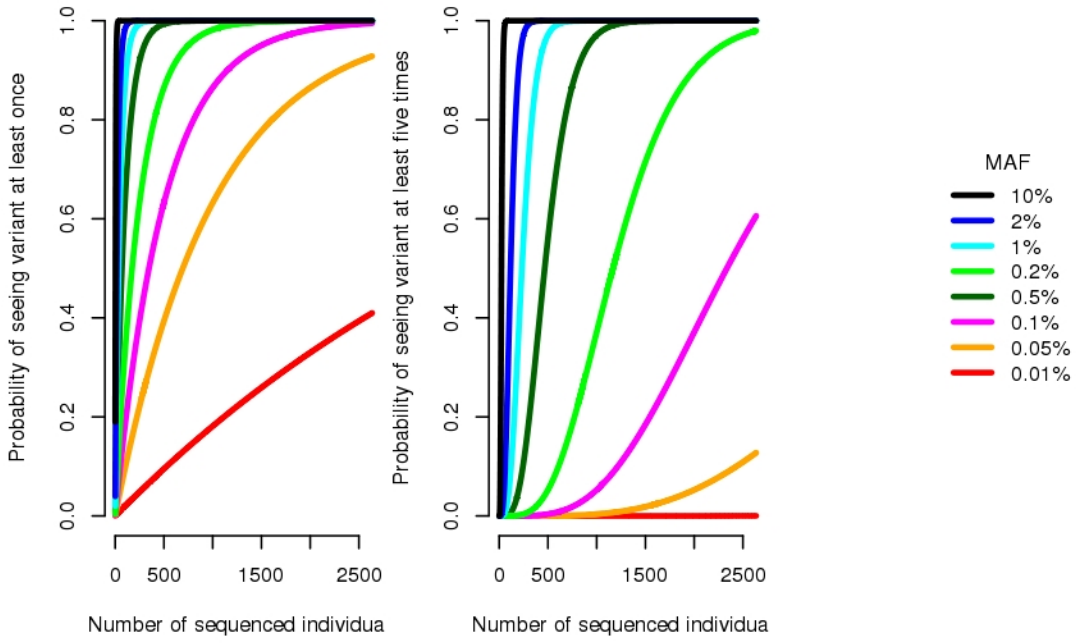
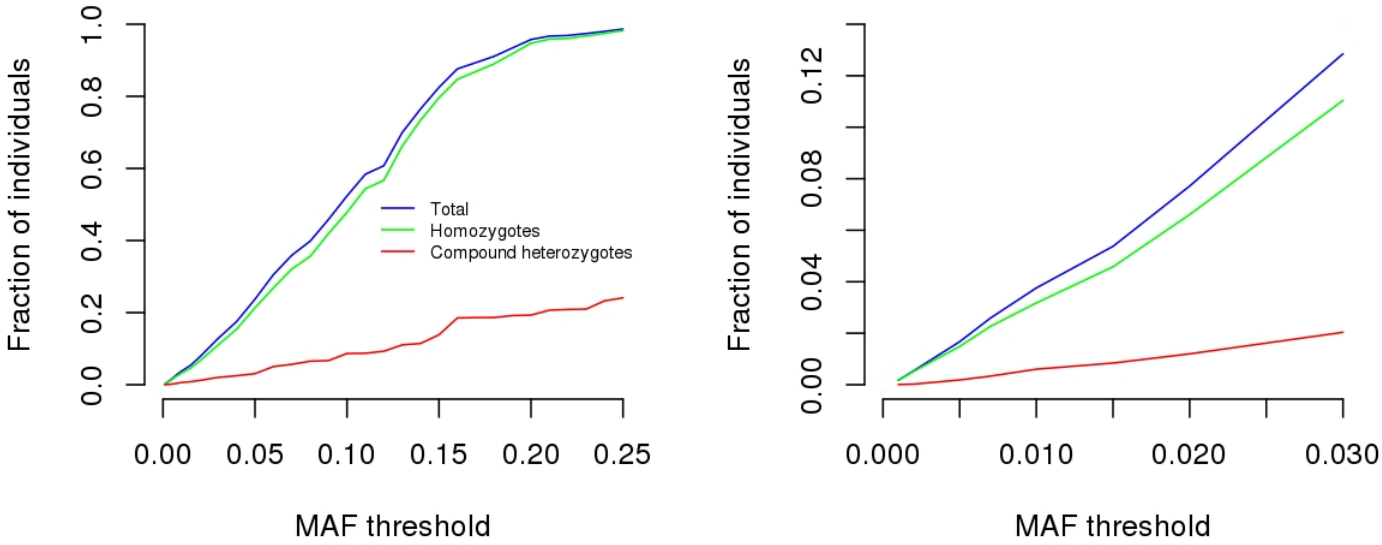


Supplementary Figure 1

The probabilities of a variant being seen once and five times as a function of the number of individuals sequenced by minor allele frequency (MAF).

Variants that are seen at least five times, corresponding to an observed allelic frequency of 0.095%, are likely to be imputed with good quality.

Nature Genetics: doi:10.1038/ng.3243

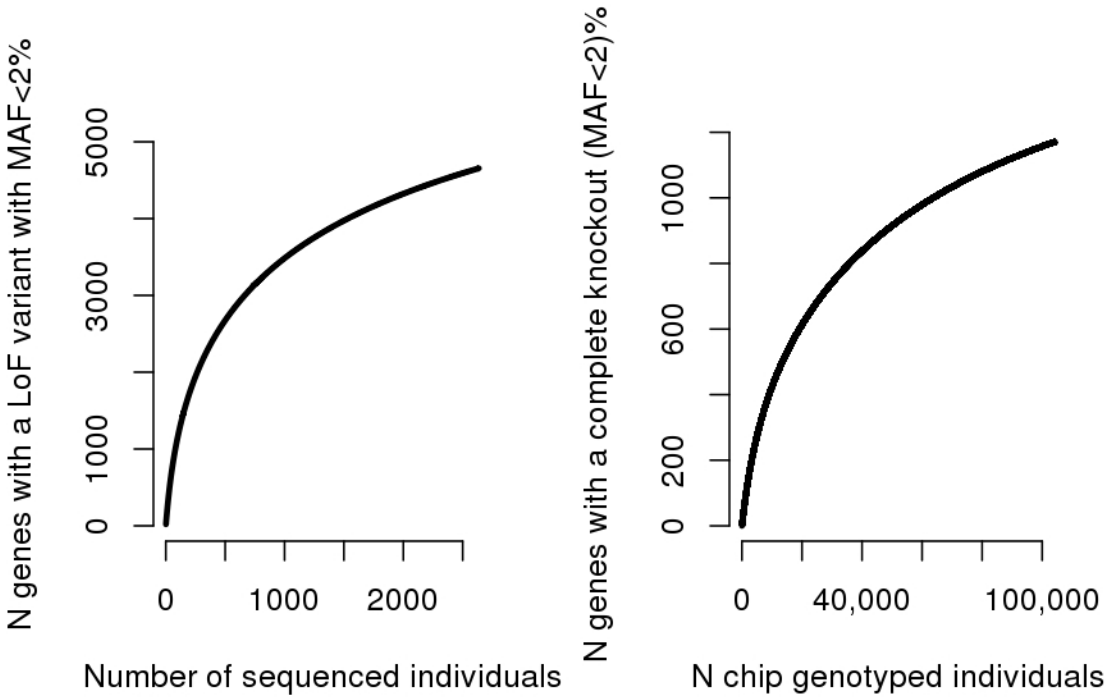


Supplementary Figure 2

The fraction of individuals among 104,220 individuals with imputed genotypes that have genes completely knockout out by LoF variants with MAF below the given threshold.

The second panel shows a magnified view of MAF below 3%.

Nature Genetics: doi:10.1038/ng.3243

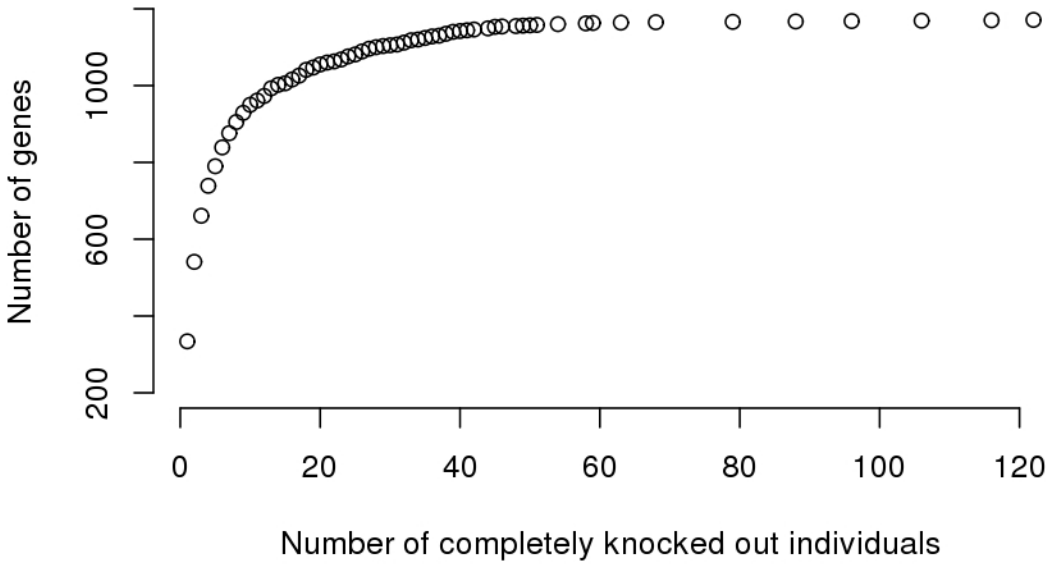


Supplementary Figure 3

The number of genes that are observed to have at least one LoF variant with MAF below 2% as a function of the number of sequenced individuals and the number of genes that are completely knocked out in at least one individual by LoF variants with MAF below 2% as a function of the number of chip-genotyped individuals.

