A systematic survey of loss-of-function variants in human protein-coding genes

Supplementary material

Daniel G. MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K. Pickrell, Stephen B. Montgomery, Cornelis A. Albers, Zhengdong Zhang, Donald F. Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A. DePristo, Eric Banks, Min Hu, Robert E. Handsaker, Jeffrey Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H.A. Barnes, Clara Amid, Denise R. Carvalho-Silva, Alexandra H Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A. Gibbs, Steven A. McCarroll, Emmanouil T. Dermitzakis, Jonathan K. Pritchard, Jeffrey C. Barrett, Jennifer Harrow, Matthew E. Hurles, Mark B. Gerstein, Chris Tyler-Smith

Table of Contents

Analysis of 1000 Genomes Pilot data 3
Indel calling from low-coverage pilot 3
Deep analysis of a single individual genome 4
Read alignment and SNV calling from HiSeq data 4
Indel calling from HiSeq data 4
Multi-nucleotide polymorphism calling from HiSeq data 4
Calling of large deletions in NA12878 5
Identification of candidate LoF variants 5
Experimental genotyping of LoF variants 6
Filtering of candidate LoF variants 7
Read-based filters 7
Annotation filters 8
Inference of ancestral state for SNVs and indels 8
Sequence context filters 9
Analysis of multi-nucleotide polymorphisms 9
Manual reannotation 10
Errors leading to miscalling of LoF variants 11
Classification of LoF variants 13
Results 13
Identification of known and predicted severe disease-causing mutations 14
Allele-specific expression analysis using RNA sequencing data 15
Imputation-based association analysis of LoF variants in complex disease cohorts 16
Comparison with signals of positive selection 17
Analysis of the properties of LoF-containing and LoF-tolerant genes 18
Evolutionary properties 18
Network properties 19
Comparison of gene sets 19
Generation of a predictive model 19
Statistical analysis 20

Supplementary Tables

Table S1. Candidate nonsense SNVs that are in fact components of multi-nucleotide variants (MNVs) with weaker predicted effects on function..... 21
Table S2. Homozygous high-confidence LoF variants in the anonymous European individual NA12878. 22
Table S3. Genes containing 5 or more independent candidate LoF variants..... 23
Table S4. Known LoF variants found in 1000 Genomes samples associated with non-Mendelian phenotypes. 24
Table S5. Known Mendelian disease-causing mutations identified in our high-confidence LoF set. 25
Table S6. Likely disease-causing mutations identified in our high-confidence LoF set..... 27
Table S7. Allele-specific expression of premature stop codon variants, using RNA sequencing data from genotype-confirmed heterozygous individuals. 29
Table S8. Gene Ontology (GO) categories significantly enriched or depleted in LoF-containing genes compared to the genome background. 31
Table S9. Gene Ontology (GO) categories significantly enriched or depleted in homozygous LoF-tolerant genes compared to the genome background..... 34
Table S10. Evidence from frequency spectrum and haplotype-based tests for positive selection on high-confidence LoF variants. 35

Supplementary Figures

Figure S1. Flow-chart indicating the process of filtering candidate LoF SNVs, indels and large deletions. 37
Figure S2. Accurate functional interpretation requires integrating multiple variants on the same haplotype. 38
Figure S3. Putative frameshift indels close to or spanning exon splice sites can be rescued by alternative splice sites. 39
Figure S4. Systematic sequencing error at the site of a reported disease-causing mutation in the *BBS7* gene. 39
Figure S5. Plots showing evidence for the LoF deletions identified in NA12878. 40

Supplementary Figures

Two additional files have been made available at *Science* online:
1215040s1.xlsx is a complete list of all candidate LoF SNVs and indels with additional annotation;
1215040s2.xlsx provides the predicted probability of recessive disease causation P(rec) for all assayable protein-coding genes in the human genome.

Analysis of 1000 Genomes Pilot data

We analyzed data from all three of the 1000 Genomes pilot projects (*2*). Briefly, these data-sets consisted of:

- low-coverage (2-4x) whole-genome sequence data from 179 individuals from four populations;
- high-coverage (>60x) whole-genome sequence data from six individuals from two families; and
- high-coverage targeted sequencing of 8,140 protein-coding exons in 697 individuals.

In all cases a combination of sequencing platforms was used to generate the data, which were then analyzed using an integrated read mapping and variant-calling pipeline to generate genotypes for single nucleotide variations (SNVs), small insertion/deletion variants (indels) and large deletions. These calls are publicly available at the 1000 Genomes FTP site, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/. With the exception of applying additionally stringent filters to small indels in coding regions (see below), we used the final call-sets from the pilot projects to generate our initial catalogue of candidate LoF variants.

Indel calling from low-coverage pilot

We used the program Dindel (*31*) to call insertions and deletions shorter than 50 bp from both the high-coverage trio and low-coverage pilots of the 1000 Genomes project. Dindel performs a probabilistic realignment of all reads mapped to a genomic region to a number of candidate haplotypes. Each candidate haplotype is a sequence of at least 120 bp that represents an alternative to the reference sequence and corresponds to the hypothesis of an indel event and potentially other candidate sequence variants such as SNVs. By assigning prior probabilities to the candidate haplotypes, the posterior probability of a haplotype and consequently an indel being present in the sample can be estimated.

Although the false-discovery rate for the indels estimated from the low-coverage data was estimated to be lower than 5% genome-wide, we found that indels in coding regions appeared to be enriched for false-positives to an even greater extent than SNVs. We therefore applied a more stringent filter for the identification of LoF indels. The stringent filter requires that the range of positions where an indel would yield the same alternative haplotype sequence as the original called indel (for instance, in a repeat, the deletion of any repeat unit would give the same alternative haplotype), plus 4 bases of reference sequence on both sides of this region, was covered by at least one read on the forward strand, and at least one read on the reverse strand, with at most one mismatch between the read and the alternative haplotype sequence resulting from the indel (regardless of base qualities). This filter removed an excess of 1-bp frameshift insertions seen in CHB+JPT with respect to CEU seen in the less stringently filtered genome-wide indel call set, but it is also expected to have removed a significant number of true positive calls as well.

For Dindel calls on the high-coverage trio pilot data we did not use read-based filters. However, calls were made jointly on all three members of each trio, and variants that failed to segregate as expected were removed prior to downstream analysis.

Deep analysis of a single individual genome

To provide a more accurate picture of the LoF variants present in a "typical" genome we performed a systematic analysis of candidate LoF variants in the genome of NA12878, an anonymous female of European ancestry sequenced as part of the 1000 Genomes high-coverage pilot. We used the following sources of data:

- genome-wide sequence data from three separate platforms (Illumina, SOLiD and 454) generated as part of the 1000 Genomes pilot (*2*);
- high-coverage (64X) whole-genome sequence data from a single platform, the Illumina HiSeq 2000 (*6*);
- array hybridization intensity data from two published sources (*23, 32*);
- published fosmid end-sequencing data (*33*).

Read alignment and SNV calling from HiSeq data

The genome of NA12878 was sequenced to a total coverage of ~64X using 100 bp reads from an Illumina HiSeq 2000 instrument. Raw sequence data was mapped using BWA, and GATK was used for quality score recalibration and realignment. SNVs were called from the realigned HiSeq data using GATK. Full details of the mapping and SNV-calling approaches are described in DePristo *et al.* 2011 (*6*).

Indel calling from HiSeq data

Dindel (*31*) was used to call indels from the HiSeq data for NA12878 assuming a diploid model. Given the high coverage, long reads, and high expected quality we applied a similar but more stringent filter as for the LoF indels from the 1000 Genomes low-coverage pilot (see above): we required no mismatches in at least one read on the forward strand and one read on the reverse strand. All indels passing this filter were taken forward for further analysis.

To generate the final set of candidate LoF indels for NA12878, we took the union of Dindel calls from the HiSeq data and calls made from the 1000 Genomes pilot data on the same individual. This union was then subjected to manual inspection and validation as described below.

Multi-nucleotide polymorphism calling from HiSeq data

Most current SNV-calling and functional annotation methods are not explicitly designed to process block substitutions of two or more adjacent bases, known as multi-nucleotide polymorphisms (MNPs). While relatively rare, such polymorphisms can have dramatic effects on coding sequence function. Importantly, incorrectly calling and annotating MNPs as independent SNVs can result in both false LoF variants and in missing true LoF variants.

We took two separate approaches to identify MNPs from the NA12878 HiSeq data. The first approach began with the SNV calls described above, but without applying the standard SNV cluster filter. A Bayesian framework (Fromer and Garimella *et al*., in preparation) was then applied to determine the most probable local haplotype given the sequencing reads. Next, phased SNVs within a genomic distance of 2 base pairs were merged into single multi-nucleotide polymorphism (MNP) records. The

RefSeq codon annotations were then compared between the unmerged SNVs and the corresponding MNP, and putative coding changes that required proper identification of the MNP to be recognized were tabulated. This analysis was performed using publicly available tools written within the Genome Analysis Toolkit (GATK) (*6, 34*). Secondly, we called MNPs using the algorithm FreeBayes (http://bioinformatics.bc.edu/marthlab/FreeBayes) on the same set of pre-aligned BAM files.

