nature genetics

Identification of a large set of rare complete human knockouts

Patrick Sulem[1,6], Hannes Helgason[1,2,6], Asmundur Oddson[1], Hreinn Stefansson[1], Sigurjon A Gudjonsson[1], Florian Zink[1], Eirikur Hjartarson[1], Gunnar Th Sigurdsson[1], Adalbjorg Jonasdottir[1], Aslaug Jonasdottir[1], Asgeir Sigurdsson[1], Olafur Th Magnusson[1], Augustine Kong[1,2], Agnar Helgason[1,3], Hilma Holm[1,4], Unnur Thorsteinsdottir[1,5], Gisli Masson[1], Daniel F Gudbjartsson[1,2] & Kari Stefansson[1,5]

Loss-of-function mutations cause many mendelian diseases. Here we aimed to create a catalog of autosomal genes that are completely knocked out in humans by rare loss-of-function mutations. We sequenced the whole genomes of 2,636 Icelanders and imputed the sequence variants identified in this set into 101,584 additional chip-genotyped and phased Icelanders. We found a total of 6,795 autosomal loss-of-function SNPs and indels in 4,924 genes. Of the genotyped Icelanders, 7.7% are homozygotes or compound heterozygotes for loss-of-function mutations with a minor allele frequency (MAF) below 2% in 1,171 genes (complete knockouts). Genes that are highly expressed in the brain are less often completely knocked out than other genes. Homozygous loss-of-function offspring of two heterozygous parents occurred less frequently than expected (deficit of 136 per 10,000 transmissions for variants with MAF <2%, 95% confidence interval (CI) = 10–261).

Loss-of-function variants, primarily stop-gain variants, frameshift indels and essential splice-site variants, are predicted to disrupt the protein encoded by the gene and therefore have the greatest pathogenic potential (**Supplementary Table 1**)[1–6]. An unanswered question in human genetics is what is the population frequency of homozygous loss-of-function mutations in the germline genome? A related question is how frequently they occur without deleterious phenotypic consequences, addressing the issue of biochemical redundancy. The recent surge in sequencing capacity allows us to address these questions[7,8]. Another opportunity provided by the new technology is to turn the classic paradigm upside down and rather than searching for sequence variants that are responsible for phenotypic characteristics to search for phenotypic characteristics that are caused by variants in the sequence. Our approach is to map out the large majority of homozygous loss-of-function mutations in the Icelandic population, allowing subsequent systematic phenotyping of the variant carriers.

We sequenced 2,636 Icelanders participating in various disease projects to a median depth of 20× (**Supplementary Tables 2** and **3**). We observed 6,795 loss-of-function mutations in 4,924 genes among the sequenced Icelanders; 3,979 SNPs and 2,816 indels (**Supplementary Table 4**). Although only 7% of all sequence variants were indels, they were over-represented in the loss-of-function category, owing to their tendency to shift the reading frame when they occur in exons[5,9,10]. A single loss-of-function mutation was observed in 3,603 genes and 2 or more mutations were observed in 1,321 genes (**Table 1**). Most loss-of-function mutations were rare, with 85% having a MAF below 0.5%. Indeed, loss-of-function mutations had the highest fraction of rare variants of all the functional annotation classes[2,3,5,11].

Mendelian autosomal recessive diseases occur when both copies of a particular gene are affected by mutations, often loss-of-function mutations. The effect of loss-of-function mutations on the sex chromosomes are different in nature to those on autosomal chromosomes because males have a single copy of the X and Y chromosomes[12] and X inactivation occurs in females[13]. Here we restricted our analysis to autosomal loss-of-function variants. Cystic fibrosis is the most common mendelian autosomal recessive disease in northern Europeans, with an incidence of 1 in 3,200 live births that corresponds to a MAF of 1.8% of the causative mutation[14]. Highly penetrant recessive variants that are more common would cause diseases with a higher incidence. Because such diseases are not known, we restricted our analysis of completely knocked out genes to variants with a MAF under 2%. A mutation causing a recessive disease will be present in unaffected heterozygotes, and its detection will not require sequencing homozygotes affected with the disease. Most variants with a MAF over 0.2% will be observed more than 5 times in the set of 2,636 sequenced Icelanders, which is usually a sufficient number for accurate imputation (**Supplementary Fig. 1**). One individual in 250,000 is expected to be homozygous for a variant with a MAF of 0.2% under Hardy-Weinberg equilibrium. On the basis of whole-genome sequencing and imputation into a total of 104,220 individuals recruited as cases, relatives of cases and controls without any exclusion criteria, through the study of several common diseases, we identified 8,041 individuals (7.7%) who had 1 gene completely knocked out by loss-of-function

[1]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]Department of Anthropology, University of Iceland, Reykjavik, Iceland. [4]Department of Internal Medicine, Landspitali The National University Hospital of Iceland, Reykjavik, Iceland. [5]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [6]These authors contributed equally to this work. Correspondence should be addressed to P.S. (patrick.sulem@decode.is) or K.S. (kari.stefansson@decode.is).