The curves are derived from the allele frequency distributions and the number of imputed complete knockouts.

Nature Genetics: doi:10.1038/ng.3243

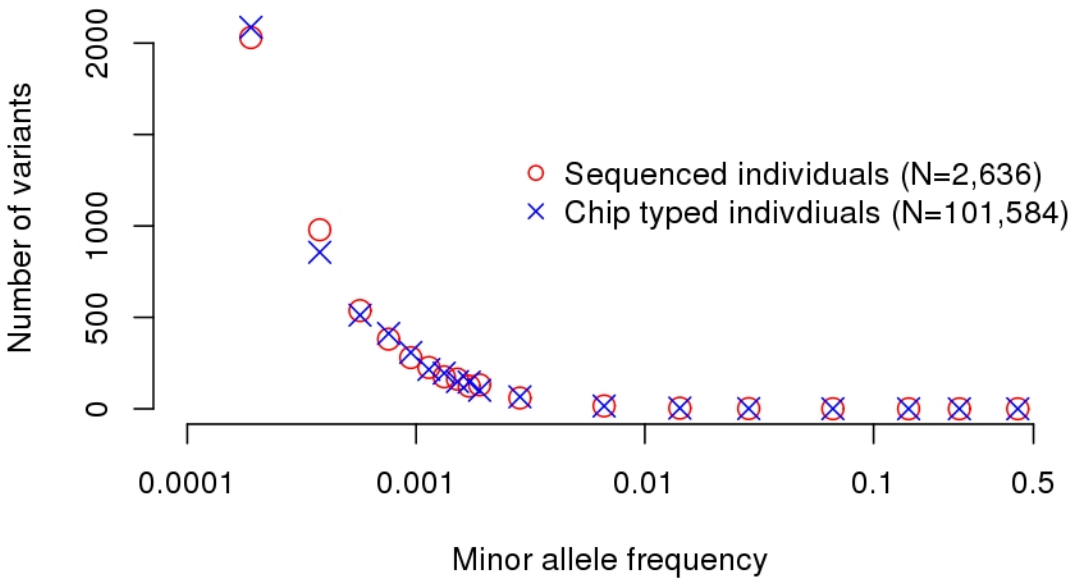


Supplementary Figure 4

The cumulative number of genes by the number of completely knocked out individuals.

The total number of genes is 1,171.

Nature Genetics: doi:10.1038/ng.3243

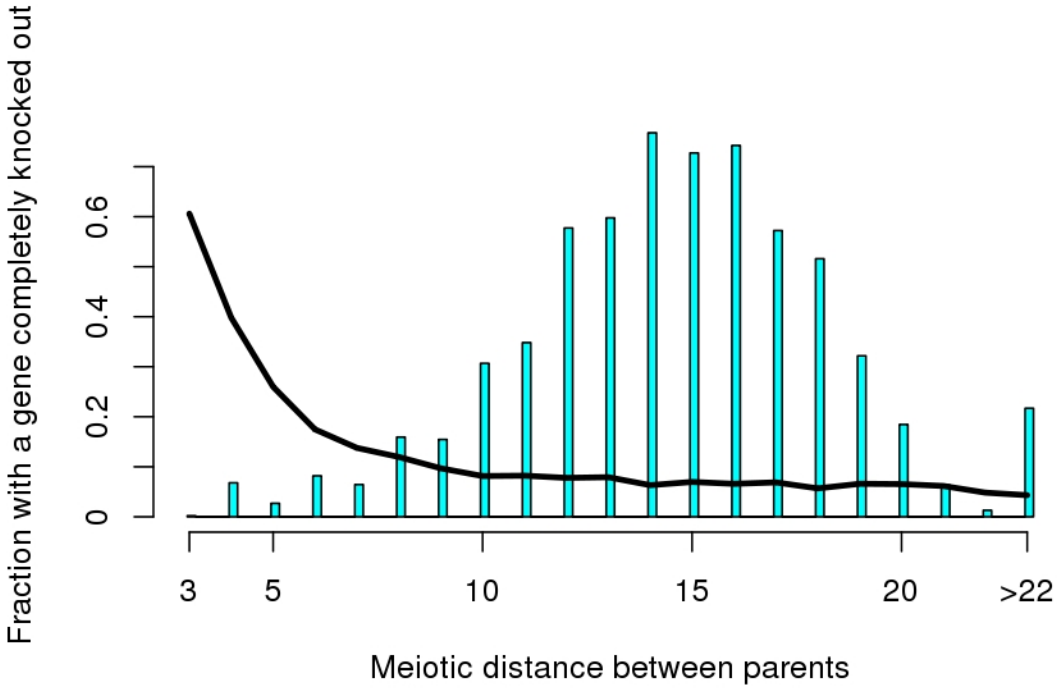


Supplementary Figure 5

The frequency distribution of the 6,795 LoF variants among sequenced and imputed individuals.

The leftmost count includes all variants with a frequency less than one and half divided by twice the number of sequenced individuals, which corresponds to the number of variants seen only once in the sequenced set.

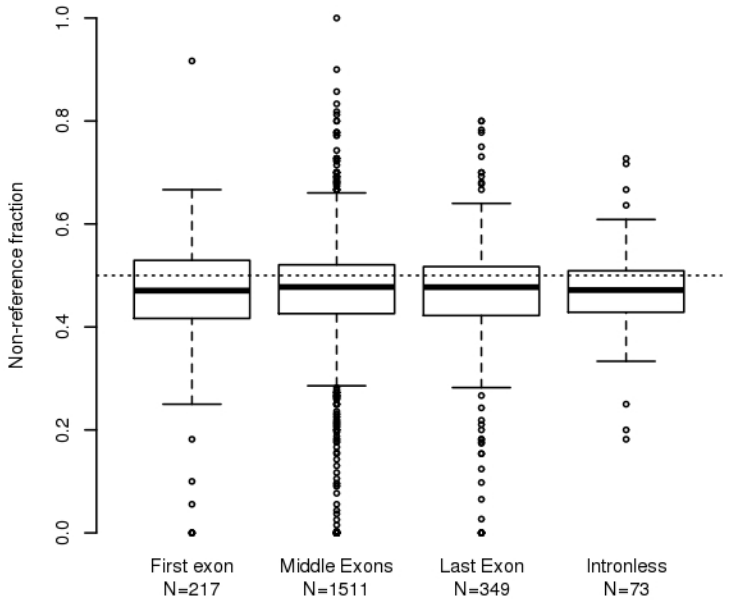
Nature Genetics: doi:10.1038/ng.3243



Supplementary Figure 6

A histogram of the number of meiosis between the parents of the 104,220 Icelanders in our study and the fraction of individuals that have at least one gene completely knocked out by rare LoF variants.

Nature Genetics: doi:10.1038/ng.3243



Supplementary Figure 7

Transcriptome effect of synonymous SNPs by exon rank.

The allele-specific expression of the non-reference allele was calculated for each variant for a set of 262 individuals with blood RNA sequence data. The top, middle and bottom of the boxes are the top quartile, median and bottom quartile values calculated over the set of variants. The whiskers show the lowest and highest datum within 1.5 times the interquartile range (IQR) from the median. The dots indicate datum more than 1.5 times the IQR from the median. The *n* values given are the number of variants in each class.

Nature Genetics: doi:10.1038/ng.3243

SUPPLEMENTARY NOTE

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[1] (Supplementary Tables 2 and 3). In addition, a total of 104,220 Icelanders have been genotyped using Illumina SNP chips[1].

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.

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.

The Icelandic Genealogy. The Icelandic genealogical database contains 819,410 individuals back to 740 AD. Of the 471,284 Icelanders recorded to have been born in the 20th century, 91.1% had a recorded father and 93.7% had a recorded mother in the database. Similarly, of the 183,896 Icelanders recorded to have been born in the 19th century, 97.5% had a recorded father and 97.8% had a recorded mother.

Illumina SNP Chip Genotyping. The chip-typed samples were assayed with the Illumina HumanHap300, HumanCNV370, HumanHap610, HumanHap1M, HumanHap660, Omni-1, Omni 2.5 or Omni Express bead chips at deCODE genetics[1]. Chip SNPs were excluded if they had (i) yield less than 95%, (ii) minor allele frequency (MAF) less than 1% in the population or (iii) significant deviation from Hardy-Weinberg equilibrium ($P < 0.001$), (iv) if they produced an excessive inheritance error rate (over 0.001), (v) if there was substantial difference in allele frequency between chip types (from just a single chip if that resolved all differences, but from all chips otherwise). All samples with a call rate below 97% were excluded from the analysis. The final set used for long-range phasing comprised 676,913 autosomal SNPs.

Long range phasing. Long range phasing of all chip-genotyped individuals was performed with methods described previously[2,3]. In brief, phasing is achieved using an iterative algorithm which phases a single proband at a time given the available phasing information about everyone else that shares a long haplotype identically by state with the proband. Given the large fraction of the Icelandic population that has been chip-typed, accurate long range

Nature Genetics: doi:10.1038/ng.3243

phasing is available genome-wide for all chip-typed Icelanders. For long range phased haplotype association analysis, we then partitioned the genome into non-overlapping fixed 0.3 cM bins. Within each bin, we observed the haplotype diversity described by the combination of all chip-typed markers in the bin.

