* Variant of *TREM2* associated with the risk of Alzheimer's disease. *N. Engl. J. Med.* **368**, 107–116 (2013).
9. Helgason, H. *et al.* A rare nonsynonymous sequence variant in *C3* is associated with high risk of age-related macular degeneration. *Nat. Genet.* **45**, 1371–1374 (2013).
10. Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.* **44**, 1326–1329 (2012).
11. Stacey, S.N. *et al.* A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43**, 1098–1103 (2011).
12. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
13. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
14. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
15. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
16. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
17. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
18. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
19. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
20. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
21. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
22. Stubbs, A. *et al.* Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection. *J. Clin. Bioinforma* **2**, 19 (2012).
23. Chen, F.C., Chen, C.J., Li, W.H. & Chuang, T.J. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* **17**, 16–22 (2007).
24. Montgomery, S.B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).

25. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
26. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
27. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
28. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
29. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
30. Zavolan, M. & van Nimwegen, E. The types and prevalence of alternative splice forms. *Curr. Opin. Struct. Biol.* **16**, 362–367 (2006).
31. Baker, K.E. & Parker, R. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.* **16**, 293–299 (2004).
32. Keller, A., Zhuang, H., Chi, Q., Vosshall, L.B. & Matsunami, H. Genetic variation in a human odorant receptor alters odour perception. *Nature* **449**, 468–472 (2007).
33. Mainland, J.D. *et al.* The missense of smell: functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **17**, 114–120 (2014).
34. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
35. Smith, N.G., Webster, M.T. & Ellegren, H. Deterministic mutation rate variation in the human genome. *Genome Res.* **12**, 1350–1356 (2002).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
37. Mi, H., Muruganujan, A. & Thomas, P.D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).
38. Ernst, J., Vainas, O., Harbison, C.T., Simon, I. & Bar-Joseph, Z. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.* **3**, 74 (2007).
39. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
40. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
41. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
42. Mayr, E. *Systematics and the Origin of Species from the Viewpoint of a Zoologist* (Columbia University Press, 1942).
43. Thorlacius, S. *et al.* A single *BRCA2* mutation in male and female breast cancer families from Iceland with varied cancer phenotypes. *Nat. Genet.* **13**, 117–119 (1996).
44. Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J. & Stefansson, K. An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37**, 90–95 (2005).
45. Small, K.S. *et al.* Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011).
46. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
47. Wallace, C. *et al.* The imprinted *DLK1-MEG3* gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. *Nat. Genet.* **42**, 68–71 (2010).
48. Abreu, A.P. *et al.* Central precocious puberty caused by mutations in the imprinted gene *MKRN3*. *N. Engl. J. Med.* **368**, 2467–2475 (2013).
49. Falls, J.G., Pulford, D.J., Wylie, A.A. & Jirtle, R.L. Genomic imprinting: implications for human disease. *Am. J. Pathol.* **154**, 635–647 (1999).
50. Go, A.S. *et al.* Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *J. Am. Med. Assoc.* **285**, 2370–2375 (2001).
51. Lloyd-Jones, D.M. *et al.* Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. *Circulation* **110**, 1042–1046 (2004).
52. Strohman, R.C., Micou-Eastwood, J., Glass, C.A. & Matsuda, R. Human fetal muscle and cultured myotubes derived from it contain a fetal-specific myosin light chain. *Science* **221**, 955–957 (1983).