Received 17 April 2014; accepted 13 February 2015; published online 25 March 2015; doi:10.1038/ng.3243

Table 1 Number of loss-of-function variants by sequence ontology (SO) term and MAF

	MAF < 0.5%		0.5% < MAF < 2%		MAF > 2%			
SO term	SNPs	Indels	SNPs	Indels	SNPs	Indels	Total variants	Genes
Splice donor	673	94	51	10	61	12	901	851
Splice acceptor	459	80	49	4	33	11	636	610
Stop gain	1,976	4	164	0	144	1	2,289	1,993
Frameshift	0	2,218	0	196	0	184	2,598	2,199
Stop loss	67	0	11	0	22	0	100	97
Initiator codon	202	2	25	0	42	0	271	269
Total	3,377	2,398	300	210	302	208	6,795	4,924

A total of 3,603 genes had a single loss-of-function variant, 956 had 2 loss-of-function variants and 365 had 3 or more loss-of-function variants.

variants with a MAF under 2% (**Table 2** and **Supplementary Fig. 2**). This constitutes a large sample from the Icelandic population, which currently numbers 320,000 individuals. We note that, although most variants that are likely to produce complete knockouts in the Icelandic population had been imputed, many more rare loss-of-function mutations will exist in the population (**Supplementary Fig. 3**). The long-range phasing of haplotypes allows determination of whether two different loss-of-function variants, observed in the same gene, are derived from different parental chromosomes, constituting a compound heterozygote. Of the 8,041 individuals with complete knockout, 6,885 were predicted to be homozygotes, 1,249 were predicted to be compound heterozygotes and 553 were predicted to have more than 1 gene completely knocked out. In these individuals, a total of 1,171 of the 19,135 RefSeq genes (6.1%)[15] were completely knocked out. We observed 5 or fewer complete knockouts for 790 of these 1,171 genes (**Supplementary Fig. 4**).

We were able to perform Sanger sequencing on 134 individuals imputed to be homozygous for rare loss-of-function variants and 155 individuals who were the sole sequenced carriers of loss-of-function mutations, and we confirmed 96% and 98% of their genotypes, respectively (**Supplementary Tables 5** and **6**).

The class of olfactory receptor genes had the highest density of loss-of-function variants and the lowest fraction of rare loss-of-function variants of all gene ontology classes[2,5]. Olfactory receptor genes constituted 2.9% (34) of the genes that were completely knocked out by loss-of-function variants with a MAF under 2%, and 3.1% (251) of the individuals had an olfactory receptor gene completely knocked out (**Supplementary Table 7**).

Currently, 1,717 genes have been linked to a condition through a recessive mode of inheritance[8,16–18]. We predicted that 546 individuals in our data set would have 1 of 88 of these 1,717 genes completely knocked out by loss-of-function variants with a MAF below 2% (**Supplementary Note**). Eighty-seven of these individuals had mutations in 26 genes that were the same as those that had been recorded as disease-causing mutations in the Human Genome Mutation Database (HGMD; DM category)[19].

We observed somewhat fewer loss-of-function carriers among the imputed individuals than among the sequenced and phased ones (**Supplementary Fig. 5** and **Supplementary Table 8**). Mutations occurring in the last few generations would be missed more often in the imputed set than in the sequenced one, and some carriers would not be imputed as such with sufficiently high certainty.

The Icelandic population was founded 1,100 years ago by 8–20 thousand settlers[20]. Since then, the population grew from 50,000 in the 1703 census to 320,000 today (Statistics Iceland; see URLs). Mating is not random in the population. In particular, Icelanders are more likely to select a spouse from their own geographical region, which leads to an excess of homozygotes for rare variants, over what Hardy-Weinberg equilibrium would predict[5]. However, consanguineous unions (between second cousins or closer) are not frequent. Of the parents of the sequenced and imputed individuals in our study, 2.7% were second cousins or closer (**Supplementary Fig. 6**). Of the 8,041 individuals with at least 1 gene completely knocked out by rare loss-of-function variants, 90.1% were not the children of parents who were second cousins or closer.

We next investigated whether the probability of a gene being completely knocked out depended on the tissue in which it was expressed. We used the results of Fagerberg *et al.*[21] to find genes that were highly expressed (FPKM (fragments per kilobase of exon per million fragments mapped) over 20) in 1 or more of 27 tissues, but excluding genes that were highly expressed in all the tissues examined. We then calculated the fraction of genes expressed in each tissue corresponding to at least one completely knocked-out individual by loss-of-function variants with a MAF below 2% in our data (**Table 3** and **Supplementary Table 9**). The lowest fraction of genes with completely knocked-out individuals was observed in the brain (3.1%) and placenta (3.9%), and the highest fraction was observed in the testis (5.8%), small intestine (5.4%) and duodenum (5.8%).