Whole-genome sequencing sample preparation. Paired-end libraries for sequencing were prepared according to the manufacturer's instructions (Illumina, TruSeq™). In short, approximately 1 μg of genomic DNA, isolated from frozen blood samples, was fragmented to a mean target size of 300 bp using a Covaris E210 instrument. The resulting fragmented DNA was end repaired using T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP followed by addition of an 'A' base at the ends using Klenow exo fragment (3′ to 5′-exo minus) and dATP (1 mM). Sequencing adaptors containing 'T' overhangs were ligated to the DNA products followed by agarose (2%) gel electrophoresis. Fragments of about 400-500 bp were isolated from the gels (QIAGEN Gel Extraction Kit), and the adaptor-modified DNA fragments were PCR enriched for ten cycles using Phusion DNA polymerase (Finnzymes Oy) and a PCR primer cocktail (Illumina). Enriched libraries were further purified using AMPure XP beads (Beckman-Coulter). The quality and concentration of the libraries were assessed with the Agilent 2100 Bioanalyzer using the DNA 1000 LabChip (Agilent). Barcoded libraries were stored at −20 °C. All steps in the workflow were monitored using an in-house laboratory information management system with barcode tracking of all samples and reagents.

Whole-genome sequencing. Template DNA fragments were hybridized to the surface of flow cells (GA PE cluster kit (v2) or HiSeq PE cluster kits (v2.5 or v3)) and amplified to form clusters using the Illumina cBot. In brief, DNA (2.5–12 pM) was denatured, followed by hybridization to grafted adaptors on the flow cell. Isothermal bridge amplification using Phusion polymerase was then followed by linearization of the bridged DNA, denaturation, blocking of 3´ ends and hybridization of the sequencing primer. Sequencing-by-synthesis (SBS) was performed on Illumina $GAII_x$ and/or HiSeq 2000 instruments. Paired-end libraries were sequenced at 2 × 101 (HiSeq) or 2 × 120 ($GAII_x$) cycles of incorporation and imaging using the appropriate TruSeq™ SBS kits. Each library or sample was initially run on a single $GAII_x$ lane for QC validation followed by further sequencing on either $GAII_x$ (≥4 lanes) or HiSeq (≥1 lane) with targeted raw cluster densities of 500–800 k/mm^2, depending on the version of the data imaging and analysis packages (SCS2.6-2-9/RTA1.6-1.9, HCS1.3.8-

Nature Genetics: doi:10.1038/ng.3243

1.4.8/RTA1.10.36-1.12.4.2). Real-time analysis involved conversion of image data to base-calling in real-time.

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[4]. Alignments were merged into a single BAM file and marked for duplicates using Picard 1.55 (http://picard.sourceforge.net/). Only non-duplicate reads were used for the downstream analyses. Resulting BAM files were realigned and recalibrated using GATK version 1.2-29-g0acaf2d[5,6].

Whole-genome SNP and INDEL calling. Multi-sample variant calling was performed with GATK version 2.3.9 using all the 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 long range phased[1,7]. Extending the previous example, if the individual shares a haplotype with another who is heterozygous given his sequence reads, then the ambiguous individual would be called as heterozygous. Conversely, if the individual shares both his haplotypes with others who are homozygous for the major allele his genotype would be called homozygous. In order to improve genotype quality and to phase the sequencing genotypes, an iterative algorithm based on the IMPUTE HMM model[8] which uses the LRP haplotypes was employed. Assume a SNP with alleles 0 and 1 is being phased. We let H be the long range phased haplotypes of the sequenced individuals and applied the following hidden Markov model (HMM) based algorithm.

Assuming that at each marker i the haplotype h has a common ancestor with a haplotype in $H\backslash\{h\}$ and denote the variable indicating this with the latent variable $z_i \in H\backslash\{h\}$, the hidden variable in the HMM. Then

$$\gamma_{h,k,i} = P(z_i = k | all\ LRP\ markers),$$

for all $k \in H\backslash\{h\}$. Given a haplotype h in H, $\gamma_{h,k}$ are calculated simultaneously for all $k \in H\backslash\{h\}$ using the same HMM model as IMPUTE[8]. Given the Markov assumptions of the HMM, the model is fully specified by emission and transition probabilities.

Nature Genetics: doi:10.1038/ng.3243

We define the emission probabilities of the HMM at each marker *i* as:

$$P(z_i = k|marker\ i) = \begin{cases} 1-\lambda, \text{if } h \text{ and } k \text{ match at } i \\ \lambda, \text{if } h \text{ and } k \text{ mismatch at } i \end{cases}$$

where λ can be thought of as a penalty for a mismatch. We used $\lambda = 10^{-7}$ in our implementation. We define the transmission probabilities of the HMM model as:

$$P(z_i|z_{i-1}, markers\ 1, \dots, i-1) = \begin{cases} e^{-\frac{\rho_i}{N}} + \frac{1-e^{-\frac{\rho_i}{N}}}{N}, \text{if } z_i = z_{i-1} \\ \frac{1-e^{-\frac{\rho_i}{N}}}{N}, \text{if } z_i \neq z_{i-1} \end{cases}$$

Where N is the number of haplotypes in $k \in H\backslash\{h\}$, which for autosomal chromosomes is $2(2{,}636-1)$ here and $\rho_i = 4N_e r_i$, where r_i is the genetic distance between markers $i-1$ and i according to the most recent version of the deCODE genetic map[9] and N_e was originally meant to be an estimate of the effective number of haplotypes in the population that our sample comes from, we used $N_e = 7{,}000$. These definitions fully specify the probability distribution $P(z_i|all\ markers)$. Calculating $\gamma_{h,k}$ for a single haplotype requires $O(MN)$ operations, where *N* is the number of haplotypes and *M* is the number of markers. Since these calculations can be performed for one haplotype at a time, the calculations can be parallelized across a computer cluster for efficiency. In practice most of the $\gamma_{h,k}$ will be close to zero and can be safely ignored (we used a threshold of 10^{-6} of the largest value at each marker for each *h*) greatly reducing storage requirements.

Now we are set to describe an iterative algorithm for the actual phasing. For every *h* in *H*, initialize the parameter θ_h, which specifies how likely the one allele of the SNP is to occur on the background of *h* from the genotype likelihoods obtained from sequencing. The genotype likelihood L_g is the probability of the observed sequencing data at the SNP for a given individual assuming *g* is the true genotype at the SNP. If L_0, L_1 and L_2 are the likelihoods of the genotypes 0, 1 and 2 in the individual that carries *h*, then set θ_h:

$$\theta_h = \frac{L_2 + \frac{1}{2}L_1}{L_2 + L_1 + L_0}.$$

For every pair of haplotypes *h* and *k* in *H* that are carried by the same individual, use the other haplotypes in *H* to predict the genotype of the SNP on the backgrounds of *h* and *k*:

$$\tau_h = \sum_{l \in H\backslash\{h\}} \gamma_{h,l}\theta_l \text{ and}$$

Nature Genetics: doi:10.1038/ng.3243

$$\tau_k = \sum_{l \in H \setminus \{k\}} \gamma_{k,l} \theta_l.$$

Combining these predictions with the genotype likelihoods from sequencing gives un-normalized updated phased genotype probabilities:

$$P_{00} = (1 - \tau_h)(1 - \tau_k) L_0,$$

$$P_{10} = \tau_h (1 - \tau_k) \tfrac{1}{2} L_1,$$

$$P_{01} = (1 - \tau_h) \tau_k \tfrac{1}{2} L_1,$$

$$\text{and } P_{11} = \tau_h \tau_k L_2.$$

Now use these values to update θ_h and θ_k to:

$$\theta_h = \frac{P_{10} + P_{11}}{P_{00} + P_{01} + P_{10} + P_{11}} \text{ and}$$

$$\theta_k = \frac{P_{01} + P_{11}}{P_{00} + P_{01} + P_{10} + P_{11}}.$$

Iterate until the maximum difference between iterations is less than a convergence threshold ε. We used $\varepsilon = 10^{-7}$.

Genotype imputation. Given the long range phased haplotypes of the sequenced individuals, H, and θ, the carrier probabilities of the haplotypes in the sequenced set, the probability that a new haplotype n, not in the set of sequenced haplotypes H, is imputed as $\sum_{l \in H} \gamma_{n,l} \theta_l$, where $\gamma_{n,l}$ is calculated as above for every $l \in H$.

In order to test the quality of our imputations, we compared chip genotypes to those imputed based on the sequence reads. Chip-based genotypes were available from 28,204 of the SNPs detected by sequencing that affect exons and splice regions, which proved to be fully concordant with genotypes from whole-genome sequencing. To assess imputation quality, we compared the imputed genotypes with those from the same set of 28,204 chip SNPs[1]. The concordance between imputed and chip-typed genotypes was high. For example, 76.4% of SNPs with a DAF of 1% were imputed almost perfectly ($r^2 > 0.99$) and 98.4% of SNPs were imputed accurately ($r^2 > 0.9$). Even for the SNPs that have only two copies of the minor allele observed in the set of sequenced individuals (DAF=0.04%) 47% were imputed almost perfectly ($r^2 > 0.99$) and 64% of SNPs were imputed accurately ($r^2 > 0.9$). We estimate the probability of a variant with MAF > 0.1% being observed at least twice among the 2,636 sequenced individuals to be 96.7%.