These two call-sets were merged, and MNPs functionally annotated relative to Gencode using a modified version of ANNOVAR (*35*). We manually examined six candidate MNPs from the union of the two call sets that were predicted to produce premature stop codons; all of these variants were found to correspond to either mapping or annotation artifacts.

Calling of large deletions in NA12878

To determine all LOF deletions for NA12878 the following data sets were taken into account: previously published deletions using array (*23, 32*) and fosmid end-sequencing (*36*) data, deletions detected in the 1000 Genomes Project (*10*) and *de novo* calling of deletions detected in the HiSeq data using BreakDancer (*37*) and Pindel (*38*). Collating all deletion sets resulted in 69690 candidate LOF deletions. For each of these candidates we searched for any overlap with coding sequences as annotated by the GENCODE project for NCBI36. Of the 8318 deletions with at least 1 bp overlap with any coding region, 5375 were predicted to result in LoF for the affected gene using the definitions described below.

To remove redundancy we then created 3709 regions of interest by clustering deletions which had at least 1 bp overlap with each other. We carefully analyzed each cluster and considered additional information such as GC content, raw read depth using reads with mapping quality zero and greater than zero, and evidence from the log2 ratio generated by CNV discovery experiments (Conrad 2009).

Finally, we manually reassigned breakpoints in each candidate region, using assembly information where available and a combination of read-pair and read depth information in other cases. After breakpoints had been assigned for each deletion, we determined whether the variants would still be predicted to cause LoF in order to arrive at a final LoF candidate set.

Using this approach we detected 31 candidate LoF deletions in NA12878, of which 5 were homozygous and 26 were heterozygous. The genes within these candidate regions were then subjected to manual reannotation as described below to examine the evidence supporting the existence of a functional transcript that would be disrupted by the predicted deletion.

Identification of candidate LoF variants

Functional annotation of SNVs and short indels was performed with reference to the GENCODE v3b annotation release (*7*) using the annotation algorithm VAT (http://vat.gersteinlab.org/). Variants were mapped on to transcripts annotated as “protein_coding” and containing an annotated START codon, and classified as synonymous, missense, nonsense (stop codon-introducing), stop codon-disrupting or splice site-disrupting (canonical splice sites). Transcripts labelled as NMD (predicted to be subject to nonsense-mediated mRNA decay) were not used.

We used a custom algorithm to annotate large deletions as gene-disrupting if they fulfilled the following criteria:

1. Removal of >50% of the coding sequence; or
2. Removal of the gene's transcriptional start site or start codon; or
3. Removal of an odd number of internal splice sites; or
4. Removal of one or more internal coding exons that would be predicted to generate a frameshift.

For large deletions with imprecise breakpoints, we conservatively required that the deletions defined by both the inner and outer confidence intervals would have the same predicted effect on gene function. For cases with microhomology at the break-point we treated the breakpoint as falling to the right-hand side of the region of microhomology.

We did not perform functional annotation for large duplications due to the challenges of inferring functional consequences. We also did not pursue small indels overlapping splice sites, as these were observed to have a relatively high rate of annotation errors. The numbers stated in the text should thus be regarded as a lower bound for the number of observed loss-of-function variants per individual genome.

Experimental genotyping of LoF variants

A variety of approaches were taken to obtain independent experimental validation that candidate LoF variants represented genuinely polymorphic sites.

Firstly, 276 LoF SNVs reported to be polymorphic by the HapMap project were regarded as independently validated. Secondly, we obtained raw intensity data from three separate custom Illumina genotyping arrays (1KG-P12, ImmunoChip and Omni2.5) that had each been run on all or most of the 1000 Genomes low-coverage pilot samples, providing genotype data for 1,135 SNVs. Finally, we designed custom Sequenom assays to genotype 243 SNVs and 537 indels identified by the 1000 Genomes Project that were not included either in the HapMap project or any of the custom arrays described above, and ran these assays on all 185 individuals analyzed by the low-coverage and high-coverage pilots of the Project.

Intensity data from both the Illumina arrays and the Sequenom assays were manually examined using the program Evoker (*39*) to determine whether the variant had support for polymorphism in the assayed samples, and also whether there was evidence for one or more individuals homozygous for the LoF allele. These intensity data were also used to provide accurate genotypes for both SNVs and indels for use in allele-specific expression analysis (see below).

For the NA12878 HiSeq variant calls we were able to use genome-wide 454 data generated as part of the 1000 Genomes high-coverage pilot for orthogonal validation. All predicted LoF SNVs and indels were

manually examined with Integrated Genomics Viewer (*40*) for evidence of polymorphism in the 1000 Genomes 454 data. As Illumina and 454 are subject to different forms of systematic error, we regarded the presence of a single 454 read supporting the same non-reference allele as the HiSeq data as sufficient to regard the variant as validated.

For 34 candidate indel sites in NA12878 there was insufficient 454 coverage to confidently call the genotype at that location. We attempted to design independent assays at all of these sites using PCR followed by capillary sequencing. These regions were nearly all extremely repetitive, so genome-unique PCR primers were only able to be designed for 24/32 variants; of these variants only 17 produced a band of the expected size, and upon sequencing only 8 reactions produced traces corresponding to a single sequence mapping to the expected region. We examined the resulting traces for NA12878 and both of her parents to confirm transmission of the candidate variants.

It is important to note that certain forms of error (particularly mapping error due to large duplicated regions) will have survived validation using some or all of these methods, so there will be some residual false positives within the validated set.

Filtering of candidate LoF variants

Read-based filters

We complemented the experimental genotyping data with two read-based filters designed to capture a variety of common sources of sequencing error in SNVs. Firstly, we compared the distribution of base quality scores between reference and non-reference alleles; a genuine polymorphism would be expected to have similar quality score distributions for the two alleles, whereas significantly different quality score distributions between the two alleles is suggestive of a sequencing error. Secondly, we looked for SNVs where non-reference base calls were found disproportionately in the 5 bp at either end of spanning reads; such cases are strongly indicative of mapping errors due to either repetitive sequence or proximity to an undetected indel. Both filters produce a rank sum *P* value for each variant, which we then compared to our experimental genotype data to determine appropriate thresholds for filtering ungenotyped variants. Both filters were found to be highly predictive of genotyping errors when using our independent genotyping data as a training set.

To lower the number of false positive indel calls in the 1000 Genomes pilot data we applied more stringent filters to the subset of indels that were called in the genome-wide set and were predicted to fall into the LoF class. The stringent filter requires that the range of positions where an indel would yield the same alternative haplotype sequence as the original called indel (for instance, in a repeat, the deletion of any repeat unit would give the same alternative haplotype), plus 4 bases of reference sequence on both sides of this region, was covered by at least one read on the forward strand, and at least one read on the reverse strand, with at most one mismatch between the read and the alternative haplotype sequence resulting from the indel (regardless of base-qualities). This filter removed an excess of 1-bp frameshift insertions seen in CHB+JPT with respect to CEU in the less stringently filtered

genome-wide indel call set, although it is expected to remove a significant number of true positive calls as well. The indels that pass these stringent filters have been annotated in the project's VCF files (*2*).

Annotation filters

Nonsense and splice-disrupting SNVs were flagged as likely representing reference error or annotation artifacts if the inferred loss-of-function (LoF) allele was also the ancestral state (see below), or if the reference (non-LoF) allele was not observed in any individual in that population. Splice-disrupting SNVs in non-canonical splice sites were also discarded, as these were found to frequently correspond to small artificial introns added during the gene annotation process to account for reference sequence errors or indels in the model transcripts relative to the reference.

For nonsense SNVs and frameshift indels, we calculated the fraction of the coding sequence of the longest affected transcript that would be disrupted by the variant. Both classes of variant showed a striking enrichment towards the 3' end of the coding sequence of affected genes (Fig. 1C,D). Putative stop and frameshift variants predicted to disrupt less than 5% of the coding sequence were excluded from further analysis. This analysis was not performed in an automated fashion for splice site SNVs due to the uncertainty in inferring the effects of splice site disruption on final transcript structure. A single frameshift falling within the final 5% of the affected CDS (in the *NOD2* gene) was manually rescued from this filter as it is a known functional variant.

Inference of ancestral state for SNVs and indels

An additional approach to identify LoF variants that are likely to represent gene annotation artifacts and/or reference sequence errors is to examine the inferred ancestral state at the site, using comparison with outgroup species. Variants where the putative LoF allele is confidently inferred to be ancestral should be regarded with suspicion: such cases must either represent an evolutionary "gain of function" mutation (that is, the loss of a stop codon, creation of a novel splice site or extension of a reading frame) that happens to be carried in the reference sequence or, much more commonly, are the result of sequencing errors present in the reference that have resulted in a mis-specification of the gene model at this location.

The ancestral state of both SNVs and indels were inferred using comparison of the human NCBI36 reference with three non-human primate species, the chimpanzee *Pan troglodytes*, orang utan *Pongo pygmaeus abelii* and rhesus macaque *Macaca mulatta*. For SNVs, we relied on ancestral state assignments generated for the 1000 Genomes pilot project (*2*) using alignments of NCBI36 to the CHIMP2.1, PPYG2 and MMUL_1 reference genomes with Enredo and Pecan (*41*).