Mouse knockouts have been used to model the effects of sequence variation on human phenotype. Of the 1,171 genes with complete knockout by rare variants (MAF < 2%), 361 were orthologs of mouse genes that have been reported to affect a mouse phenotype when mutated, according to the Mouse Genome Informatics (MGI) database[22] (**Supplementary Table 10**). Of the 28 classes of mouse phenotypes in this database, the taste/olfaction class (9.2%) by far showed the greatest proportion of completely knocked-out genes with mouse orthologs. Conversely, the classes with the smallest proportions of genes with complete knockout were the craniofacial (2.3%), pigmentation (2.5%) and embryogenesis (2.8%) mouse phenotype classes.

Table 2 Counts of completely knocked-out genes and individuals by loss-of-function variants among 104,220 genotyped Icelanders

		Genes linked to conditions under a recessive mode of inheritance[19]	
	RefSeq genes	Severe diseases[18]	Other conditions[c]
Number of genes	19,135	437	1,280
Completely knocked-out genes[a]	1,171 (775)	20 (15)	68 (46)
Individuals with completely knocked-out genes[b]	8,041 (1,741)	70 (29)	476 (86)
Restricted to HGMD-reported variants[20]			
Completely knocked-out genes	–	9 (8)	17 (11)
Completely knocked-out individuals	–	19 (15)	68 (17)

Counts are given for variants with MAF below 2% and for variants with MAF below 0.5% (in parentheses).

[a]In total, 1,485 (907) distinct loss-of-function variants with MAF <2% (<0.5%) contribute to knockout of the 8,041 (1,741) individuals in 1 of the 1,171 (775) genes. [b]For loss-of-function variants with MAF <2% (<0.5%), the individuals with at least 1 completely knocked-out gene consist of 6,885 (1,550) homozygotes for rare loss-of-function variants and 1,249 (193) compound heterozygotes for rare loss-of-function variants; in total, 7,488 (1,663) individuals have 1 gene completely knocked out and 553 (78) individuals have 2 or more genes completely knocked out. [c]These genes originated from the Human Phenotype Ontology (HPO) recessive gene list after excluding recessive severe genes. The HPO recessive gene list is based on Online Mendelian Inheritance in Man (OMIM).

Table 3 Counts of genes completely knocked out by rare loss-of-function variants that are well expressed in some but not all tissues

			Knocked-out genes			
			MAF < 2%		MAF < 0.5%	
Tissue	*n* genes	Median coding size	*n*	Percentage (95% CI)	*n*	Percentage (95% CI)
Brain	2,023	1,452	62	3.1 (2.3–3.8)	44	2.2 (1.6–2.8)
Placenta	2,005	1,356	78	3.9 (3.1–4.7)	50	2.5 (1.8–3.2)
Adrenal gland	1,888	1,245	78	4.1 (3.2–5.0)	48	2.5 (1.8–3.2)
Ovary	1,918	1,470	79	4.1 (3.2–5.0)	49	2.6 (1.9–3.2)
Endometrium	2,017	1,434	87	4.3 (3.4–5.2)	52	2.6 (1.9–3.2)
Esophagus	2,182	1,355	100	4.6 (3.7–5.4)	62	2.8 (2.2–3.5)
Thyroid gland	2,213	1,392	102	4.6 (3.8–5.5)	65	2.9 (2.2–3.6)
Kidney	2,033	1,306	95	4.7 (3.8–5.6)	60	3.0 (2.2–3.7)
Lymph node	2,261	1,386	107	4.7 (3.9–5.6)	80	3.5 (2.8–4.3)
Pancreas	616	1,221	29	4.7 (3.0–6.4)	15	2.4 (1.2–3.7)
Salivary gland	1,366	1,327	64	4.7 (3.6–5.8)	41	3.0 (2.1–3.9)
Heart	1,727	1,270	83	4.8 (3.8–5.8)	53	3.1 (2.3–3.9)
Urinary bladder	2,448	1,344	117	4.8 (4.0–5.6)	75	3.1 (2.4–3.7)
Adipose tissue	1,722	1,314	84	4.9 (3.9–5.9)	51	3.0 (2.2–3.7)
Prostate	1,914	1,309	93	4.9 (3.9–5.8)	57	3.0 (2.2–3.7)
Appendix	2,090	1,395	105	5.0 (4.1–5.9)	73	3.5 (2.7–4.3)
Colon	2,101	1,290	104	5.0 (4.0–5.9)	64	3.0 (2.3–3.8)
Bone marrow	1,705	1,432	87	5.1 (4.1–6.1)	54	3.2 (2.4–4.0)
Lung	2,088	1,395	108	5.2 (4.2–6.1)	68	3.3 (2.5–4.0)
Liver	1,516	1,221	80	5.3 (4.2–6.4)	52	3.4 (2.5–4.3)
Gall bladder	2,323	1,386	126	5.4 (4.5–6.3)	82	3.5 (2.8–4.3)
Spleen	2,095	1,491	118	5.6 (4.7–6.6)	80	3.8 (3.0–4.6)
Stomach	1,854	1,329	103	5.6 (4.5–6.6)	68	3.7 (2.8–4.5)
Skin	1,752	1,545	99	5.7 (4.6–6.7)	60	3.4 (2.6–4.3)
Testis	2,911	1,503	169	5.8 (5.0–6.6)	118	4.1 (3.4–4.7)
Small intestine	1,968	1,327	126	6.4 (5.3–7.5)	81	4.1 (3.3–5.0)
Duodenum	1,910	1,305	132	6.9 (5.8–8.0)	83	4.3 (3.5–5.2)