Nature Genetics: doi:10.1038/ng.3243

Genotype imputation information. The informativeness of genotype imputation is estimated by the ratio of the variance of imputed expected allele counts and the variance of the actual allele counts:

$$\frac{Var(E(\theta|chip\ data))}{Var(\theta)},$$

where θ is the allele count. $Var(E(\theta|chip\ data))$ is estimated by the observed variance of the imputed expected counts.

In the sequenced set the information was calculated on a genotype level so that $\theta \in \{0, 1, 2\}$ and $Var(\theta) = 2p(1-p)$, assuming Hardy-Weinberg equilibrium, where p is the allele frequency. In the case of correct and accurate genotype calls the estimated information will be close to 1. Ambiguous genotype calls will usually lead to information estimates below 1. However, in some cases, due to artifacts, all observed allele calls will be 0 or 2 and the estimated information can go up to 2.

The imputations into chip typed individuals is done on a haplotype basis so that $\theta \in \{0, 1\}$ and the estimated information is bounded between 0 and 1, where 1 represents exact imputations and ambiguous imputations yield values below 1.

Whole-genome variant quality filtering. The variants identified by GATK were filtered using thresholds on GATK variant call annotations [10]. SNPs were discarded if at least one of the following thresholds for their GATK call annotation parameter was violated: QD (Variant confidence/quality by depth) < 2.0, MQ (RMS mapping quality) < 40.0, FS (Fisher strand) > 60.0, HaploTypeScore > 13.0, MQRankSum < -12.5, and ReadPosRankSum< -8.0. Indels were discarded if at least one of the inequalities QD < 2.0, FS >200.0, or ReadPosRankSum < -20.0 were satisfied. The thresholds for these parameters were adopted from GATK Best Practices (see URLs). In addition, SNPs and indels were discarded if one of the following thresholds were violated: DP (sequencing coverage/depth) > 110,000, AN (total number of alleles in called genotypes) < 4,200, and HW (Hardy-Weinberg P among sequenced samples) < 10^{-7}, SI (genotype information among sequenced individuals) > 1.4, or if SI < 0.6 for SNPs and SI < 0.9 for indels. The GATK call annotation AN corresponds to the total number of chromosomes in called genotypes which equals 2 × 2,636 =5,272 when all chromosomes can be called. The additional filtering removes 2% of the remaining SNPs but 50% of the remaining indels. It is primarily the SI condition that removes indels, which usually indicates

Nature Genetics: doi:10.1038/ng.3243

that the failing indel is not called with a high degree of certainty and that its calling cannot be resolved in a coherent manner based on haplotype sharing.

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[11]. Variants called in these regions were ignored in the analysis.

Liftover between hg18 and hg19. Coordinates of variants and regions were converted between hg18 and hg19 using the liftOver tool from UCSC[12].

Gene and variant annotation. Variants were annotated with information from Ensembl release 70 using Variant Effect Predictor (VEP) version 2.8[13,14]. Only protein coding transcripts from RefSeq Release 56[15] were considered. Variants were annotated with the classification categories of impact LoF, MODERATE, LOW, and OTHER based on the Sequence Ontology (SO)[16] annotation from VEP (see Supplementary Table 1 for the definition of the classification of impact categories). Sequence variants that could be assigned to more than one category (primarily due to their impact on more than one gene transcript) were assigned to the most severe of the applicable categories.

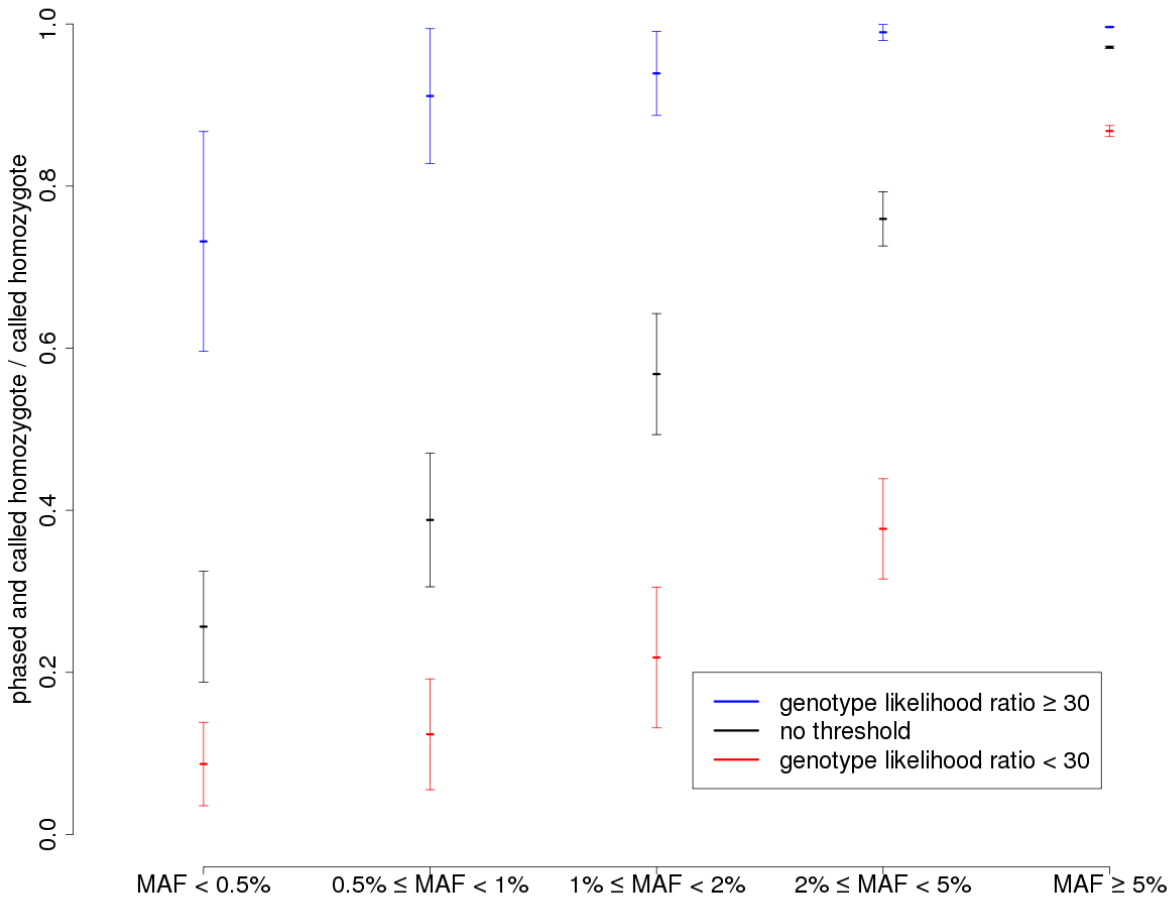
Estimation of carrier status based on imputation data. An imputed individual was counted as a carrier of the minor allele of a variant if the imputation probability was greater than 0.9 for at least one of his two haplotypes and as homozygous carrier if the imputation probability was greater than 0.9 for both of his haplotypes.

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 based on the RefSeq set. The exons of multi-exon genes were further divided into three groups: (i) first exons (those that are the first exon in at least one multi-exon transcript), (ii) last exons (those that are the last exon in at least one multi-exon transcript), and (iii) middle exons (those that are never the first, the last or the only exon in a transcript).

Overlap with ESP and dbSNP. We assessed the overlap between the variants we discovered in Iceland with those in dbSNP[17] and with those reported by the ESP[14,15]. The Icelandic variants were counted as existing in the ESP or dbSNP if a variant at the same position and with the same allele was present in the database.

Nature Genetics: doi:10.1038/ng.3243

Calling rare homozygotes. With the depth of sequencing that we have employed, there is a chance that only one allele of a sequence variant will be represented in the sequencing reads from a truly heterozygous individual[18]. The probability that the homozygous minor allele genotype is the most likely genotype given only the sequencing reads of the individual when he is truly heterozygous goes up with lower sequencing depth at the variant and lower MAF. This has the practical consequence that for low MAFs, a large proportion of individuals who are most likely minor allele homozygotes given the sequencing data are truly heterozygotes. We do however call genotypes based on combining genotype likelihoods and haplotype sharing with other sequenced individuals which addresses this issue.



Relatedness of parents in Iceland. The Icelandic genealogical database was used to estimate the shortest distance between the parents of the 104,220 genotyped Icelanders. The distance between the parents was measured in the number of meiosis between them.

Nature Genetics: doi:10.1038/ng.3243

Transmission. Sharing between parent offspring pairs was tested and pairs with less than 99% haplotype sharing were excluded from the analysis. Haplotype sharing with other close family members was also tested and inconsistencies removed. We restricted our analysis to chip typed and long range phased individuals and excluded sequenced individuals to avoid observation bias. For example, if only the offspring of a triad has been sequenced, a variant that is transmitted is more likely to be imputed into the parent than a variant that was not transmitted. After exclusion, 35,024 father offspring, 47,769 mother offspring pairs and 26,188 triads remained. All the individuals are long-range phased which allows identification of recombinations. We excluded regions within 1Mb of recombinations in the offspring since these regions are most likely to suffer from phasing uncertainty and errors. For markers within these regions where both parents were imputed to be heterozygous (each allele being imputed to be the major or allele with 99% probability) and counted how many times the minor allele was transmitted to the offspring. This method does not rely on imputation into the offspring and should be robust to errors in phasing and genotyping, as such errors will shuffle the transmitted allele and are not likely to cause a bias.