For indels, ancestral states were freshly calculated for this project, using alignments of NCBI36 with the panTro2, ponAbe2 and rheMac2 reference genomes. An indel was polarized if (i) at least two of the three primate outgroups had aligning sequence present at the variant site in the relevant UCSC BlastZ two-way alignment, (ii) either all aligning sequence showed a matching indel at the variant site, or no aligning sequence showed a matching indel at the variant site, and (iii) no non-matching indel was found at the variant site in any of the primate outgroups.

Provided these criteria were fulfilled, when no aligning sequence showed a matching indel at the variant site, the indel was annotated as "derived". Otherwise, the indel was annotated as "ancestral". Indels were deemed to match if their length and type (deletion or insertion with respect to the reference) matched. To allow for substitutions, the inserted or deleted sequences themselves were not required to be identical. To allow for alignment ambiguity and possible substitutions in any of the primate sequences, the site of any indel was defined to be the segment of possible positions of the gap characters in a consistent alignment of the two sequences, plus 5 bp on either end.

Manual inspection was also used to classify a number of very high frequency indels and SNVs for which ancestral state could not be inferred (typically due to missing non-human primate sequence at these locations).

For both SNVs and indels, cases where the ancestral state could not be reliably inferred were not filtered.

A total of 80 SNVs and 154 indels were found to have an ancestral LoF allele and were thus excluded from further analysis. In support of the notion that these variants frequently represent errors in the reference sequence, there was a striking enrichment of very high-frequency variants in this filtered class relative to other LoF variants (indel mean derived allele frequency 0.64 *vs* 0.09; SNV average DAF 0.42 vs 0.06). This enrichment can be explained partly by sequencing errors in the reference that are consequently called as non-reference in all (or nearly all, depending on genotyping power) of the individuals in the sample.

Sequence context filters

We excluded variants that were present in a segmental duplication, as well as SNVs found within a variable number tandem repeat, based on annotation from the UCSC Genome Browser. Candidate LoF SNVs were also excluded if they were found within 3bp of the location of a known indel (seen either in dbSNP or in the 1000 Genomes pilot calls), as manual inspection of read data indicated that the majority of these represented read mapping artifacts.

Analysis of multi-nucleotide polymorphisms

Multi-nucleotide polymorphisms (MNPs) are events in which variations from the reference genome are present on the same haplotype at multiple adjacent bases. When an MNP affects multiple bases within the same codon it can have substantially different functional effects than its component SNVs annotated individually. We explored the effects of MNPs in two ways: firstly, for all stop codon SNVs identified in this project, we looked for evidence that the variant was actually part of an MNP that would result in a different functional outcome; and secondly, we systematically called and tested the functional effects of all candidate MNPs in the genome of NA12878.

To identify "stop-disrupting" MNPs we looked for additional SNV calls either 1 or 2 bases away from each candidate stop-gain SNV (2 bases being the maximum distance within which an additional variant could still affect the same codon) in any 1000 Genomes pilot sample. For all cases where a neighboring SNV was present we then manually examined the read data in the relevant population to determine if

reads carrying the stop-gain SNVs also always carried the neighboring SNV. Finally, we determined the overall outcome for manually validated MNPs in terms of effects on protein sequence. This process identified 33 cases where an apparent stop-gain SNV was in fact a component of an MNP, all of which resulted in a missense rather than nonsense prediction overall (see example, Fig. S2A). Of the 28 of these SNVs that had been subjected to experimental genotyping there were nine failures, presumably due to interference with genotyping probes by the neighboring SNV.

We manually inspected the evidence for seven autosomal candidate MNPs identified in the HiSeq data from NA12878 and annotated as creating a stop codon in either the Gencode v3b or RefSeq gene sets. Four of these were likely mapping errors: two in MHC genes *HLA-B* and *HLA-DPB1,* one in the artifact-rich gene *CDC27*, and one in *NBPF9*. One MNP in the *IRF2BP1* gene had very weak read evidence in the HiSeq data, had no support from either 454 or GA2 data from NA12878 or her parents, and was excluded as a likely variant-calling error. Finally, two MNPs were annotation artifacts: one was identified in the gene *RAB36*, but in a transcript with an in-frame upstream stop codon; a second, in the predicted gene *AL122127*, was found in a weakly-supported exon flanked by non-canonical splice sites.

Manual reannotation

Full manual annotation was undertaken on loci containing 884 candidate LoF variants. The purpose of this exercise was twofold; firstly to confirm the validity of the annotation of the protein-coding model on which the LoF variants were called (i.e. to reduce the number of false-positive LoF calls made on loci and splice variants for which either the structure of the model or the annotated CDS was incorrect), and secondly to fully characterize the locus in terms of its alternative splicing and functional potential of any splice variants (i.e. to place true positive LoF calls in their context with regard to whether the exon affected by a LoF variant is constitutive or alternatively spliced and to try and predict the effect of the SNV on the functional potential of the locus).

Manual annotation was performed according to the guidelines of the HAVANA (Human And Vertebrate Analysis and Annotation) group; the current set can be accessed at ftp://ftp.sanger.ac.uk/pub/annotation. In summary, the HAVANA group produces annotation of protein-coding genes, pseudogenes, and non-coding transcripts largely based on the alignment of transcriptomic (ESTs and mRNAs) and proteomic data from GenBank and Uniprot. These data were aligned to the individual BAC clones that make up the reference genome sequence using BLAST (*42*), with a subsequent realignment of transcript data by Est2Genome (*43*). Gene models were manually extrapolated from the alignments by annotators using the otterlace annotation interface (*44*). Alignments were navigated using the Blixem alignment viewer (*45*). Visual inspection of the dot-plot output from the Dotter tool [4] was used to resolve any alignment with the genomic sequence that was unclear or absent from Blixem. Short alignments (<15 bases) that cannot be visualized using Dotter were detected using Zmap DNA Search (essentially a pattern matching tool). The construction of exon-intron boundaries required the presence of canonical splice sites and any deviations from this rule were given clear explanatory tags. All non-redundant splicing transcripts at an individual locus were used to build

transcript models, and all splice variants were assigned an individual biotype based on their putative functional potential. Once the correct transcript structure had been ascertained the protein-coding potential of the transcript was determined on the basis of similarity to known protein sequences, the sequences or orthologous and paralogous proteins, the presence of Pfam functional domains (*46*) possible alternative ORFs, the presence of retained intronic sequence and the likely susceptibility of the transcript to nonsense-mediated mRNA decay (NMD) (*47*). The biotype of the locus was derived from the individual biotypes of the splice variants it incorporates.

The geneset created by the GENCODE consortium will ultimately include manually annotated transcripts for all human genes, but this process is not yet complete; hence the GENCODE geneset is currently represented by merge of HAVANA manual annotation and automated Ensembl gene predictions (*48*) to achieve a better coverage of loci and alternative splice variants (including all CCDSs (*49*)). Consequently, checking and reannotation of manually annotated loci and complete annotation of automatically curated loci is advantageous to ensure calling of LoF variants is made on consistent, high quality annotation.

Errors leading to miscalling of LoF variants

Reference errors. Where manual annotation identifies likely errors in the reference human genome sequence in the same position a LoF variant is called, the variant is flagged as a genome sequence error and excluded from subsequent analysis. Where a genome sequence error that affects the annotation of the locus is identified, but it is still possible to annotate gene models with sufficient information to fully interpret the functional impact of the LoF variant, these variants are included in analysis. Putative genome sequence errors are initially identified on the basis of their disruptive effects on CDSs and splice junctions, and subsequently on their lack of transcriptional support, lack of cross-species support i.e. the human sequence is different to all other primate and mammalian genomes, and lack of a high confidence SNV called at the position. All suspected genome sequence errors were reassessed to determine whether SNVs could be confidently called using data from the 1000 genomes project. Those that were found to be true LoF variants were analyzed in this light; those that were not were reported to the Genome Reference Consortium (GRC) for further investigation with the view of correcting the human reference genome where necessary.

Gene annotation errors. Where re-annotation reveals that a locus or splice variant has a misannotated CDS, the annotation was corrected and the new annotation affects the interpretation of a putative LoF variant, the variant is flagged as being unlikely to have any functional effect. At the locus level cases often result from a change of interpretation e.g. the locus is now believed to represent either a pseudogene of a protein-coding parent or functional non-coding gene (lncRNA). At the transcript level, changes are made where an individual transcript variant initially possesses a CDS that does not fulfill the requirements for annotation according to our manual annotation guidelines. Sources of such errors include problems with the alignment of supporting evidence and the quality of the supporting evidence itself.