53. Cohen-Haguenauer, O. *et al.* Chromosomal assignment of two myosin alkali light-chain genes encoding the ventricular/slow skeletal muscle isoform and the atrial/fetal muscle isoform (*MYL3*, *MYL4*). *Hum. Genet.* **81**, 278–282 (1989).
54. Nicolaou, M. *et al.* Canalicular ABC transporters and liver disease. *J. Pathol.* **226**, 300–315 (2012).
55. Davit-Spraul, A., Gonzales, E., Baussan, C. & Jacquemin, E. Progressive familial intrahepatic cholestasis. *Orphanet J. Rare Dis.* **4**, 1 (2009).
56. Dixon, P.H. *et al.* Heterozygous *MDR3* missense mutation associated with intrahepatic cholestasis of pregnancy: evidence for a defect in protein trafficking. *Hum. Mol. Genet.* **9**, 1209–1217 (2000).
57. Gudmundsson, J. *et al.* Discovery of common variants associated with low TSH levels and thyroid cancer risk. *Nat. Genet.* **44**, 319–322 (2012).
58. Sathasivam, S. Brown-Vialetto–Van Laere syndrome. *Orphanet J. Rare Dis.* **3**, 9 (2008).
59. Chan, W.M. *et al.* Expanded polyglutamine domain possesses nuclear export activity which modulates subcellular localization and toxicity of polyQ disease protein via exportin-1. *Hum. Mol. Genet.* **20**, 1738–1750 (2011).
60. Johnson, J.O. *et al.* Exome sequencing reveals riboflavin transporter mutations as a cause of motor neuron disease. *Brain* **135**, 2875–2882 (2012).
61. Ciccolella, M. *et al.* Riboflavin transporter 3 involvement in infantile Brown-Vialetto-Van Laere disease: two novel mutations. *J. Med. Genet.* **50**, 104–107 (2013).
62. Haack, T.B. *et al.* Impaired riboflavin transport due to missense mutations in *SLC52A2* causes Brown-Vialetto–Van Laere syndrome. *J. Inherit. Metab. Dis.* **35**, 943–948 (2012).
63. Green, P. *et al.* Brown-Vialetto–Van Laere syndrome, a ponto-bulbar palsy with deafness, is caused by mutations in *c20orf54*. *Am. J. Hum. Genet.* **86**, 485–489 (2010).
64. Johnson, J.O., Gibbs, J.R., Van Maldergem, L., Houlden, H. & Singleton, A.B. Exome sequencing in Brown-Vialetto–van Laere syndrome. *Am. J. Hum. Genet.* **87**, 567–569, author reply 569–570 (2010).
65. Bosch, A.M. *et al.* Brown-Vialetto–Van Laere and Fazio Londe syndrome is associated with a riboflavin transporter defect mimicking mild MADD: a new inborn error of metabolism with potential treatment. *J. Inherit. Metab. Dis.* **34**, 159–164 (2011).
66. da Silva-Júnior, F.P., Moura Rde, D., Rosemberg, S., Marchiori, P.E. & Castro, L.H. Cor pulmonale in a patient with Brown-Vialetto–Van Laere syndrome: a case report. *J. Neurol. Sci.* **300**, 155–156 (2011).
67. Dakhil, F.O., Bensreiti, S.M. & Zew, M.H. Pontobulbar palsy and sensorineural deafness (Brown-Vialetto–van Laere syndrome): the first case from Libya. *Amyotroph. Lateral Scler.* **11**, 397–398 (2010).
68. Lombaert, A., Dom, R., Carton, H. & Bruchler, J.M. Progressive ponto-bulbar palsy with deafness. A clinico-pathological study. *Acta Neurol. Belg.* **76**, 309–314 (1976).
69. van Bogaert, L. & van der Broeck, J. Sclérose latérale amyotrophique ou myasthénie bulbospinal avec exaltation des réflexes tendineux et cntractions fibrillaires. *J. Neurol. Psychiatry* **6**, 380–382 (1929).
70. Rotowski, J. & McHarg, J.F. A case of amyotrophic lateral sclerosis complicated by progressive lipodystrophy. *Edin. Med. J.* **60**, 281–293 (1953).
71. Gudbjartsson, D.F. *et al.* Sequence variants from whole genome sequencing a large group of Icelanders. *Sci. Data* **2**, 150011 doi:10.1038/sdata201511 (2015).

ONLINE METHODS

The Icelandic study population. This study is based on whole-genome sequence data from the white blood cells of 2,636 Icelanders participating in various disease projects at deCODE Genetics (**Supplementary Tables 1** and **2**). In addition, a total of 104,220 Icelanders have been genotyped using Illumina SNP chips (**Supplementary Table 6**).

All participating individuals, or their guardians, gave their informed consent before blood samples were drawn. The family history of participants donating blood was incorporated into the study by including the phenotypes of first- and second-degree relatives and integrating over their possible genotypes. This integration is performed without the genotypes being stored.