Confidence intervals and significance were estimated using studentized bootstrap sampling[19] of parent offspring pairs and trios for the heterozygous and double heterozygous transmissions, respectively. The standard error of the estimated transmission probability was estimated by creating 10,000 boostrap samples and the distribution of the transmission probability estimate was assumed to be approximately normally distributed around the true transmission probability. The 95% confidence interval around the observed transmission probability t with estimated standard error σ_B is then $(t - 1.96\sigma_B, t + 1.96\sigma_B)$.

Validation of rare genotypes. In order to validate some of the imputed homozygous rare LOF genotypes, we attempted to Sanger sequence over 211 SNP or indel variants in up to 302 individuals. Dye-terminator Sanger sequencing was performed with the Applied Biosystems BigDye Terminator v3.1 Cycle Sequencing Kit, with Agencourt Ampure XP and Agencourt CleanSeq for cleanup of the PCR and cycle sequencing product, respectively. AmpureXP and CleanSeq bead cleaning was performed on a Zymark Sciclone ALH-500 liquid-handling robot system. Tray dilutions for PCR setup and cycle-sequencing setup were prepared on a Packard MultiprobeII HTEX liquid-handling robot system, and genomic DNA and PCR product were transferred into their respective trays using the Zymark Sciclone ALH-500. PCR and cycle sequencing reactions were performed on MJ Research PTC-225 thermal cyclers. For signal detection, Applied Biosystems 3730xl DNA analyzers were used.

Nature Genetics: doi:10.1038/ng.3243

Quantifying allele specific expression in white blood cells for LoF variants. We estimated allele specific expression in blood for 262 individuals that been RNA sequenced. All of these individuals had imputed and phased genotypes.

Preparation of Poly-A cDNA sequencing libraries. The quality and quantity of isolated total RNA samples was assessed using the Total RNA 6000 Nano chip for the Agilent 2100 Bioanalyzer. cDNA libraries derived from Poly-A mRNA were generated using Illumina´s TruSeq™ RNA Sample Prep Kit. Briefly, Poly-A mRNA was isolated from total RNA samples (1-4 µg input) using hybridizaton to Poly-T beads. The Poly-A mRNA was fragmented at 94°C, and first-strand cDNA was prepared using random hexamers and the SuperScript II reverse transcriptase (Invitrogen). Following second-strand cDNA synthesis, end repair, addition of a single A base, adaptor ligation, AMPure® bead purification, and PCR amplification, the resulting cDNA was measured on a Bioanalyzer using the DNA 1000 Lab Chip.

RNA Sequencing. Samples were clustered on to flowcells using Illumina‘s cBot and the TruSeq PE cluster kits v2, respectively. Paired-end sequencing (2x76 cycles) was performed on GAIIx instruments, equipped with paired-end modules using the TruSeq SBS kits v5 from Illumina. Approximately 125-175 million forward reads (250-350 M total reads) were sequenced per sample.

Protocol and software used for RNA read alignment. Read alignment and gene expression estimation were based on the recently published Tuxedo protocol described in[20]. The Tuxedo protocol describes how to use the open-source software tools TopHat[21] and Cufflinks[22] which have both been developed specifically for analysis of RNA-seq data. Sequencing reads were aligned to Homo sapiens Build 36 with TopHat version 1.4.1 with a supplied set of known transcripts in GTF format (RefSeq hg18; Homo sapiens, NCBI, build36.3). With this option, TopHat will first try to align reads to the provided transcriptome and only the reads that do not fully map to the transcriptome will then be mapped on the genome.

Quantifying allele specific expression based on phased heterozygous loci. To estimate the allele specific expression of a gene in which an individual was heterozygous for a LoF variant, reads for each allele were counted (using samtools mpileup[23]). Loci covered by skipped reads were ignored.

Complete knockouts by human expression pattern. Gene counts for tissue-specific expression in Table 3 and Supplementary Table 5 were based on publicly available data that

Nature Genetics: doi:10.1038/ng.3243

classifies tissue-specific espression of genes across twenty-seven tissues[24] (Supplementary Dataset 1). Counts in Table 3 are based on genes and tissues that have FPKM greater than or equal to 20 and are in the 'Expressed in all high' category in Fagerberg et al.[24]; the counts in Supplementary Table 6 are based the list of specific and enriched genes are autosomal genes in the categories 'Highly tissue enriched' and 'Moderately tissue enriched' from Fagerberg et al. [24]. The 'Highly tissue enriched' ('Moderately tissue enriched') category consists of genes-tissue pairs where the FPKM level for the gene expression in one of the tissues is 50-fold (5-fold) higher than for any of the other 26 tissues.

Complete knockouts by phenotypes of mouse orthologs. Genes were annotated with phenotypes from the Mouse Genome Database (MGD)[25] at the Mouse Genome Informatics (MGI) website. The MGD gene annotation was based on the tables in the files HMD_HumanPhenotype.rpt and VOC_MammalianPhenotype.rpt,by joining the tables on the Mammalian Phenotype (MP) accession number. Genes were annotated as recessive based on the Human Phenotype Ontology (HPO)[26] recessive gene list (based on Online Mendelian Inheritance in Man (OMIM)) and a recently published list for severe recessive diseases with pediatric presentations[27]; the genes on the latter list were annotated as 'recessive-severe-genes'. Twenty eight mouse phenotype classes were considered. The 'normal' phenotype was not included in our analysis.

Confidence intervals for expression patterns and mouse phenotype orthologs. Confidence intervals for the number of completely knocked out genes by class were estimated using studentized bootstrap sampling (B=10,000)[19] of the genes in the class and estimated as described above.

Nature Genetics: doi:10.1038/ng.3243

SO term	SO description	SO accession	Impact
transcript ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	LoF
splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	LoF
splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	LoF
stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	LoF
frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	LoF
stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	LoF
initiator codon variant	A codon variant that changes at least one base of the first codon of a transcript	SO:0001582	LoF
inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	MODERATE
inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	MODERATE
missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	MODERATE
transcript amplification	A feature amplification of a region containing a transcript	SO:0001889	MODERATE
splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	MODERATE
incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	MODERATE

Supplementary Table 1. Classification of the impact of sequence variant annotations from VEP [13]. The sequence variant annotation is based on the Sequence Ontology (SO) [16]. This classification is consistent with the order of severity of the sequence variants estimated by Ensembl.

Nature Genetics: doi:10.1038/ng.3243

Disease	N
Coronary Artery Disease	474
Chronic Kidney Disease	424
Obesity	394
Hypertension	371
Type 2 diabetes	307
Osteoporosis	260
Atrial Fibrillation	259
Myocardial Infarction	246
Alzheimer's disease	214
Asthma	211
Osteoarthritis	209
Urinary Tract Infection	208
Systemic Lupus Erythematous	207
Alcohol Dependence	199
Breast Cancer	196
Depression	171
Sleep Apnea	165
Autism Spectrum Disorders	160
Kidney Stones	160
Nicotine Dependence	158
Prostate Cancer	158
Gallstones	156
Schizophrenia	155
Attention Deficit Hyperactivity Disorder	154
Glaucoma	138
Mental Retardation	138
Colorectal Adenoma	134
Migraine	132
Age Related Macular Degeneration	127
Ischaemic Stroke	122
Psoriasis	120
Epilepsy	118
Tuberculosis	118
Basal Cell Carcinoma Of The Skin	115
Parkinson's Disease	108
Diverticular Disease	107
Cataract	106
Bronchitis	98
Tourette Syndrome	84
Heart Failure	81
Emphysema	80
Dyslexia	79
Hypothyroidism	79
Sick Sinus Syndrome	78
Peripheral Artery Disease	77
Rheumatoid Arthritis	73
Benign Prostatic Hyperplasia	70
Congenital Heart Disease	70
Hypertension In Pregnancy	69
Panic Disorder	69

Supplementary Table 2. The 50 most prevalent conditions among the 2,636 sequenced Icelanders.

Nature Genetics: doi:10.1038/ng.3243

Demographic	Sequenced	Chip-typed	Relatives of chip-typed
N	2,636	104,220	294,212
Female (%)	54.0	55.1	46.8
YOB[a] (SD)	1950 (23)	1953 (23)	1964 (36)
Alive[b] (%)	72.5	84.9	72.9
Age[c] (SD)	55 (20)	56 (70)	33 (20)
Lifespan[d] (SD)	79 (13)	80 (13)	59 (28)

Supplementary Table 3. The demographics of the sequenced, chip-typed and relatives of chip-typed individuals. [a]Year of birth. [b]Fraction currently alive. [c]Current age for the living. [d]Age at death.