Predicted incomplete reduction in functional potential. The potential effect on functional potential of all putative LoF variants that were confirmed to affect valid coding gene models were assessed. Nonsense SNVs and small frameshift-inducing indels introduce alternative stop codons into the CDS. Novel stop codons result in either truncations or, rarely, extensions of the reference CDS. Truncations that possess the positional characteristics signaling their targeting by the NMD pathway (*47*) are likely to lead to a significant reduction in the amount/stability of the transcript and suggest the protein it encodes is likely to be non-functional. The functional effect of truncations that do not induce NMD are more difficult to characterize, however, we have used the disruption of a Pfam A domain (*46*) as a second proxy for loss of function. Where a truncation led to the loss of >=1 residue of a Pfam domain it was considered to be disrupted. Where no Pfam A domains were disrupted, or the reference CDS possessed no Pfam A domains, a third criterion was used; by which truncation and extension were characterized by the proportion of the reference sequence lost (or gained) in the variant CDS. Truncations were grouped according to whether they lost >50%, 50%-5%, or <5% of the length of the reference CDS. Intuitively CDSs with larger truncations seem more likely to have lost the function of the reference CDS than those with smaller losses, however, there are many well characterized examples where a small terminal truncation leads to abolition of protein function e.g. olfactory receptors (*50*).

Splice junction SNVs. Variation at both donor and acceptor splice sites affects the complex dynamics of splicing can and potentially lead to loss of function due to erroneous exon skipping or inclusion of non-exonic sequence which can lead to inclusion of a premature stop codon either directly or via a frameshift (see (*5*) for summary). Predicting the consequences of splice site disruption can be difficult, particularly in the case of splice donor site disruption) (*51-54*). As such, all predictions of the effect of splice junction SNVs on the functional potential of a transcript were conservative where no additional evidence for novel splice sites was available. For splice acceptor SNVs, the next confidently identifiable splice acceptor is presumed to be used. Practically, this equates to a prediction that the exon immediately proceeding the affected splice acceptor being skipped unless there is transcriptional support for the use of an alternative downstream splice acceptor within that exon. The impact of splice donor SNVs are more difficult to predict as they can have an effect on the splicing of 5' as well as 3' exons. As such where a splice donor variant was identified it was deemed that any attempt to evaluate its impact on the functional potential of the locus would be unreliable. The one exception to this is where transcriptional evidence possessing the disrupted donor site can be used to support an alternative splice model e.g. where the disrupted splice donor is read through and the transcript either reads through the intron completely or utilizes a cryptic downstream splice donor.

Alternative splicing. Consideration of alternative splicing is of great importance in assessing the functional impact of a potential LoF variant. Where an affected exon, or part of an exon, is alternatively spliced it is very likely that the LoF SNV will only affect the function of those transcripts that contain it. If the transcripts that do not contain the LoF SNV are unaffected it is reasonable to assume that they possess the same functional potential as the same variant with the reference allele. As such, where an affected exon is alternatively spliced the effect on function can only be considered at the level of the transcript rather than the locus, complicating any assessment of its impact. In the simplest example, if

the affected the locus is subject to tissue specific alternative splicing, any loss of function would obviously affect only those cells where the LoF variant containing exon was included. The manual annotation of putative LoF loci ensured that all supported alternative splice variants were built, giving some context to the analysis of functional potential. Some indication as to the proportion of transcripts including/excluding the affected exon/portion of an exon is given in the spreadsheets, however, this information should be taken as indicative at best due to uneven coverage of tissues, conditions and developmental stages in the transcript databases. More transcriptomic data will be required in order to more precisely refine the predictions for those loci flagged.

Classification of LoF variants

In summary, LoF SNVs were grouped as follows according to the impact of the variant on the CDS: severe impact (NMD, Pfam A domain break, >50% truncation), moderate impact (50%-5% truncation), minor impact (<5% truncation), uncertain impact (splice donor SNVs with no transcriptional data, final exon splice acceptor SNVs) and mixed impact (variant had different effects on different transcripts e.g. rescuing NMD variants). Furthermore, all potential LoF variants were classified according to whether they were mapped to a constitutive exon or an exon (or part of an exon) skipped by at least one piece of transcript data.

Results

The genes affected by 884 putative LoF variants were fully reannotated, comprising 296 SNVs and 213 indels from the 1000 Genomes pilot data and 375 variants from NA12878. Variants selected for annotation were not uniformly ascertained, with the pilot SNVs in particular being drawn predominantly from the higher end of the frequency spectrum, so the results below should not be regarded as representative of genome-wide error rates for such variants.

Detailed manual annotation led to the identification of 44 genome sequence errors and 243 errors in gene models, fewer than 5% of which affected previously manually annotated gene models.

Of the 597 putative LoF variants confirmed to affect protein-coding gene models 213 (~36%) were found to be present in exons that were subject to alternative splicing capable of excluding the LoF variant from the final transcript. In total 315 LoF variants at both constitutive and alternatively spliced loci led to changes in the CDS that broke a Pfam A domain, generated a transcript likely to be targeted by the NMD pathway, or truncated the CDS by more than 50%. This category represented the largest set for both constitutive and alternatively spliced variants in every set investigated. Overall, 114 variants led to a truncation of the CDS of between 5 and 50% while 101 variants introduced at truncation or extension to the CDS of less than 5% compared to the reference genome.

In 53 cases the effect of the LoF variant could not be established because the variant represented either a splice donor SNV without specific transcript evidence that confirmed its effect or a splice acceptor SNV at the final exon of the locus where the next reasonable splice acceptor could not be determined due to lack of transcript evidence or uncertainty over candidates. A further 11 SNVs had potentially mixed effects i.e. the presence of the putative LoF variant affected different alternative splice variants in

different way e.g. a splice junction skip could predict the exclusion of an exon, the result of which would be to knock all annotated coding variants at the locus out of frame, leading to inclusion of a premature stop codon and NMD, but the same exon skipping event could shift one or more models predicted to be subject to NMD in the reference genome back into frame, enabling them to encode a full-length CDS.

Identification of known and predicted severe disease-causing mutations

We compared our high-confidence LoF variants to the largest currently available database of known Mendelian disease-causing mutations, the Human Gene Mutation Database (HGMD). HGMD missense, nonsense, regulatory, splice site and coding indel variants tagged as "damaging mutations" and confidently associated with disease were included (entries with a question mark in the "disease" field, representing low-confidence disease associations, were excluded). The final HGMD set consisted of 91,193 variants in 2,375 genes, all obtained by manual curation of the disease literature, and all with positional coordinates available relative to the reference genome.

Firstly, we identified high-confidence LoF variants that were also annotated as disease-causing by HGMD. We found 26 overlapping sites, but manual inspection revealed 3 of these (in the *MST1R*, *HTN3* and *RAGE* genes) to have unconvincing associations with disease phenotypes. In addition, we identified two known disease-causing mutations from our own surveys of the literature that were not present in HGMD: a stop SNV in *PCSK9* associated with low LDL cholesterol levels, and a frameshift insertion in *NOD2* associated with the risk of Crohn's disease. The surviving 26 disease-causing mutations are summarized in Table S5. Only one of these variants was identified in a homozygous state: a frameshift deletion in the *P2RX5* gene associated with graft-versus-host disease in bone marrow transplant patients.

Secondly, we looked for novel candidate disease-causing mutations by inspecting high-confidence LoF variants in genes annotated as carrying disease mutations in HGMD. In total we found 223 LoF variants in HGMD disease genes, but only 89 of these survived filtering (including the 25 LoF sites annotated as disease-causing noted above). We manually inspected the evidence for potential disease causation for the 29 of these variants that were predicted to cause loss-of-function for all known transcripts of the affected gene, investigating both the strength of the predicted effects on gene function and the literature supporting a role for the gene in disease causation. After removing variants in genes with weak evidence of disease causation we were left with 21 LoF variants we regard as strong candidates for novel Mendelian disease-causing mutations (Table S6).

While none of the strong candidate mutations were seen in the homozygous state, two were identified with relatively high allele frequencies in at least one population: in one case the disease in question only manifests itself in response to drug exposure (pravastatin-induced myopathy due to mutations in the *SLCO1B1* gene), while in the second case the disease phenotype is relatively mild (congenital stationary night blindness, associated with mutations in the *TRPM1* gene).

Allele-specific expression analysis using RNA sequencing data

To quantify the effect of putative LoF variants on RNA expression from the affected gene, we used previously published (*24, 25*) lymphoblastoid cell line RNA sequencing data from 60 CEU and 59 YRI individuals also included in the 1000 Genomes low-coverage pilot to examine allele-specific expression (ASE) from the LoF and reference alleles in heterozygous individuals.

This analysis was performed only for stop SNVs: SNVs in splice sites by definition fall outside exonic sequences and are thus not included in transcriptomic data, and analysis of indels was complicated by mapping bias (that is, a strong tendency for reads corresponding to the non-reference allele to not be correctly placed during read mapping) for these polymorphisms. To ensure the genotypes used in this analysis were of high confidence, only stop SNVs that had been genotyped on an independent platform (either one of the custom Illumina arrays or Sequenom assays analyzed in this study, or as part of the HapMap project) were used. For SNVs genotyped on the custom arrays or the Sequenom assays, heterozygous individuals were identified by manual inspection of genotype intensity data using custom software (Pyvoker) provided by T. Shah.

In total, of 598 SNVs that passed all of the validation, annotation and informatic filters described above, we were able to obtain independent genotype data for 388, of which 347 possessed at least one independently validated heterozygous individual. For SNVs where data were available from both chip intensity and HapMap data, we used the manual chip-based calls; where data were available from multiple chips, we favored the chip with the highest number of genotyped individuals.