Shown are the number of genes in the class, the median size of the coding regions (the size of the union of all transcripts of the gene) of the genes in the class, and the number of genes and the percentage of all the genes in the class (with 95% CI) that are completely knocked out for at least one individual for variants with MAF below both 2% and 0.5%.

Complete knockouts of several genes are known to die early or suffer from severe disease[17]. To control for the excess of homozygotes for rare variants in the Icelandic population due to non-random mating[5], we estimated the probability of parents who are both heterozygous for the same variant transmitting two copies of the minor allele to their offspring. We restricted the analysis to variants with accurate imputation (info > 0.9) and only considered genotypes that were imputed below 0.01 or above 0.99 for each allele of the individual. We calculated transmission probabilities for loss-of-function variants (n = 3,235 variants with MAF <2%), moderate-impact variants (missense, in-frame indel and splice-region variants, n = 63,601 variants with MAF <2%) and intergenic variants (variants more than 5 kb from a RefSeq gene, n = 5,193,617 variants with MAF <2%). We examined transmissions to 26,188 offspring on the basis of genotypes imputed into chip-typed individuals (**Fig. 1** and **Supplementary Table 11**). In comparison to the expectation under mendelian inheritance (25%), we observed a deficit of 136 double transmissions of the minor allele of loss-of-function variants with a MAF below 2% per 10,000 transmissions from a pair of heterozygous parents (P = 0.034, 95% CI = 10–261 per 10,000 transmission), with 1,149 of 4,860 possible double transmissions corresponding to a double-transmission probability of 23.6%. This deficit in loss-of-function variants was greater than was observed for the intergenic variants with MAF below 2% (P = 0.045). The observed number of transmissions from single heterozygous parents would predict deficits of 2 per 10,000 transmissions for loss-of-function variants with a MAF below 2%. We did not observe a significant deficit in double transmissions of moderate-impact or intergenic variants. We partitioned the transmissions of loss-of-function variants with a MAF below 2% by Residual Variation Intolerance Score (RVIS; available for 18,329 genes) percentile[23] and percentile in a negative selection screen for essential genes (available for 7,114 genes)[24]. For both scores, the most sensitive genes (first quintiles) had the lowest double-transmission rates (**Fig. 1**). Double transmissions of loss-of-function variants with MAF <2% in genes linked to severe recessive diseases[17] (437 genes) were estimated at 19% (P = 0.33).

The observed deficit in double transmissions must be because either homozygotes are missing from the population, owing to early death or the variants being embryonic lethal in the homozygous state, or homozygotes are undersampled, possibly because of illness or disability. For each of the loss-of-function variants, we give the number of homozygotes, the number of deaths among homozygotes, the earliest age at death for a known homozygote, the age of the longest-lived known homozygote, the number of offspring had by known carrier couples, the number of offspring of carrier couples who died before 15 years and the expected number of homozygotes in the population of 104,220 under Hardy-Weinberg equilibrium (**Supplementary Table 4**). We never observed a splice acceptor variant, c.964–1G>C (rs138649167, MAF = 1.4%), in *DHCR7* (encoding 7-dehydrocholesterol reductase), in the homozygous state although we predicted 19.1 homozygotes to be present among the 104,220 individuals ($P = 5.1 \times 10^{-9}$). Mutations in *DHCR7* cause Smith-Lemli-Optiz syndrome under a recessive mode of inheritance[25], a syndrome causing embryonic and early-age lethality[26]. In our data, 2 of 38 children of heterozygous parents were recorded to die in their first year (**Supplementary Table 4**).