Nature Genetics: doi:10.1038/ng.3243

							Sanger genotype		
Gene	**Chr**	**Pos (build 36)**	**Type**	**Consequence**	**Info**	**N_{seq}**	**Hom**	**Het**	**Fail**
AKR1A1	1	45,804,882	Indel	frameshift	0.98	56	3	0	1
LRRIQ3	1	74,394,016	Indel	frameshift	0.98	22	3	0	0
GPR113	2	26,394,482	Indel	frameshift	0.98	18	3	0	0
GCKR	2	27,598,876	SNP	stop gained	1.00	23	4	0	0
RETSAT	2	85,431,608	Indel	frameshift	0.97	44	2	0	1
NCKAP5	2	133,199,764	Indel	frameshift	0.97	10	3	1	0
RBM43	2	151,817,528	Indel	frameshift	0.97	46	4	0	0
LRTM1	3	54,927,549	Indel	frameshift	1.00	36	2	0	0
THPO	3	185,573,262	Indel	frameshift	0.99	21	1	2	0
SLC6A19	5	1,270,062	SNP	splice donor	0.98	40	1	0	0
IL31RA	5	55,231,568	Indel	frameshift	0.99	38	4	0	1
GPR98	5	90,043,957	SNP	stop gained	0.98	22	4	0	0
ERAP1	5	96,147,326	Indel	frameshift	0.98	41	4	0	0
C2	6	32,010,044	Indel	splice donor	0.99	95	11	0	1
CUL9	6	43,282,034	Indel	splice acceptor	0.99	41	4	0	0
AHI1	6	135,801,205	SNP	splice donor	0.94	10	1	1	0
GNAT3	7	79,946,241	SNP	stop gained	0.97	22	2	0	0
PPP1R3A	7	113,306,396	Indel	frameshift	0.97	30	3	0	0
RP1L1	8	10,507,566	Indel	frameshift	0.96	21	1	0	0
CNGB3	8	87,725,124	Indel	frameshift	1.00	28	4	0	1
FANCG	9	35,065,702	Indel	frameshift	1.00	7	1	0	0
DOLK	9	130,749,402	Indel	frameshift	0.94	50	2	0	0
PNPLA7	9	139,564,448	Indel	frameshift	0.97	45	1	0	0
CCDC7	10	32,896,804	Indel	frameshift	0.99	27	4	0	0
HPS1	10	100,168,005	SNP	splice acceptor	0.88	5	2	0	0
OR51G1	11	4,901,932	Indel	frameshift	0.98	43	1	0	0
PYGM	11	64,283,799	SNP	stop gained	0.99	19	1	0	0
CTSW	11	65,403,926	Indel	frameshift	0.99	22	4	0	0
NLRX1	11	118,555,861	Indel	frameshift	1.00	50	4	0	0
KLRF1	12	9,888,308	Indel	frameshift	1.00	18	3	0	0
GPRC5D	12	12,993,748	Indel	frameshift	1.00	39	2	0	0
RAD9B	12	109,434,981	Indel	frameshift	0.99	20	2	0	0
ATP7B	13	51,432,390	Indel	frameshift	0.99	31	3	0	0
ABHD4	14	22,145,205	Indel	frameshift	0.99	17	2	0	0
ARID4A	14	57,901,772	SNP	splice donor	1.00	16	2	0	0
SNAPC1	14	61,329,314	Indel	frameshift	0.98	51	7	0	0
MLH3	14	74,568,565	Indel	frameshift	0.98	21	2	0	0
SPTBN5	15	39,945,301	Indel	frameshift	0.97	24	2	0	0
TGM5	15	41,336,054	Indel	splice donor	0.98	10	2	0	0
RLBP1	15	87,554,642	SNP	stop gained	0.95	11	2	0	0
SYNM	15	97,490,739	Indel	frameshift	0.98	25	1	0	0
NOD2	16	49,321,279	Indel	frameshift	0.97	32	1	0	0
KIF19	17	69,852,027	Indel	frameshift	0.99	28	3	2	0
PGLS	19	17,492,874	Indel	frameshift	0.98	35	5	0	1
TUBB1	20	57,028,006	Indel	frameshift	1.00	16	1	0	0
SLCO4A1	20	60,766,798	Indel	frameshift	0.99	22	2	0	0
TMPRSS3	21	42,682,220	Indel	frameshift	0.95	10	2	0	0
Total	-	-	-	-	-	-	128	6	6

Supplementary Table 5. The results of Sanger sequencing individuals predicted to be homozygous for a set of rare LoF variants, enriched for indels over SNPs. Shown are the affected gene, chromosome, build 36 position, variant type (SNP or Indel), maximal

Nature Genetics: doi:10.1038/ng.3243

consequence predicted by VEP, imputation information, number of whole-genome sequenced carriers, and the number of Sanger sequenced individuals genotyped as homozygous, heterozygous (not matching the predicted genotype) and that could not be determined.

Nature Genetics: doi:10.1038/ng.3243

					Sanger genotype		
Gene	**Chr**	**Pos (build 36)**	**Type**	**Consequence**	**Match**	**Mismatch**	**Fail**
RPF1	1	84,728,000	SNP	splice donor	1	0	0
COL24A1	1	86,080,386	SNP	stop gained	0	0	1
GBP1	1	89,294,359	Indel	frameshift	1	0	0
KIAA1107	1	92,416,480	Indel	frameshift	1	0	0
FNBP1L	1	93,769,013	Indel	frameshift	1	0	0
ARHGAP29	1	94,469,582	Indel	frameshift	1	0	0
FNDC7	1	109,061,921	SNP	splice acceptor	1	0	0
ADAMTSL4	1	148,794,615	Indel	frameshift	1	0	0
ASH1L	1	153,584,271	SNP	stop gained	1	0	0
SLC25A44	1	154,444,412	Indel	frameshift	1	0	0
USF1	1	159,279,245	SNP	splice donor	1	0	0
RC3H1	1	172,196,936	SNP	stop gained	1	0	0
JMJD4	1	225,988,267	Indel	frameshift	1	0	0
OBSCN	1	226,513,846	Indel	frameshift	1	0	0
HEATR1	1	234,785,579	SNP	splice acceptor	1	0	0
OR14A16	1	246,045,218	Indel	frameshift	1	0	0
PGBD2	1	247,178,343	SNP	stop gained	1	0	0
PXDN	2	1,637,273	SNP	stop gained	1	0	0
LAPTM4A	2	20,100,645	SNP	stop gained	1	0	0
GCKR	2	27,595,227	Indel	frameshift	1	0	0
CRIM1	2	36,557,636	Indel	frameshift	1	0	0
SRBD1	2	45,469,964	Indel	frameshift	1	0	0
VPS54	2	64,052,875	Indel	frameshift	1	0	0
GAD1	2	171,408,810	SNP	stop gained	1	0	0
CCDC141	2	179,517,533	SNP	stop gained	1	0	0
TMEM194B	2	191,089,263	SNP	stop gained	1	0	0
HSPD1	2	198,067,124	SNP	splice donor	1	0	0
RPE	2	210,592,681	Indel	frameshift	1	0	0
ANKZF1	2	219,805,329	SNP	splice donor	1	0	0
ANKZF1	2	219,808,833	Indel	frameshift	1	0	0
AGAP1	2	236,622,528	SNP	stop gained	1	0	0
UBE2F	2	238,603,953	SNP	stop gained	1	0	0
OR6B3	2	240,633,259	Indel	frameshift	1	0	0
ATP2B2	3	10,345,775	Indel	frameshift	1	0	0
KRBOX1	3	42,957,805	SNP	stop gained	1	0	0
SNRK	3	43,359,929	Indel	frameshift	1	0	0
RASSF1	3	50,343,885	SNP	splice acceptor	1	0	0
KBTBD8	3	67,137,323	SNP	stop gained	1	0	0
SLC35A5	3	113,782,121	Indel	frameshift	1	0	0
CEP63	3	135,760,852	Indel	frameshift	1	0	0
PLSCR4	3	147,401,511	SNP	splice donor	1	0	0
GNB4	3	180,619,896	SNP	stop gained	1	0	0
CLDN1	3	191,513,354	SNP	splice donor	1	0	0
ATP13A4	3	194,703,123	SNP	splice acceptor	1	0	0
LRRC15	3	195,562,901	Indel	frameshift	1	0	0
ZNF721	4	428,035	SNP	stop gained	1	0	0
BOD1L1	4	13,210,220	Indel	frameshift	1	0	0
DHX15	4	24,138,728	Indel	frameshift	1	0	0
FRYL	4	48,286,518	SNP	splice donor	1	0	0
UGT2B11	4	70,114,689	Indel	frameshift	0	0	1
SEC31A	4	83,959,376	SNP	stop gained	1	0	0
PDE5A	4	120,703,435	SNP	splice donor	1	0	0
KIAA1109	4	123,458,761	Indel	frameshift	1	0	0