For each stop variant we predicted whether the variant was likely to trigger nonsense-mediated decay using the rule proposed by Nagy and Maquat (*26*): if a stop variant was found more than 50 bp upstream of the final exon-exon boundary in a transcript it was regarded as an NMD-predicted variant. To define exon boundaries we used the longest transcript for which the variant would be predicted to cause loss-of-function.

The RNA expression of these variants was then assessed. Briefly, RNA sequencing in both of these studies was performed using the Illumina GAII, and reads were mapped to the reference genome using either BWA v.0.5.8 (YRI) or MAQ v.0.6.6 (CEU). For each heterozygous individual, reads mapping to the reference and non-reference allele were counted, and these numbers were then summed across all experimentally validated heterozygotes to give a global read count for each allele. Only variants with a total read count ≥ 5 were included in downstream analysis, as illustrated in Fig. 2B. In this figure (and below), the Wilson score interval method (*55*) was used to estimate the most likely proportion and the 95% confidence interval for each variant.

Two variants had sufficient read counts for analysis in both populations, and showed high consistency between the two samples: alternate read count fractions were 0.423-0.602 in CEU and 0.444-0.620 in YRI for the *CARD8* stop variant rs2043211; and 0.356-0.518 in CEU and 0.428-0.593 in YRI. We thus merged the data from the two populations into a single combined analysis.

Predicted NMD and non-NMD variants showed no significant difference in total read coverage (Mann-Whitney U test, *P* = 0.96), but, as expected, NMD-predicted variants had significantly lower proportions of reads mapping to the LoF allele (mean 27.8% vs 47.5% for NMD-predicted and non-predicted variants respectively, MWU *P* = 0.0023). For each variant we examined the probability of obtaining that proportion of reads under a binomial distribution with a true sampling proportion of 0.5; variants showing a significant *P* value with a one-tailed binomial test were classified as showing evidence for NMD. Using a nominal *P* value of 0.05, 14 variants showed reduced expression of the stop allele: 11/28 NMD-predicted (39.3%) and 3/21 non-predicted variants (14.3%). Using a Bonferroni-corrected *P* value of 0.00102, eight variants showed evidence for reduced expression of the stop allele: 7/28 NMD-predicted (25.0%) and 1/21 non-predicted variants (4.8%).

Given the poor accuracy of the standard NMD prediction method, we explored whether the addition of information about the fraction of the coding sequence truncated by a variant could provide additional predictive information. Contrary to this hypothesis, we found no significant correlation between the position of the stop variant within the coding sequence and the proportion of reads mapping to the alternate allele. However, we note that our power to detect such an association here is small, and that combining genome and RNA sequencing information from larger samples will be required to definitively test this hypothesis.

Imputation-based association analysis of LoF variants in complex disease cohorts

To assess whether LoF variants were enriched for effects on complex disease risk, we imputed all SNVs and indels genotyped in the CEU population in the 1000 Genomes low-coverage pilot into the complete Wellcome Trust Case Control Consortium 1 (WTCCC1) data-set (*22*), comprising 3,004 controls and 13,990 cases from seven complex disease cohorts, of which 2,938 controls and 13,241 cases remained following sample QC.

Genotypes for CEU SNVs and indels were obtained from the July 2010 release (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/), and were merged with SNV genotypes from HapMap3 release 2. Imputation of these variants into the WTCCC1 data-set was performed using impute2 version 2.1.0, using a k value of 80 and an effective population size (N_e) of 14000. The data were split up into segments of either 5Mb or 20K reference SNVs (whichever was smaller), with a 500Kb buffer on either side of each segment.

We investigated potential associations with complex disease risk for 625 high-confidence LoF variants identified as polymorphic in the CEU population. Of these variants, 417 imputed well enough in both controls and at least one cohort to go ahead with association (using an info score threshold of 0.2), resulting in a total of 2901 association tests in the seven disease cohorts. Only 3 variants were close enough to the threshold to be assessed in some cohorts but not others.

We performed a frequentist association analysis using the program SNPTest, version 2.2.0. We used an additive model of risk, and a likelihood score test to account for uncertainty in imputed genotypes. Matched synonymous and missense sets were calculated using allele frequencies in controls, taking random draws without replacement of synonymous and missense variants from the same 1% frequency bin as each LoF variant. In both cases, five random draws were made; the values plotted in Fig. 2B are the median values from the 5 draws.

The major caveat of this analysis is that the systematically low frequencies of LoF variants result in a decrease in imputation accuracy, and a subsequent drop in power to detect association. However, we note that the *NOD2* frameshift indel, with an allele frequency of <3% and an odds ratio of approximately 4, achieved a *P* value of 1.78×10^{-14} for association with Crohn's disease despite having a relatively low info score for imputation (0.25). This suggests that our analysis would have successfully identified other LoF variants with large effects, even where allele frequency and imputation accuracy was relatively low.

There were no significant detectable enrichments of associations for LoF variants compared to missense variants at *P* value thresholds of 10^{-5}, 10^{-4} or 10^{-3} (Fisher's exact *P* values 0.4994, 0.1245 and 0.8034, respectively), suggesting that common LoF variants are not substantially over-represented among complex disease risk variants compared to other functional coding polymorphisms.

In addition to the *NOD2* variant that achieved genome-wide significance, two LoF variants achieved Bonferroni-corrected significance: rs16380, a frameshift indel in *ZNF3* (associated in type 1 diabetes), and a novel frameshift indel at chr1:152018423 in the gene *SLC27A3* (associated in hypertension). We pursued the evidence for association for the *ZNF3* variant using data from a meta-analysis of genome-wide association studies of type 1 diabetes incorporating 7,514 cases and 9,045 controls (*56*). We identified 3 SNVs in strong linkage disequilibrium with rs16380 based on 1000 Genomes pilot data that were also examined in the meta-analysis; these showed only nominal significance in the meta-analysis (*P* = 0.03-0.04), and this association was driven entirely by the samples overlapping with the WTCCC1 analysis: looking only at samples that were not overlapping with WTCCC1, the *P* value was 0.4012. This suggests that the marginally significant association in the WTCCC1 samples is a chance finding rather than a genuine association.

Comparison with signals of positive selection

To test the hypothesis that loss of gene function has played a major role in the recent evolutionary history of modern humans we explored the overlap between our high-confidence LoF variants and signals of positive selection derived from the 1000 Genomes low-coverage pilot. We reasoned that under a model of adaptive gene loss, LoF variants with a high derived allele frequency should frequently be associated with signals of recent positive selection.

The approaches used to identify candidate regions of recent positive selection have been described previously (*2*). Briefly, we applied several allele frequency spectrum-based tests – Tajima's D (*57*), Fay and Wu's H (*58*) and Nielsen's Composite Likelihood Ratio (CLR) (*59*) – to the sequence data generated

by the low-coverage pilot to identify regions showing deviations from the expectations under a neutral model of evolution. Simulations under best-fit demographic models (*60*) for European, East Asian and African populations were used to identify appropriate *P* value cut-offs to identify candidate selected regions, using 10 kb bins. The *P* values from the three tests were calculated based on the distribution of 1,000 neutral simulations in each population, and then the scores for all three tests were combined into a single *P* value using Fisher's method. Finally, we classified a region as a candidate for positive selection if two or more significant bins were seen within a 150 kb interval, a criterion estimated to produce a 2% false discovery rate based on our simulations.

We next looked at the overlap between these candidate regions and our final high-confidence set of LoF variants. As frequency spectrum-based tests have very limited power to detect selected variants at a low frequency, we restricted our analysis to the 36 high-confidence LoF variants with a derived allele frequency greater than 0.5 in at least one population. Of these variants, a total of 11 were found in regions overlapping with signals of selection.

To assess signals of positive selection derived from haplotype-based tests, we first retrieved all autosomal SNPs with known phase from the 1000 Genomes Pilot Project for the CEU, CHBJPT and YRI populations. Using these data we calculated two tests of haplotype homozygosity, XP-EHH (*61*) and iHS (*62*) as previously described using tools provided by J. Pickrell (available at http://hgdp.uchicago.edu). XP-EHH was calculated for all three possible population pairs, while iHS was calculated independently for all three populations. Physical and genetic distances were retrieved from the 1000 Genomes Pilot Project data; cM distances between SNPs were averaged across all three population-specific recombination maps to avoid biasing test calculations towards any given population. Ancestral and derived states for each SNP were determined using the same procedure described in the "Inference of ancestral state for SNVs and indels" section above. Scores for each test and population were normalized to have a mean of 0 and an SD of 1. We considered regions in the 2.5% tail at either end of the genome-wide distributions to show nominal evidence of positive selection.

Analysis of the properties of LoF-containing and LoF-tolerant genes

Here we define "LoF-tolerant" genes as genes for which at least one individual in the 1000 Genomes cohort was homozygous for a high-confidence LoF variant; in other words, genes that can apparently be entirely inactivated without causing a fatal early-onset disease.