All sample identifiers were encrypted in accordance with the regulations of the Icelandic Data Protection Authority. Approval for these studies was provided by the National Bioethics Committee and the Icelandic Data Protection Authority.

Illumina SNP chip genotyping. The chip-typed samples were assayed with Illumina chips (**Supplementary Table 6**).

Whole-genome sequencing. Template DNA fragments were hybridized to the surface of flow cells (Genome Analyzer Paired-End Cluster Kit (v2) or HiSeq Paired-End cluster Kits (v2.5 or v3)) and amplified to form clusters using the Illumina cBot. Paired-end libraries were sequenced for 2×101 (HiSeq) or 2×120 (GAIIx) cycles of incorporation and imaging using the appropriate TruSeq SBS kits. Each library or sample was initially run on a single GAIIx lane for quality control validation, and further sequencing was performed on either the GAIIx (≥4 lanes) or HiSeq (≥1 lane) platform.

Whole-genome alignment. Reads were aligned to NCBI Build 36 (hg18) of the human reference sequence using Burrows-Wheeler Aligner (BWA) 0.5.7-0.5.9 (ref. 72). Alignments were merged into a single BAM file and marked for duplicates using Picard 1.55. Only non-duplicate reads were used for the downstream analyses. Resulting BAM files were realigned and recalibrated using GATK version 1.2-29-g0acaf2d (refs. 17,73).

Whole-genome SNP and indel calling. Multi-sample variant calling was performed with GATK version 2.3.9 using all 2,636 BAM files together.

Genotype calls made solely on the basis of next-generation sequence data yield errors at a rate that decreases as a function of sequencing depth. Thus, for example, if sequence reads at a heterozygous SNP position carry one copy of the alternative allele and seven copies of the reference allele, then without further information the genotype would be called homozygous for the reference allele. To minimize the number of such errors, we used information about haplotype sharing, taking advantage of the fact that all the sequenced individuals had also been chip typed and undergone long-range phasing (**Supplementary Fig. 3**)[18].

Whole-genome variant quality filtering. The variants identified by GATK were filtered using thresholds on GATK variant call annotations based on GATK best practices and other quality criteria (**Supplementary Fig. 2**).

Simple-repeat regions were defined by combining the entire Simple Tandem Repeats by TRF track in UCSC hg18 with all homopolymer regions in hg18 of length 6 bp or more[74]. Variants called in these regions were ignored in the analysis.

Coordinates of variants and regions were converted between hg18 and hg19 using the liftOver tool from UCSC[75].

Gene and variant annotation. Variants were annotated with information from Ensembl release 72 using VEP version 2.8 (refs. 20,76). Only protein-coding transcripts from RefSeq Release 56 (ref. 19) were considered. For transcripts occurring both in RefSeq and Ensembl for which the coding parts of the alignments were not identical, the RefSeq functional annotations were replaced by the corresponding Ensembl annotations, on the basis of claims by Ensembl that their alignments are more accurate and our verification of this. Variants were annotated with the classification categories of impact loss of function, moderate impact, low impact and other on the basis of SO[21] annotation from VEP (see **Supplementary Table 3** for the definitions by which the impact categories were classified). Sequence variants that could be assigned to more than one category (primarily because of their impact on more than one gene transcript) were assigned to the most severe of the applicable categories.

Determination of ancestral state. The inference of ancestral states for SNPs and indels was based on the Ensembl Compara ancestral sequences for *Homo sapiens* (GRCh37) corresponding to release 65 of Ensembl[28]. These sequences are created using the Enredo-Pecan-Ortheus (EPO) pipeline[77,78] for multiple-sequence alignment and inference of ancestor alignments using sequences from six primates. For polarization of indels, ancestral sequences were retrieved from the MAF files using bx-python.

Estimation of the transition/transversion ratio. Transition/transversion (Ts/Tv) ratios were calculated for various sets of identified SNPs. SNPs with alleles A and G or C and T were caused by transition mutations, and all other SNPs were caused by transversions. The Ts/Tv ratio within a set of SNPs was estimated as the number of SNPs caused by transitions divided by the number of SNPs caused by transversions.