Nature Genetics: doi:10.1038/ng.3243

SPRY1	4	124,543,023	Indel	frameshift	1	0	0
ANKRD31	5	74,449,746	Indel	frameshift	1	0	0
ARSK	5	94,944,390	Indel	frameshift	1	0	0
SEPT8	5	132,126,120	Indel	frameshift	1	0	0
PCDHB7	5	140,534,842	SNP	stop gained	0	1	0
PCDHGA1	5	140,690,906	Indel	frameshift	1	0	0
TCERG1	5	145,823,310	Indel	frameshift	1	0	0
RIPK1	6	3,034,821	Indel	frameshift	1	0	0
TJAP1	6	43,581,406	Indel	frameshift	1	0	0
TDRD6	6	46,766,887	Indel	frameshift	1	0	0
PKHD1	6	52,038,772	Indel	frameshift	0	0	1
DOPEY1	6	83,923,681	Indel	frameshift	1	0	0
COQ3	6	99,924,209	Indel	frameshift	1	0	0
METTL24	6	110,750,769	SNP	splice acceptor	1	0	0
MED23	6	131,965,093	SNP	stop gained	1	0	0
MED23	6	131,967,064	Indel	frameshift	0	0	1
UTRN	6	145,144,830	Indel	frameshift	1	0	0
INTS1	7	1,478,549	SNP	splice acceptor	1	0	0
TRIM50	7	72,376,322	SNP	splice donor	0	1	0
ABCB1	7	87,012,099	Indel	frameshift	1	0	0
ZKSCAN5	7	98,967,205	Indel	frameshift	1	0	0
PUS7	7	104,930,102	SNP	splice donor	1	0	0
HBP1	7	106,627,976	Indel	frameshift	1	0	0
TUSC3	8	15,575,717	SNP	splice donor	1	0	0
TNFRSF10B	8	22,937,664	SNP	stop gained	1	0	0
EBF2	8	25,774,592	Indel	frameshift	1	0	0
ESCO2	8	27,702,411	Indel	frameshift	1	0	0
SDCBP	8	59,654,887	SNP	stop gained	1	0	0
ASPH	8	62,659,055	Indel	splice donor	1	0	0
GDAP1	8	75,438,881	Indel	frameshift	1	0	0
DCAF13	8	104,508,657	SNP	splice donor	1	0	0
FOCAD	9	20,897,164	Indel	frameshift	1	0	0
IFNW1	9	21,131,250	Indel	frameshift	1	0	0
POLR1E	9	37,476,805	SNP	splice donor	1	0	0
CNTNAP3	9	39,168,221	Indel	frameshift	0	0	1
RMI1	9	85,805,947	Indel	frameshift	1	0	0
PHF2	9	95,465,694	Indel	frameshift	1	0	0
RNF20	9	103,363,232	SNP	stop gained	1	0	0
EPB41L4B	9	111,060,351	Indel	frameshift	1	0	0
RC3H2	9	124,653,538	SNP	splice acceptor	1	0	0
CAMSAP1	9	137,881,802	Indel	splice donor	1	0	0
APBB1IP	10	26,842,478	Indel	frameshift	1	0	0
BAMBI	10	29,010,935	SNP	stop gained	1	0	0
KIAA1462	10	30,358,117	SNP	stop gained	1	0	0
C10orf68	10	33,057,932	SNP	splice donor	1	0	0
GJD4	10	35,936,511	SNP	splice acceptor	1	0	0
PCDH15	10	55,238,694	SNP	stop gained	1	0	0
PCDH15	10	56,094,027	SNP	initiator codon	1	0	0
CCAR1	10	70,202,759	Indel	frameshift	1	0	0
MMP26	11	4,967,664	Indel	frameshift	1	0	0
TRIM48	11	54,792,401	SNP	splice acceptor	1	0	0
UBE2L6	11	57,084,385	SNP	splice donor	1	0	0
PATL1	11	59,182,941	SNP	stop gained	1	0	0
TMEM135	11	86,708,068	SNP	splice donor	1	0	0
PPP2R1B	11	111,117,994	Indel	frameshift	1	0	0

Nature Genetics: doi:10.1038/ng.3243

ARHGEF12	11	119,805,652	SNP	stop gained	1	0	0
PRB4	12	11,352,511	SNP	stop gained	0	0	1
HEBP1	12	13,019,510	SNP	stop lost	1	0	0
LMBR1L	12	47,778,054	Indel	frameshift	1	0	0
MYL6B	12	54,835,239	Indel	frameshift	1	0	0
STAB2	12	102,680,237	Indel	frameshift	1	0	0
CCDC63	12	109,826,980	SNP	splice donor	1	0	0
COQ5	12	119,426,177	Indel	frameshift	1	0	0
ZNF140	12	132,193,175	SNP	stop gained	1	0	0
TRPC4	13	37,164,429	SNP	stop gained	1	0	0
PHF11	13	48,990,187	Indel	frameshift	1	0	0
PHF11	13	48,998,596	Indel	splice donor	1	0	0
NOP9	14	23,842,872	Indel	frameshift	1	0	0
BAZ1A	14	34,301,136	Indel	frameshift	1	0	0
RALGAPA1	14	35,217,012	SNP	stop gained	1	0	0
ABHD12B	14	50,438,332	Indel	frameshift	1	0	0
CCDC175	14	59,081,653	SNP	stop gained	1	0	0
SYNE2	14	63,618,026	Indel	frameshift	1	0	0
SEL1L	14	81,013,275	SNP	stop gained	1	0	0
ATG2B	14	95,852,687	SNP	stop gained	1	0	0
ANKDD1A	15	63,036,499	SNP	stop gained	1	0	0
SLC24A1	15	63,703,604	Indel	frameshift	1	0	0
AKAP13	15	83,990,144	Indel	frameshift	1	0	0
SLX4	16	3,580,804	SNP	stop gained	1	0	0
DOC2A	16	29,925,332	SNP	splice acceptor	1	0	0
NOD2	16	49,321,314	SNP	splice donor	1	0	0
CES3	16	65,555,260	Indel	frameshift	1	0	0
FA2H	16	73,307,795	SNP	stop gained	1	0	0
DNAH2	17	7,645,627	SNP	stop gained	1	0	0
KRBA2	17	8,213,920	Indel	frameshift	1	0	0
ABCC3	17	46,107,722	Indel	frameshift	1	0	0
STXBP4	17	50,479,512	Indel	splice donor	1	0	0
FBF1	17	71,430,875	Indel	frameshift	1	0	0
SMCHD1	18	2,740,487	Indel	frameshift	1	0	0
PPP4R1	18	9,573,114	SNP	splice donor	1	0	0
AFG3L2	18	12,367,037	SNP	stop gained	0	0	1
TRAPPC8	18	27,765,477	Indel	frameshift	1	0	0
SERPINB13	18	59,406,864	SNP	splice acceptor	1	0	0
R3HDM4	19	851,108	Indel	frameshift	1	0	0
DENND1C	19	6,419,658	SNP	splice acceptor	1	0	0
C19orf45	19	7,476,323	SNP	splice donor	1	0	0
MYO1F	19	8,501,250	SNP	splice acceptor	1	0	0
MUC16	19	8,934,495	SNP	stop gained	1	0	0
JAK3	19	17,814,916	Indel	frameshift	1	0	0
ZNF793	19	42,705,271	Indel	splice acceptor	1	0	0
FCGBP	19	45,049,601	SNP	splice acceptor	1	0	0
PSG11	19	48,220,940	SNP	stop gained	0	1	0
TMEM150B	19	60,524,154	SNP	stop gained	1	0	0
CDS2	20	5,113,502	SNP	splice acceptor	1	0	0
DZANK1	20	18,383,966	Indel	frameshift	1	0	0
KIF3B	20	30,361,887	Indel	frameshift	1	0	0
LSS	21	46,435,572	Indel	frameshift	1	0	0
GSTT1	22	22,706,821	SNP	splice donor	1	0	0
SLC5A4	22	30,950,354	Indel	frameshift	1	0	0
Total	-	-	-	-	152	3	7

Nature Genetics: doi:10.1038/ng.3243

Supplementary Table 6. The results of Sanger sequencing individuals who were observed to be the only carrier of a LoF variant among the 2,636 whole-genome sequenced Icelanders. Shown are the affected gene, chromosome, build 36 position, variant type (SNP or Indel), maximal consequence predicted by VEP and indicators of whether the Sangers sequencing was successful and matched the whole-genome sequencing.

Nature Genetics: doi:10.1038/ng.3243

	All frequencies	MAF < 2%	MAF < 0.5%
Number of genes	402		
Number of genes with LoF variants	117	79	67
Number of LoF variants	140	94	77
Completely knocked out genes	-	34	20
Completely knocked out individuals	-	251	43
Variants contributing to knockout	-	39	21

Supplementary Table 7. The number of sensory perception of smell genes affected by LoF variants and individuals that have these genes completely knocked out by LoF variants (GO:0007608).

Nature Genetics: doi:10.1038/ng.3243

	Sequenced and phased		Imputed	
	Homozygote	**Heterozygote**	**Homozygote**	**Heterozygote**
MAF < 0.5%	0.018	9.4	0.016	8.5
0.5% < MAF < 2%	0.066	9.5	0.056	8.9
MAF > 2%	21.0	111.1	20.5	110.0
Total	21.1	130.0	20.6	127.4

Supplementary Table 8. Average LoF minor allele genotype counts per individual stratified on minor allele frequency (MAF).

Nature Genetics: doi:10.1038/ng.3243

Tissue	N genes	Median coding size	Knocked out genes			
			MAF < 2 %		MAF < 0.5 %	
			N	Percentage (95% CI)	N	Percentage (95% CI)
brain	318	1,407	4	1.3 (0.0-2.5)	3	0.9 (0.0-2.0)
placenta	71	1,008	4	5.6 (0.1-11.2)	4	5.6 (0.1-11.1)
skin	104	1,078	6	5.8 (1.2-10.4)	6	5.8 (1.2-10.4)
kidney	62	1,413	4	6.5 (0.2-12.7)	2	3.2 (0.0-7.7)
pancreas	44	906	3	6.8 (0.0-14.5)	2	4.5 (0.0-11.0)
liver	171	1,377	13	7.6 (3.5-11.7)	12	7.0 (3.1-11.0)
esophagus	61	1,207	5	8.2 (1.1-15.3)	4	6.6 (0.1-13.1)
bone marrow	82	768	7	8.5 (2.2-14.9)	1	1.2 (0.0-3.6)
salivary gland	46	627	4	8.7 (0.1-17.2)	3	6.5 (0.0-13.9)
testis	885	1,156	102	11.5 (9.4-13.7)	72	8.1 (6.3-9.9)
heart	98	1,260	12	12.2 (5.4-19.1)	9	9.2 (3.2-15.1)

Supplementary Table 9. Counts of genes completely knocked out by rare LoF variants for genes with human tissue-specific expression [28]. Counts are shown for tissues that have at least 40 (autosomal) tissue-specific genes.