We compared the functional and evolutionary properties of 1,035 LoF-containing and 253 LoF-tolerant genes with a set of 858 known recessive disease genes obtained from the OMIM database, as well as with a set of 18,797 protein-coding genes from the Gencode annotation set.

Evolutionary properties

dN/dS data for chimp, macaque and mouse were downloaded from Ensembl. Genomic Evolutionary Rate Profiling (GERP) (*63*) score was downloaded from EBI. Two summed GERP values, one for coding sequence and the other for promoter region, defined as bases within [-100, 100) window centered at

transcription start site, were then calculated for all human protein-coding transcripts according to Ensembl annotations and summarized by gene using the median values. We also calculated the GERP score for conserved non-coding elements (CNCs) obtained from Ensembl within 50 kb of a protein-coding gene; for this calculation, CNCs that overlapped with any protein-coding annotation were excluded. The number and sequence identity of paralogs were downloaded from Ensembl.

Network properties

Two interaction networks were used. One is a binary protein-protein interaction network integrated from a number of sources (*64-67*). The other is a probabilistic gene interaction network (a network of 470,217 links among 16,375 human genes calculated using methods previously described for yeast (*68*) and worm (*69*) and derived from 22 publicly available genomics datasets including DNA microarray data, protein-protein interactions, genetic interactions, literature mining, comparative genomics, and orthologous transfer of gene-gene functional associations from fly, worm, and yeast where the weight of a link is the log likelihood score of the interaction (*68*). Measures of centrality (degree, betweenness) and modularity (cluster coefficient) were calculated using MCL (*70*). Shortest path distance and sum of weight of interactions (*69*) were calculated as measures of proximity to a group of 'seed' genes. We note that the inclusion of both human and non-human data in the interaction data may have introduced some non-conservative bias in the comparison between LoF-tolerant and known recessive disease genes in the event that there is unequal conservation of orthologues between these two categories, so the network results should be treated with some caution. However, this caveat does not affect the interpretation of the results of the predictive model described below.

Comparison of gene sets

For continuous variables, the two-tailed Mann-Whitney U test was performed to assess if positive (haploinsufficient) and negative (haplosufficient) training data have the same median value for potential predictor variables. For two-class categorical features, Fisher's exact tests were performed. Statistical tests were performed using R (http://www.r-project.org).

Generation of a predictive model

We assessed different potential sets of predictor variables for input into the predictive model using the following criteria: (i) they allowed prediction for at least half the genes in the genome, (ii) the Spearman correlation between all pairs of predictor variables was less than 0.3, (iii) they were drawn from different broad categories (genomic, evolutionary, functional and network), and iv) they achieved best performance in model assessment (see below).

The sensitivity of the prediction was plotted against (1 - specificity) and the area under the ROC curve (AUC) [44] was used as quantitative measure of the performance of the model, where sensitivity = $TP/(TP+FN)$, and specificity = $TN/(TN+FP)$. The other measure used is the Matthews correlation coefficients (MCC) [45], defined as:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

To avoid over-fitting, the sensitivity and specificity were calculated using 10-fold cross-validation. To overcome the variability caused by random partition involved in 10-fold cross-validation, each such assessment was repeated 30 times and the mean values were reported.

We tested the model both including and excluding olfactory receptor genes from the LoF-tolerant set; both results are shown in Fig. 3C.

Statistical analysis

Except where otherwise specified, we used the Mann-Whitney-Wilcoxon test implemented in R (wilcox.test) for all comparisons between continuous variables. To assess the effect of ties in the analysis of network connectivity data we also used the wilcox.exact test from the R package exactRankTests; none of the associations that were significant in our initial analysis lost their significance with the use of the exact test.

Table S1. **Candidate nonsense SNVs that are in fact components of multi-nucleotide variants (MNVs) with weaker predicted effects on function.**

Bases altered by the MNVs are indicated in upper case in the "actual codon change" field. In all cases the overall result of the MNV was a missense substitution. Note that in some cases the MNV is interrupted by a single unaltered base.

chr	pos	LoF reference allele	LoF non-ref allele	actual codon change	actual protein change
1	45845948	C	T	CAg>TGg	Gln>Trp
1	159742828	C	T	CAg>TGg	Gln>Trp
1	171793211	C	T	TGg>CAg	Trp>Gln
1	226536526	A	T	AGa>TTa	Arg>Leu
3	195543601	C	T	TGg>CAg	Trp>Gln
5	41097472	C	T	TGg>CAg	Trp>Gln
6	71345909	C	T	CAg>TGg	Gln>Trp
6	111694003	T	A	tTA>tAT	Leu>Tyr
7	21549488	G	T	GAg>TTg	Glu>Leu
7	100402871	G	A	TgG>AgA	Trp>Arg
8	599879	A	C	tAT>tGG	Tyr>Trp
8	144593530	G	T	TCa>AAa	Ser>Lys
9	133375256	C	T	CAg>TGg	Gln>Trp
11	5963847	C	T	TGg>CAg	Trp>Gln
11	122437120	A	C	TaT>CaG	Tyr>Gln
14	44044862	G	T	TCa>GAa	Tyr>Glu
16	82541393	C	T	TGg>CAg	Trp>Gln
19	63065900	A	T	AGa>TTa	Arg>Leu
22	17292677	C	T	TGg>CAg	Trp>Gln
1	40545737	G	A	CAg>TGg	Gln>Trp
2	26553793	C	A	GAg>TTg	Glu>Leu
2	220127584	G	A	CAg>TGg	Gln>Trp
2	241204534	G	A	tGG>tCA	Trp>Ser
4	114494796	C	G	TCa>AGa	Ser>Arg
6	30019219	T	A	TTg>CAg	Leu>Gln
6	155619409	T	A	TTg>CAg	Leu>Gln
9	37767620	C	A	tAC>tTA	Tyr>Leu
11	59237528	G	A	CAg>TGg	Gln>Trp
14	62827368	T	A	TTg>AAg	Leu>Lys
14	63629845	G	A	TGg>CAg	Trp>Gln
19	13861171	G	A	CAg>TGg	Gln>Trp

Table S2. Homozygous high-confidence LoF variants in the anonymous European individual NA12878.

variant type	gene	notes
stop	*FUT2*	known nonsense variant, associated with protection against viral infection
stop	*ACTN3*	known nonsense variant in muscle gene, associated with reduced strength and sprint performance
splice	*HTR3B*	type 3 serotonin receptor subunit
frameshift	*CYP4B1*	known null variant in gene involved in inflammation and xenobiotic metabolism
frameshift	*SMPDL3B*	sphingomyelin phosphodiesterase
frameshift	*TIGD6*	tigger transposable element-derived protein
frameshift	*MS4A14*	likely membrane protein of unknown function
frameshift	*CELA1*	chymotrypsin-like elastase, only reported to be expressed in skin keratinocytes
frameshift	*P2RX5*	associated with graft-versus-host disease in bone marrow transplant recipient
frameshift	*ZNF681*	zinc finger protein of unknown function
frameshift, stop	*OR2T4, OR11G2, OR5K4, OR2L8, OR4X1*	olfactory receptors
large deletion	*LCE1D*	known to create fusion gene with *LCE1E*
large deletion	*TUBA3E*	alpha-tubulin protein primarily expressed in testis
large deletion	*SPINK14*	possible serine protease inhibitor
large deletion	*KRTAP9-6, KRTAP9-7*	may create fusion gene between two single-exon genes encoding keratin-associated proteins

Table S3. Genes containing 5 or more independent candidate LoF variants.
Numbers are the sum of all independent candidate LoF variants seen across the three 1000 Genomes pilot projects and in the high-depth NA12878 genome. In most cases the observed LoF variants are a consequence of large-scale read-mapping errors.

gene name	LoF variants before filtering	LoF variants after filtering
AC131157.4	16	0
SSPO	15	0
AC092143.1	12	0
MAN1B1	12	0
CDC27	10	0
MUC19	10	0
AC009063.1	7	0
AC073957.1	7	0
C17orf57	7	0
C11orf40	6	0
OR4C5	6	2
ABCA10	5	0
AC009113.1	5	0
AC010634.1	5	0
AC073995.1	5	0
AC091435.2	5	0
C6orf10	5	0
PKD1L3	5	3

Table S4. Known LoF variants found in 1000 Genomes samples associated with non-Mendelian phenotypes.
All coordinates are relative to the GRCh36 reference build.

chr	pos	dbSNP	ref	alt	type	gene	phenotype	hets	homs	notes
1	25464555-25534879[a]	.	+	-	large del	*RHD*	Rhesus negative blood group	38 all	10 all	~70 kb deletion
1	110024361-110046935[a]	.	+	-	large del	*GSTM1*	loss of enzyme activity; many reported trait associations	66 all	73 all	~22 kb deletion
1	110024361-110046935[a]	.	+	-	large del	*LCE3B, LCE3C*	susceptibility to psoriasis	79 all	44 all	~32 kb deletion
2	162832842	rs35732034	C	T	splice	*IFIH1*	protection from type 1 diabetes	2 CEU	0	
2	162844751	rs35337543	C	G	splice	*IFIH1*	protection from type 1 diabetes	1 CEU	0	
3	38323747	rs753331	A	C	splice	*SLC22A14*	confirmed effects on mRNA splicing	22 CHB+JPT	6 CHB+JPT	
4	69076626-69093238	.	+	-	large del	*UGT2B17*	altered metabolism of testosterone	58 all	53 all	~16 kb deletion
4	70933511	rs17147990	T	A	stop	*HTN3*	truncated histatin protein	19 YRI	2 YRI	
4	154844848	rs62323857	C	T	stop	*TLR2*	decrease in TLR2 protein function	1 JPT	0	
10	96530400	rs4986893	G	A	stop	*CYP2C19*	altered drug metabolism	5 JPT	1 JPT	
11	66084671	rs1815739	C	T	stop	*ACTN3*	altered muscle function	75 all	31 all	LoF in reference
17	36804625	.	C	T	splice	*KRT31*	truncated (functional) keratin	2 CEU	0	
19	53898835	rs1799761	AC	A	del	*FUT2*	non-secretion of ABO/Lewis antigens	2 YRI	0	
19	56226942	rs3745540	A	G	splice	*KLK12*	loss of protease activity	61 all	63 all	

[a] Approximate coordinates provided for large deletions

Table S5. Known Mendelian disease-causing mutations identified in our high-confidence LoF set.