Variants that are predicted to affect a particular gene transcript may have no effect on other transcripts of the same gene. Of the loss-of-function variants, 74.2% had a loss-of-function effect on all the transcripts of the affected gene. The frequency of variants in genes and genomic regions gives information about the nature and strength of selection acting on them[5,23,27]. 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%. The FRV among loss-of-function variants that had an effect on all transcripts of the affected gene was 88.0%, whereas the FRV among loss-of-function variants that only affected a subset of the transcripts of the gene was 81.4%. The FRV among both groups was thus substantially greater than for the moderate-impact variants, which had an FRV of 74.2% (ref. 5).

We examined the allele-specific expression of 215 stop-gain SNPs in 262 Icelanders with white blood cell mRNA sequencing data. Substantially less than half the alleles observed through mRNA sequencing were non-reference alleles (0.36, 95% CI = 0.33–0.38). This proportion is consistent with nonsense-mediated decay of transcripts with premature stop codons[2,28,29] (**Fig. 2**) and is substantially lower than the allele-specific expression of frequency-matched synonymous SNPs (0.463, 95% CI = 0.458–0.468; **Supplementary Fig. 7**). Stop-gain mutations in the middle exons of genes have lower non-reference allele fractions than stop-gain mutations in the first (P = 0.0041) or last ($P = 1.1 \times 10^{-6}$) exon (**Fig. 3**), in contrast to frequency-matched synonymous SNPs whose allele-specific expression did not depend on the position of the variant (**Supplementary Fig. 7**). These results are consistent with the lowest density and highest fraction of rare loss-of-function variants in the middle exons of genes[5]. The fraction of non-reference alleles among stop-gain SNPs in middle exons correlated positively with MAF (the correlation between log(MAF) and the non-reference allele fraction was 0.31, $P = 4.9 \times 10^{-4}$).