Nature Genetics: doi:10.1038/ng.3243

			Knocked out genes							
			MAF < 2%				MAF < 0.5%			
			RefSeq genes		Genes linked to conditions under a recessive mode of inheritance [26]		RefSeq genes		Genes linked to conditions under a recessive mode of inheritance [26]	
Mouse phenotype	N	Median coding size	N	Percentage (95% CI)	Severe diseases [27]	Other conditions[c]	N	Percentage (95% CI)	Severe diseases [27]	Other conditions[c]
craniofacial	1,005	1,647	23	2.3 (1.4,3.2)	3	5	17	1.7 (0.9,2.5)	3	3
pigmentation	362	1,608	9	2.5 (0.9,4.1)	2	4	8	2.2 (0.7,3.7)	1	4
embryogenesis	1,530	1,575	43	2.8 (2.0,3.6)	4	5	33	2.2 (1.4,2.9)	4	4
digestive/alimentary	1,142	1,554	35	3.1 (2.1,4.1)	4	3	26	2.3 (1.4,3.1)	4	3
skeleton	1,483	1,509	49	3.3 (2.4,4.2)	1	8	37	2.5 (1.7,3.3)	1	6
mortality/aging	3,897	1,554	140	3.6 (3.0,4.2)	13	15	92	2.4 (1.9,2.8)	10	9
adipose tissue	647	1,509	24	3.7 (2.2,5.2)	2	6	19	2.9 (1.6,4.3)	2	6
growth/size/body	3,142	1,560	116	3.7 (3.1,4.3)	12	18	76	2.4 (1.9,2.9)	9	12
muscle	1,224	1,569	45	3.7 (2.6,4.7)	2	5	29	2.4 (1.5,3.2)	1	3
cardiovascular system	2,004	1,567	77	3.8 (3.0,4.7)	5	13	55	2.7 (2.0,3.4)	4	11
nervous system	2,814	1,599	108	3.8 (3.1,4.5)	11	17	78	2.8 (2.2,3.4)	9	12
integument	1,519	1,581	60	3.9 (3.0,4.9)	4	7	45	3.0 (2.1,3.8)	3	7
vision/eye	1,187	1,566	46	3.9 (2.8,5.0)	7	12	38	3.2 (2.2,4.2)	7	11
endocrine/exocrine gland	1,474	1,467	59	4.0 (3.0,5.0)	5	8	46	3.1 (2.3,4.0)	4	7
cellular	2,792	1,527	114	4.1 (3.4,4.8)	8	10	85	3.0 (2.4,3.7)	6	8
respiratory system	1,149	1,513	48	4.2 (3.0,5.3)	5	4	31	2.7 (1.8,3.6)	3	3
behavior/neurological	2,462	1,527	107	4.3 (3.6,5.1)	11	13	76	3.1 (2.4,3.7)	8	10
limbs/digits/tail	776	1,611	34	4.4 (2.9,5.8)	0	5	23	3.0 (1.8,4.1)	0	3
liver/biliary system	992	1,515	46	4.6 (3.3,6.0)	5	4	34	3.4 (2.3,4.6)	3	4
hearing/vestibular/ear	529	1,593	25	4.7 (2.9,6.5)	2	11	19	3.6 (2.0,5.2)	2	8
hematopoietic system	2,458	1,446	119	4.8 (4.0,5.7)	7	14	80	3.3 (2.6,3.9)	5	10
tumorigenesis	773	1,519	37	4.8 (3.3,6.3)	2	3	26	3.4 (2.1,4.6)	2	3
reproductive system	1,579	1,512	79	5.0 (3.9,6.1)	5	11	62	3.9 (3.0,4.9)	4	9
immune system	2,651	1,450	136	5.1 (4.3,5.9)	7	13	90	3.4 (2.7,4.1)	5	11
renal/urinary system	1,004	1,573	53	5.3 (3.9,6.7)	7	11	41	4.1 (2.9,5.3)	6	9
homeostasis/metabolism	3,422	1,467	186	5.4 (4.7,6.2)	13	22	129	3.8 (3.2,4.4)	10	18

Nature Genetics: doi:10.1038/ng.3243

other	270	1,638	17	6.3 (3.3,9.3)	3	2	9	3.3 (1.2,5.5)	2	2
taste/olfaction	120	1,522	11	9.2 (3.7,14.6)	0	0	10	8.3 (3.2,13.5)	0	0
Any mouse phenotype[a]	7,174	1,485	361 (2,240)	5.0 (4.6,5.4)	16 (47)	44 (237)	262 (583)	3.7 (3.3,4.0)	13 (27)	32 (58)

Supplementary Table 10. Counts of gene and individual knock-outs with rare variants (DAF < 2%) for human orthologs of mouse genes with phenotypes in the Mouse Genome Informatics (MGI) database. [a]Number of individuals with a complete knockout in a human ortholog of a gene with mouse phenotype is shown in parenthesis.

Nature Genetics: doi:10.1038/ng.3243

Class	N markers	MAF (%)			N het. parent	N minor transm.	Minor transm. frequency (SE)	N both parents het.	N double minor transm.	Double minor transm. frequency (SE)
		Thresh	Med	Mean						
LoF	2,741	<0.5	0.11	0.15	566,454	282,556	0.49882 (0.00068)	932	208	0.2232 (0.0151)
LoF	3,059	<1	0.12	0.20	867,035	432,649	0.49900 (0.00055)	2,205	504	0.2286 (0.0098)
LoF	3,235	<2	0.13	0.27	1,204,213	601,849	0.49979 (0.00046)	4,860	1,149	0.2364 (0.0064)
MODERATE	50,079	<0.5	0.12	0.16	11,146,218	5,570,282	0.49975 (0.00017)	19,383	4,772	0.2462 (0.0038)
MODERATE	57,943	<1	0.15	0.23	18,809,718	9,404,723	0.49999 (0.00013)	51,641	12,760	0.2471 (0.0024)
MODERATE	63,601	<2	0.17	0.34	29,899,791	14,951,803	0.50006 (0.00010)	140,233	34,860	0.2486 (0.0014)
Intergenic	3,849,601	<0.5	0.12	0.16	930,068,954	464,961,218	0.49992 (0.00007)	1,677,095	414,847	0.2474 (0.0019)
Integenic	4,571,969	<1	0.15	0.25	1,681,403,646	840,702,580	0.50000 (0.00006)	4,914,196	1,220,935	0.2485 (0.0012)
Intergenic	5,193,617	<2	0.19	0.39	2,986,623,050	1,493,458,184	0.50005 (0.00005)	15,552,151	3,878,138	0.2494 (0.0007)
LoF:										
RVIS 0-20% [29]	429	<2	0.11	0.20	115,940	57,783	0.49839 (0.00150)	290	52	0.1793 (0.0266)
RVIS 20-40%	380	<2	0.11	0.22	121,432	60,692	0.49980 (0.00146)	464	96	0.2069 (0.0214)
RVIS 40-60%	388	<2	0.13	0.28	155,663	77,726	0.49932 (0.00127)	752	211	0.2806 (0.0156)
RVIS 60-80%	529	<2	0.14	0.26	211,012	105,381	0.49941 (0.00112)	770	162	0.2104 (0.0167)
RVIS 80-100%	1,131	<2	0.15	0.31	487,101	243,595	0.50009 (0.00072)	2,184	542	0.2482 (0.0095)
Essential 0-20% [30]	150	<2	0.13	0.21	47,533	23,550	0.49545 (0.00231)	150	26	0.1733 (0.0376)
Essential 20-40%	214	<2	0.15	0.27	73,165	36,851	0.50367 (0.00184)	285	54	0.1895 (0.0274)
Essential 40-60%	220	<2	0.13	0.30	90,284	45,206	0.50071 (0.00169)	369	95	0.2575 (0.0226)
Essential 60-80%	215	<2	0.15	0.30	84,794	42,258	0.49836 (0.00174)	346	87	0.2514 (0.0241)
Essential 80-100%	203	<2	0.12	0.29	82,759	41,411	0.50038 (0.00175)	357	91	0.2549 (0.0221)
Severe recessive [27]	95	<2	0.13	0.21	29,112	14,607	0.50175 (0.00348)	72	14	0.1944 (0.0572)

Supplementary Table 11. Transmission from single heterozygouse parents and two heterozyougs parents. For the Residual Variation Intolerance Score (RVIS), essential, recessive and Kingsmore classes only LoF variants were considered. Genes were divided into five groups by their RVIS (available for 18,329 genes) and the essential score (available for 7,114 genes) percentiles.

Nature Genetics: doi:10.1038/ng.3243

Supplementary References

1. Gudbjartsson, D.F. Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics* **In press**(2015).
2. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068-1075 (2008).
3. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868-874 (2009).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
5. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
6. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
7. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
8. 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**, 906-13 (2007).
9. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
10. Gudbjartsson, D. The Sequence of Icelanders. *Submitted to Nature Genetics* (2014).
11. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
12. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
13. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
14. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40**, D84-90 (2012).
15. 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**, D130-5 (2012).
16. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* **6**, R44 (2005).
17. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
18. Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-51 (2011).
19. Efron, B. & Tibshirani, R. *An introduction to the bootstrap*, xvi, 436 p. (Chapman & Hall, New York, 1993).
20. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-78 (2012).
21. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
22. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
24. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**, 397-406 (2014).
25. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A. & Richardson, J.E. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res* **42**, D810-7 (2014).

Nature Genetics: doi:10.1038/ng.3243

26. Kohler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* **42**, D966-74 (2014).
27. Saunders, C.J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med* **4**, 154ra135 (2012).
28. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
29. 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).
30. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-4 (2014).

Nature Genetics: doi:10.1038/ng.3243