Disease-causing mutations identified using the Human Gene Mutation Database and additional literature searches. All coordinates are relative to the GRCh36 reference build.

chr	pos	dbSNP	ref	alt	type	gene	disease	hets	homs	notes
1	42997627	.	G	A	stop	*LEPRE1*	osteogenesis imperfecta	1 CEU	0	
1	55284810	.	C	G	stop	*PCSK9*	low LDL cholesterol	2 YRI	0	
1	150551510	.	G	A	stop	*FLG*	atopic dermatitis	2 YRI	0	Chinese proband
1	195657157	.	C	T	stop	*CRB1*	Leber congenital amaurosis	1 CHD[b]	0	Korean proband
2	166556297	.	G	A	stop	*SCN1A*	Myoclonic epilepsy of infancy	1 CEU	0	
2	215560703	.	G	A	stop	*ABCA12*	harlequin ichthyosis	1 CEU	0	
3	33146467	.	C	T	stop	*CRTAP*	osteogenesis imperfecta	1 LWK[b]	0	African-American proband
5	39377971	rs34000044	G	T	stop	*C9*	complement C9 deficiency	1 CEU	0	
5	41185792	.	A	G	splice	*C6*	complement C6 deficiency, partial	1 CEU	0	
5	41194621	rs61469168	TC	T	del	*C6*	complement C6 deficiency	2 YRI	0	
6	161006077	rs41272114	C	T	splice	*LPA*	Lp(a) deficiency	7 CEU	0	
8	94867689	.	C	T	stop	*TMEM67*	Meckel-Gruber syndrome	1 JPT[b]	0	
9	138689082	.	T	C	splice	*AGPAT2*	Berardinelli-Seip lipodystrophy	2 YRI	0	
11	6594961	rs56144125	C	T	splice	*TPP1*	Neuronal ceroid lipofuscinosis, late infantile	2 CEU	0	father/daughter in CEU trio
11	7017553	.	A	T	stop	*NLRP14*	spermatogenic failure	1 CHB, 3 JPT	0	
11	64275394	.	G	A	stop	*PYGM*	McArdle's disease	1 CHD[b]	0	
11	64283799	.	G	A	stop	*PYGM*	McArdle's disease	1 CEU	0	

chr	pos	dbSNP	ref	alt	type	gene	disease	hets	homs	notes
11	73392519	rs45476292	C	T	splice	*UCP3*	severe obesity with diabetes	10 YRI	0	weak evidence
11	73394537	.	G	A	stop	*UCP3*	severe obesity with diabetes	1 YRI	0	weak evidence
11	76545663	rs4129813	C	T	stop	*MYO7A*	Usher syndrome 1b	1 CEU	0	
12	42453038	.	C	T	stop	*IRAK4*	predisposition to childhood bacterial infections	1 YRI[b]	0	
15	70427442	.	C	T	splice	*HEXA*	Tay-Sachs disease	2 CEU	0	
16	163705-167490[a]	.	+	-	large del	*HBA1*	alpha thalassaemia	3 YRI	0	~3.8 kb deletion removing last three exons
16	49321279	rs2066847	G	GC	ins	*NOD2*	Crohn's disease	2 CEU	0	
17	3541025	rs5818907	TG	T	del	*P2RX5*	graft-versus-host disease in bone marrow transplant recipient	69 all	66 all	
21	18607163	.	G	C	stop	*PRSS7*	enteropeptidase deficiency	1 CEU	0	

[a] Approximate coordinates; [b] Mutation found in exon capture pilot samples only

Table S6. Likely disease-causing mutations identified in our high-confidence LoF set.
Table shows all high-confidence LoF variants predicted to affect all known transcripts of a known recessive disease-causing gene. All coordinates are relative to the GRCh36 reference build.

chr	pos	dbSNP	ref	alt	type	gene	disease	hets	homs	notes
1	150554463	.	C	A	stop	*FLG*	ichthyosis vulgaris	1 CEU	0	
1	152514361	rs41313932	G	A	splice	*HAX1*	severe congenital neutropenia	1 CEU	0	
1	158519357	.	C	T	splice	*PEX19*	peroxisomal biogenesis disorder	1 CEU	0	
2	31443356	.	A	AT	ins	*XDH*	xanthinuria, type 1	1 YRI	0	
2	71758199-71760693[a]	.	+	-	large del	*DYSF*	adult-onset limb-girdle muscular dystrophy	1 CEU	0	2.5 kb deletion removing exon 52
3	112825290	.	G	A	splice	*CD96*	Opitz trigonocephaly	1 YRI	0	
4	25287192	.	C	A	stop	*SLC34A2*	pulmonary alveolar microlithiasis	1 CHD[b]	0	relatively benign
8	134177729	.	G	A	stop	*TG*	goitre and hypothyroidism	1 YRI[b]	0	
10	100370456	.	C	T	splice	*HPSE2*	urofacial syndrome	1 CEU	0	
11	95235137	.	G	A	stop	*MTMR2*	Charcot-Marie-Tooth type 4B	1 CHD[b]	0	
12	21221099	.	G	T	splice	*SLCO1B1*	pravastatin-induced myopathy	6 YRI	0	environmentally-induced disorder
12	38990629	.	C	T	stop	*LRRK2*	Parkinson disease	1 YRI	0	
12	52105535	.	C	T	stop	*AMHR2*	persistent Müllerian duct syndrome	1 LWK[b]	0	
12	94905063	rs34457757	G	A	stop	*HAL*	histidinemia	2 YRI	0	relatively benign

chr	pos	dbSNP	ref	alt	type	gene	disease	hets	homs	notes
12	94912841	.	C	T	splice	*HAL*	histidinemia	1 CEU	0	relatively benign
12	122805039	.	T	C	splice	*ATP6V0A2*	cutis laxa type 2	1 JPT	0	
15	29082006	rs3784589	C	A	stop	*TRPM1*	stationary night blindness	8 CEU, 4 YRI	0	relatively benign
15	87648098	.	G	A	splice	*FANCI*	Fanconi anaemia	1 CEU	0	
16	1352535	.	A	AG	ins	*GNPTG*	mucolipidosis III	1 CHB	0	
17	36276396	.	A	G	splice	*KRT12*	Meesmann corneal dystrophy	1 CEU	0	relatively benign
18	27247215	.	C	T	stop	*DSG4*	disorders of hair structure	1 CEU	0	relatively benign

[a] Approximate coordinates; [b] Mutation found in exon capture pilot samples only

Table S7. Allele-specific expression of premature stop codon variants, using RNA sequencing data from genotype-confirmed heterozygous individuals.
P value calculated using a one-tailed binomial test of deviation from a proportion of 0.5. All coordinates are relative to the GRCh36 reference build.