Indel summary. We used three statistics to quantify the properties of various sets of insertions and deletions: the geometric mean length, exp(mean(log(length))), the relative number of deletions to insertions and the ratio of the number of insertions or deletions whose length was not a multiple of three to the number whose length was a multiple of three.

FRV and variant density. The density of variants in a subset of the genome is defined as the number of variants in the subset divided by the length of the subset. We report density as the number of variants per kb. The FRV is defined as the number of variants with DAF under 0.5% divided by the total number of variants in the set.

Exon position. We divided genes into two categories: multi-exon genes, those with two or more exons, and intronless genes, those with only one exon according to the RefSeq set. The exons of multi-exon genes were further divided into three groups: (i) first exons (those that were the first exon in at least one multi-exon transcript), (ii) last exons (those that were the last exon in at least one multi-exon transcript) and (iii) middle exons (those that were never the first, the last or the only exon in a transcript). We also split the intronless genes into olfactory receptors, according to Gene Ontology[36] classification, and other intronless genes.

OMIM. We defined the overlap between the RefSeq genes and OMIM[25] previously linked to disease as OMIM disease-related genes[27]. All other RefSeq genes were classified as not in OMIM. On the basis of keyword, OMIM genes had been further classified as 'haploinsufficient', 'dominant negative', '*de novo* disease causing' and 'recessive' (ref. 27). We defined as recessive the genes that were only classified as recessive in OMIM. We defined as dominant the genes that were classified in OMIM as either haploinsufficient, dominant negative or *de novo*. The remaining OMIM disease-related genes were designated as 'other'.

Gene ontology. We used the subset of Gene Ontology[36] terms defined by the PANTHER GO slim version 8.1 (ref. 37) on the set of RefSeq genes. Genes were counted as members of all GO classes that could be reached from their designated class by "is a" and "part of" relationships.

Chromatin states. A hidden Markov model has been used to cluster chromatin immunoprecipitation and sequencing (ChIP-seq) data for nine histone antibodies from nine cell types[79]. Using the trained model, regions of the genome could be assigned to 1 of the 15 learned chromatin states at a resolution of 200 bp (UCSC browser ChromHMM track). Leaving out the two states corresponding to repetitive elements, we tested the remaining 13 states for enrichment of rare variants in all 9 cell types. We report the density and FRV for the 13 tested states averaged over the 9 available cell types.

Sensitive regions. We followed the published definition of subsets of the genome as sensitive and ultra-sensitive[26]. These designations were based on using FRVs to estimate the strength of purifying selection on different functional categories. The FRV was calculated for a selection of DNA elements

from the ENCODE Project using variants from the 1,092 individuals in the 1000 Genomes Project (Phase 1)[16].

Overlap with ESP and dbSNP. We assessed the overlap between the variants we discovered in Iceland with those in dbSNP[41] and with those reported by ESP[13,14]. The Icelandic variants were counted as existing in ESP or dbSNP if a variant at the same position and with the same allele was present in the database.

Frequency distribution comparison. The frequency distribution of Icelandic coding-sequence SNPs was compared to that of coding-sequence SNPs in the European-American portion of ESP. To make the frequency distributions comparable, we restricted ourselves to SNPs (indel calling is much less consistent) in regions with good coverage in both sets (15× for Iceland and 40× for ESP). Also, only SNPs with a MAF over 0.1% were used because the number of rarer variants is highly dependent on sample size and the sequencing and calling methods used.

Homozygote count. Given the allelic frequency of an autosomal sequence variant, Hardy-Weinberg equilibrium specifies the homozygote frequency. We binned variants according to their MAF and, within each bin, calculated the expected number of homozygotes under Hardy-Weinberg equilibrium, where a haplotype was considered to carry the minor allele if the imputed value for the haplotypes was greater than 0.9, and the number of homozygotes, where an imputed individual was considered to be homozygous if both of this individual's haplotypes had imputed values greater than 0.9.

See the **Supplementary Note** for further details.

72. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
73. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
74. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
75. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
76. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
77. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
78. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
79. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).