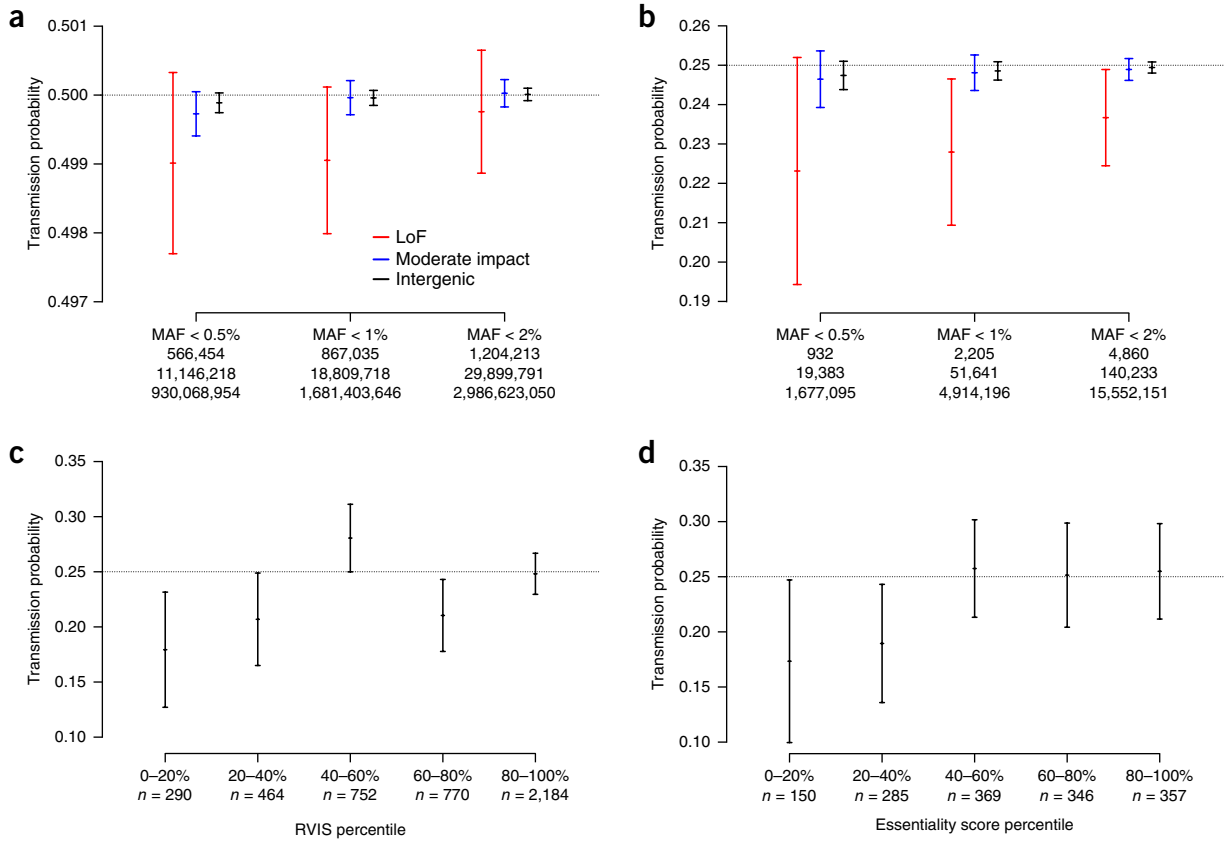

Figure 1 Transmission probabilities from carrier parents. (**a**–**d**) Transmission probabilities from a single heterozygous parent (**a**) and two heterozygous parents (**b**), with the transmissions of loss-of-function variants being further stratified by Residual Variation Intolerance Score (RVIS) (**c**) and essentiality score (**d**) percentiles. Transmissions observed in 35,024 father-offspring pairs and 47,769 mother-offspring pairs were used for the single-heterozygous-parent calculations, and transmissions observed in 26,188 triads were used for the two-heterozygous-parent calculations. The numbers of informative transmissions, where both parents were heterozygous for an indicated type of variant, are shown below each graph. Shown are the transmission probabilities of loss-of-function (LoF; red), moderate-impact (blue) and intergenic (black) sequence variants with MAF below 0.5%, 1% and 2% in **a** and **b** and the transmission probabilities for loss-of-function variants with MAF below 2% in **c** and **d**. The middle tick of each colored segment indicates the observed transmission probability, and the extreme ticks indicate the 95% confidence intervals estimated by bootstrap sampling. The dotted lines correspond to transmission probabilities under mendelian inheritance.

Misidentification of the ATG start codon would lead to erroneous annotation of sequence variants in the first exon as loss-of-function rather than 5′-UTR variants, and loss-of-function variants in the last exon are less likely to lead to nonsense-mediated decay[30]. Stop-gain mutations in intronless genes had a similar fraction of non-reference alleles as ones in first and last exons.

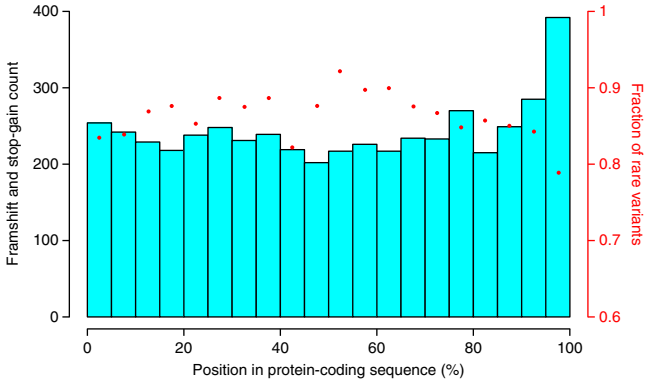

Figure 2 A histogram of the number of frameshift and stop-gain variants by percentage position within the affected protein sequence and the fraction of rare variants within each bin (derived allele frequency (DAF) < 0.5%). For variants affecting multiple transcripts, the mean percentage position was used.

The combination of sequencing a large set of individuals (n = 2,636; median depth of 20×) and imputing genotypes into a much larger set (n = 104,220) has allowed us to find 8,041 complete knockouts for 1 of 1,171 genes by variants with MAF below 2%. In comparison, the previous study of MacArthur *et al.* was effectively restricted to genes completely knocked out (n = 253) by common loss-of-function variants because of the small sample size (n = 185 whole genome–sequenced individuals; mostly sequenced to a depth of 2–4×)[2]. Indeed, only six of the loss-of-function variants released that were seen in the homozygous state had a MAF below 2% in at least one population (see Supplementary Data Set 1 of ref. 2). Identifying individuals with a rare genotypic combination predicted to have high biological impact is particularly vulnerable to erroneous annotation and experimental errors[2]. Through assessing the impact of stop-gain mutations on allele-specific mRNA expression, we have shown the effect of these variants at the transcriptome level, and using Sanger sequencing we have validated the whole-genome sequencing and imputation processes.

The next phase of this work will be to phenotype the complete knockouts, guided by the organ of expression of the knocked-out gene, its biochemical pathways and available phenotype data for animal knockouts[31,32]. This will not only shed light on the nature of diseases

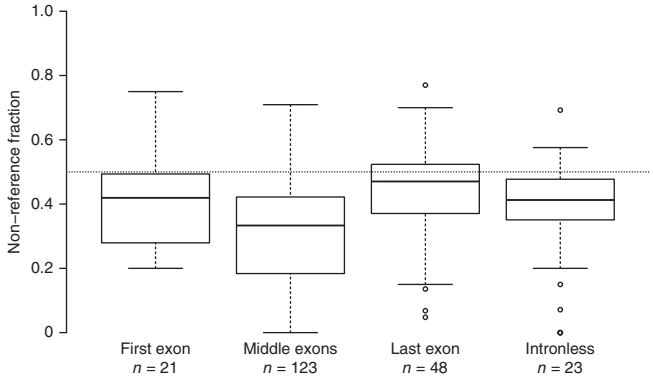

Figure 3 Transcriptome effect of stop-gain SNPs by exon rank. The allele-specific expression of the non-reference (stop-gain) allele was calculated for each variant for a set of 262 individuals with blood RNA sequence data. The top, middle and bottom of the boxes are the top quartile, median and bottom quartile values calculated over the set of variants. The whiskers show the lowest and highest data points within 1.5 times the interquartile range (IQR) from the median. The dots indicate data points more than 1.5 times the IQR from the median. The *n* values given are the number of variants in each class.

but also on hitherto undiscovered aspects of physiological function. Mutations in two genes that encode proteins believed to be critical to the function of the inner ear provide an example of how this could work. We have identified 11 individuals who are homozygous or compound heterozygous for loss-of-function mutations in *LRIG3*, which is considered to be critical to formation of the lateral semicircular canal in mouse[33]. We have also identified six individuals homozygous for a loss-of-function mutation in the *OTOP1* gene that encodes a protein necessary to the formation of the otoliths in mouse[34]. We intend to recruit together complete knockouts for both genes and to phenotype them meticulously, especially for their sense of balance.

URLs. Statistics Iceland (accessed April 2014), http://www.statice.is/Statistics/Population/Overview; dbSNP (Build 137), http://www.ncbi.nlm.nih.gov/SNP/; dbSNP (Build 158), http://www.ncbi.nlm.nih.gov/SNP/; National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (accessed October 2013), http://evs.gs.washington.edu/EVS/; Human Phenotype Ontology (HPO) (accessed March 2014; hpo.annotations.monthly.build_52), http://human-phenotype-ontology.org/; Mouse Genome Database (MGD) at the Mouse Genome Informatics website (data retrieved 1 April 2014), http://www.informatics.jax.org/. Gene annotation was based on the files HMD_HumanPhenotype.rpt and VOC_MammalianPhenotype.rpt (retrieved 1 April 2014) from ftp://ftp.informatics.jax.org/pub/reports.

METHODS

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

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

P.S., H. Helgason, A.O., U.T., G.M., D.F.G. and K.S. designed the experiment. H.S., H. Holm and U.T. collected the samples. Adalbjorg Jonasdottir, Aslaug Jonasdottir, A.S. and O.T.M. performed the sequencing experiments. P.S., H. Helgason, S.A.G., F.Z., E.H., G.T.S., A.K., G.M. and D.F.G. analyzed the data. P.S., H. Helgason, A.H., D.F.G. and K.S. wrote the first draft of the manuscript. All authors contributed to the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

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

1. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
2. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
3. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
4. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
5. Gudbjartsson, D. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* doi:10.1038/ng.3247 (25 March 2015).
6. Lim, E.T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235–242 (2013).
7. Bamshad, M.J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
8. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
9. 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).
10. 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).
11. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
12. Helgason, A. *et al.* The Y-chromosome point mutation rate in humans. *Nat. Genet.* doi:10.1038/ng.3171 (25 March 2015).
13. Lyon, M.F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
14. Rosenstein, B.J. & Cutting, G.R. The diagnosis of cystic fibrosis: a consensus statement. Cystic Fibrosis Foundation Consensus Panel. *J. Pediatr.* **132**, 589–595 (1998).
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–D135 (2012).
16. Robinson, P.N. & Mundlos, S. The human phenotype ontology. *Clin. Genet.* **77**, 525–534 (2010).
17. Saunders, C.J. *et al.* Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* **4**, 154ra135 (2012).
18. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, D966–D974 (2014).
19. Cooper, D.N. & Krawczak, M. Human Gene Mutation Database. *Hum. Genet.* **98**, 629 (1996).
20. Helgason, A. *et al.* Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland. *Am. J. Hum. Genet.* **67**, 697–717 (2000).
21. 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).
22. Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A. & Richardson, J.E. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* **40**, D881–D886 (2012).
23. 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).
24. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
25. Tint, G.S. *et al.* Defective cholesterol biosynthesis associated with the Smith-Lemli-Opitz syndrome. *N. Engl. J. Med.* **330**, 107–113 (1994).
26. Löffler, J., Trojovsky, A., Casati, B., Kroisel, P.M. & Utermann, G. Homozygosity for the W151X stop mutation in the δ7-sterol reductase gene (*DHCR7*) causing a lethal form of Smith-Lemli-Opitz syndrome: retrospective molecular diagnosis. *Am. J. Med. Genet.* **95**, 174–177 (2000).
27. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
28. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7**, e1002144 (2011).
29. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
30. Baker, K.E. & Parker, R. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Curr. Opin. Cell Biol.* **16**, 293–299 (2004).
31. Kettleborough, R.N. *et al.* A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496**, 494–497 (2013).
32. Ayadi, A. *et al.* Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm. Genome* **23**, 600–610 (2012).
33. Abraira, V.E. *et al.* Cross-repressive interactions between Lrig3 and netrin 1 shape the architecture of the inner ear. *Development* **135**, 4091–4099 (2008).
34. Hurle, B. *et al.* Non-syndromic vestibular disorder with otoconial agenesis in tilted/mergulhador mice caused by mutations in otopetrin 1. *Hum. Mol. Genet.* **12**, 777–789 (2003).

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 2** and **3**). Selection for sequencing was not dominated by any disease. We also note that, because of the size of the human genome and its linkage disequilibrium (LD) structure, most of the genome will not be substantially affected by selection for a specific phenotype. That is, selecting patients with a single monogenic disease will only influence the region around the disease-associated gene. Our selection was primarily based on common adult diseases (the mean year of birth was 1950). Coronary artery disease was the most common disease among the sequenced individuals (n = 474, 18%).

In addition, a total of 104,220 Icelanders have been genotyped using SNP chips (Illumina). 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.

Whole-genome sequencing. Sequencing by synthesis (SBS) was performed on GAIIx and/or HiSeq 2000 instruments (Illumina). Paired-end libraries were sequenced with 2 × 101 (HiSeq) or 2 × 120 (GAIIx) cycles of incorporation and imaging using the appropriate TruSeq SBS kits.

Whole-genome SNP and indel calling. Multi-sample variant calling was performed with Genome Analysis Toolkit (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 the 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.

Whole-genome variant quality filtering. The variants identified by GATK were filtered using thresholds on GATK variant call annotations. 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 (https://www.broadinstitute.org/gatk/guide/best-practices). 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) < 1×10^{-7}, and 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 removed 2% of the remaining SNPs but 50% of the remaining indels. It is primarily the SI condition that removed indels, which usually indicates that the failing indel was not called with a high degree of certainty and that its calling could not be resolved in a coherent manner on the basis of 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 6 bp or more in length. Variants called in these regions were ignored in the analysis.

Of the loss-of-function variants we detected in Icelanders, 1,216 SNPs (39.8%) and 693 indels (31.6%) with MAF below 2% were present in either the Exome Sequencing Project (ESP)[3,11] or dbSNP[35] database, in comparison to 201 SNPs (100%) and 165 indels (95.9%) with MAF over 2%. A higher fraction of missense SNPs (57.7%) and in-frame indels (45.2%) with MAF below 2% in Icelanders were present in these databases, and almost all missense SNPs and most in-frame indels (95.5%) with MAF over 2% were present.

Transmission. Sharing between parent-offspring pairs was tested, and pairs with less than 99% haplotype sharing were excluded from the analysis. Individuals of foreign ancestry (those who shared less than 90% of their genome with sequenced individuals) and sequenced individuals were excluded from the analysis. This left us with 26,188 trios where both the parents and an offspring had been chip typed and undergone long-range phasing. We identified regions in the genome that were more than 1 Mb away from an inferred recombination event in the child. For markers within these regions where both parents were imputed to be heterozygous (with each allele being imputed as being the major or minor allele with 99% probability), we counted how many times the minor allele was transmitted to the offspring.

Similarly, for heterozygous transmissions, we identified 35,024 father-offspring and 47,769 mother-offspring pairs of chip-typed and long range–phased individuals and examined markers imputed to be heterozygous in regions more than 1 Mb away from a recombination site in the offspring.

Validation of rare genotypes. First, we performed Sanger sequencing on a set of individuals predicted to be homozygous for rare loss-of-function variants, enriched for indels over SNPs (**Supplementary Table 5**). We observed genotypes for 47 of 49 loss-of-function variants (96%). Of the 140 individuals imputed to be homozygous for 1 of these 47 variants, genotypes were obtained for 134 (96%), 128 of whom were homozygotes (96%). Second, we performed Sanger sequencing on a set of individuals who were the only carriers of a loss-of-function variant among the 2,636 whole genome–sequenced Icelanders (**Supplementary Table 6**). We observed genotypes for 155 of 162 individuals (96%), and 152 of these matched the genotype determined by whole-genome sequencing (98%).

Quantifying allele-specific expression in white blood cells for loss-of-function variants. We estimated allele-specific expression in blood for 262 individuals who had undergone RNA sequencing. All of these individuals had imputed and phased genotypes. Allele-specific expression was estimated by comparing the number of non-reference (stop-gain) alleles to the number of reference alleles seen at the same position for all mRNA-sequenced heterozygous carriers of each variant.

Complete knockouts by human expression pattern. The gene counts for tissue-specific expression in **Table 3** and **Supplementary Table 10** were based on publicly available data that classify the tissue-specific expression of genes across 27 tissues.

Further detailed information on methods is available in the **Supplementary Note**.

35. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).