chr	pos	ref	alt	dbSNP	gene	NMD	fraction of CDS disrupted	reference reads	non-ref reads	fraction non-ref reads	*P* value	population
18	648001	G	A	.	*C18orf56*	0	0.331	9	15	0.625	0.920	YRI
10	1055710	C	T	rs1044261	*IDI2*	0	0.367	4	6	0.600	0.830	CEU
11	5733060	A	T	rs4910844	*OR52N4*	0	0.466	3	2	0.400	0.500	YRI
12	8180728	G	A	.	*CLEC4A*	0	0.257	10	8	0.444	0.410	YRI
19	21505299	C	T	.	*ZNF429*	0	0.902	1	4	0.800	0.970	YRI
4	38452499	G	A	rs62617795	*TLR10*	0	0.545	21	16	0.432	0.260	YRI
17	39609781	C	T	rs7224330	*C17orf65*	0	0.368	15	9	0.375	0.150	YRI
19	42002191	G	A	rs1227794	*ZNF790*	0	0.531	3	4	0.571	0.770	YRI
20	43944664	G	A	rs35972756	*ZSWIM1*	0	0.982	4	3	0.429	0.500	CEU
6	57025866	C	T	rs61748913	*KIAA1586*	0	0.742	6	6	0.500	0.610	YRI
19	57808623	C	T	rs17855778	*ZNF83*	0	0.349	19	10	0.345	0.068	YRI
16	80591311	G	A	rs11542462	*SDR42E1*	0	0.925	3	5	0.625	0.860	CEU
16	88638451	C	T	rs1048149	*AC133919.7*	0	0.867	28	26	0.481	0.450	YRI
3	1.21E+08	C	T	.	*ADPRH*	0	0.176	15	6	0.286	0.039	CEU
7	1.5E+08	C	T	.	*C7orf29*	0	0.880	19	8	0.296	0.026	YRI
22	40666118	G	A	rs5758511	*CENPM*	0	0.960	56	20	0.263	2.2×10^{-5}	CEU
7	64076102	G	A	rs1404453	*ZNF117*	0	0.115	4	8	0.667	0.930	YRI
15	72115043	C	T	.	*PML*	0	0.120	9	10	0.526	0.680	YRI
17	77664739	G	A	rs11653662	*CCDC57*	0	0.215	4	1	0.200	0.190	CEU
16	87248714	C	T	.	*MVD*	0	0.075	4	5	0.556	0.750	YRI
12	1.19E+08	G	C	.	*COQ5*	0	0.355	5	6	0.545	0.730	YRI
11	2278405	C	T	.	*C11orf21*	1	0.614	3	3	0.500	0.660	YRI
19	4279755	G	C	.	*STAP2*	1	0.624	24	0	0.000	6.0×10^{-8}	CEU
19	17698952	C	T	.	*MAP1S*	1	0.446	5	3	0.375	0.360	YRI

chr	pos	ref	alt	dbSNP	gene	NMD	fraction of CDS disrupted	reference reads	non-ref reads	fraction non-ref reads	*P* value	population
16	20699889	G	A	rs52817836	*ACSM3*	1	0.502	11	0	0.000	4.9×10^{-4}	YRI
14	23749768	G	A	.	*CHMP4A*	1	0.327	31	0	0.000	4.7×10^{-10}	CEU
6	26077610	C	T	.	*TRIM38*	1	0.649	6	0	0.000	0.016	CEU
1	45735570	G	A	.	*RP11-291L19.1*	1	0.425	6	1	0.143	0.063	YRI
3	51980678	G	A	.	*ABHD14B*	1	0.922	8	1	0.111	0.020	YRI
12	52863985	G	A	rs2233919	*SMUG1*	1	0.991	2	4	0.667	0.890	YRI
19	54861076	C	T	.	*BCL2L12*	1	0.816	14	13	0.481	0.500	YRI
17	77955236	G	A	.	*C17orf101*	1	0.431	5	3	0.375	0.360	YRI
4	1.14E+08	G	A	.	*ALPK1*	1	0.522	5	1	0.167	0.110	YRI
11	1.14E+08	G	C	.	*C11orf71*	1	0.871	20	12	0.375	0.110	YRI
4	1.3E+08	C	T	rs10009430	*AC093826.1*	1	0.280	11	3	0.214	0.029	YRI
1	1.57E+08	G	A	.	*MNDA*	1	0.126	17	0	0.000	7.6×10^{-6}	YRI
5	1.69E+08	C	T	.	*CCDC99*	1	0.194	23	0	0.000	1.2×10^{-7}	CEU
11	6548630	C	T	.	*DNHD1*	1	0.066	3	3	0.500	0.660	YRI
22	25192041	G	A	rs3747129	*HPS4*	1	0.667	4	1	0.200	0.190	CEU
6	31232828	C	T	rs3130453	*CCHCR1*	1	0.910	66	39	0.371	0.005	YRI
21	43196789	C	T	rs4148974	*NDUFV3*	1	0.579	17	1	0.056	7.2×10^{-5}	CEU
19	53429518	A	T	rs2043211	*CARD8*	1	0.977	113	124	0.523	0.780	CEU+YRI
15	66284651	G	A	rs11071990	*CALML4*	1	0.799	12	6	0.333	0.120	YRI
17	71589392	C	T	rs1043149	*ZACN*	1	0.320	146	131	0.473	0.200	CEU+YRI
6	74076059	G	A	rs16883571	*KHDC1*	1	0.859	3	2	0.400	0.500	CEU
16	88172869	C	G	.	*CPNE7*	1	0.757	6	5	0.455	0.500	YRI
9	1.14E+08	C	T	rs3780513	*SUSD1*	1	0.958	8	9	0.529	0.690	YRI
12	1.2E+08	G	C	.	*ANAPC5*	1	0.956	25	25	0.500	0.560	CEU
1	2.35E+08	T	A	rs2273865	*LGALS8*	1	0.410	93	3	0.031	2.2×10^{-16}	YRI

Table S8. **Gene Ontology (GO) categories significantly enriched or depleted in LoF-containing genes compared to the genome background.**
Corrected *P* values were generated by Bonferroni correction for multiple tests. * Indicates whether category was still Bonferroni-corrected significant when analysis was repeated excluding olfactory receptor genes.

GO category	number LoF genes	number genome-wide genes	raw P	corrected P	direction	significant without ORs*	type	category description
GO:0007606	84/669	414/13911	5.35×10^{-26}	9.62×10^{-23}	enrichment	no	BP	sensory perception of chemical stimulus
GO:0005515	245/731	7501/14883	2.80×10^{-19}	6.61×10^{-17}	depletion	yes	MF	protein binding
GO:0004930	114/731	968/14883	1.10×10^{-16}	2.59×10^{-14}	enrichment	no	MF	G-protein coupled receptor activity
GO:0004888	144/731	1400/14883	2.19×10^{-16}	5.12×10^{-14}	enrichment	no	MF	transmembrane receptor activity
GO:0007600	99/669	779/13911	4.53×10^{-17}	8.14×10^{-14}	enrichment	no	BP	sensory perception
GO:0043231	267/721	7683/15296	4.13×10^{-12}	5.82×10^{-10}	depletion	yes	CC	intracellular membrane-bounded organelle
GO:0004872	163/731	1900/14883	5.34×10^{-12}	1.24×10^{-09}	enrichment	no	MF	receptor activity
GO:0007186	122/669	1307/13911	8.86×10^{-12}	1.59×10^{-08}	enrichment	no	BP	G-protein coupled receptor protein signaling pathway
GO:0044249	131/669	4171/13911	3.05×10^{-9}	5.48×10^{-6}	depletion	no	BP	cellular biosynthetic process
GO:0016021	309/721	5047/15296	7.18×10^{-8}	1.01×10^{-5}	enrichment	no	CC	integral to membrane
GO:0001653	48/731	389/14883	4.76×10^{-8}	1.10×10^{-5}	enrichment	no	MF	peptide receptor activity
GO:0008528	48/731	389/14883	4.76×10^{-8}	1.10×10^{-5}	enrichment	no	MF	peptide receptor activity, G-protein coupled
GO:0031224	314/721	5146/15296	8.14×10^{-8}	1.13×10^{-5}	enrichment	no	CC	intrinsic to membrane
GO:0005886	215/721	3301/15296	4.75×10^{-7}	6.55×10^{-5}	enrichment	no	CC	plasma membrane

GO category	number LoF genes	number genome-wide genes	raw P	corrected P	direction	significant without ORs*	type	category description
GO:0071944	218/721	3365/15296	5.67×10^{-7}	7.77×10^{-5}	enrichment	no	CC	cell periphery
GO:0007166	151/669	2070/13911	2.66×10^{-7}	4.77×10^{-4}	enrichment	no	BP	cell surface receptor linked signaling pathway
GO:0009891	5/669	568/13911	3.87×10^{-7}	6.94×10^{-4}	depletion	yes	BP	positive regulation of biosynthetic process
GO:0016020	391/721	6962/15296	5.13×10^{-6}	6.98×10^{-4}	enrichment	no	CC	membrane
GO:0031328	5/669	560/13911	5.53×10^{-7}	9.92×10^{-4}	depletion	yes	BP	positive regulation of cellular biosynthetic process
GO:0006366	11/669	780/13911	7.10×10^{-7}	1.27×10^{-3}	depletion	yes	BP	transcription from RNA polymerase II promoter
GO:0045941	3/669	468/13911	8.53×10^{-7}	1.53×10^{-3}	depletion	yes	BP	positive regulation of transcription
GO:0044446	138/721	3986/15296	2.27×10^{-5}	3.06×10^{-3}	depletion	no	CC	intracellular organelle part
GO:0005654	12/721	702/15296	4.21×10^{-5}	5.64×10^{-3}	depletion	yes	CC	nucleoplasm
GO:0048856	67/669	2296/13911	3.51×10^{-6}	6.28×10^{-3}	depletion	no	BP	anatomical structure development
GO:0009893	11/669	737/13911	3.57×10^{-6}	6.39×10^{-3}	depletion	no	BP	positive regulation of metabolic process
GO:0044444	154/721	4305/15296	5.23×10^{-5}	6.96×10^{-3}	depletion	no	CC	cytoplasmic part
GO:0044428	35/721	1342/15296	1.02×10^{-4}	0.014	depletion	no	CC	nuclear part
GO:0003676	111/731	3158/14883	6.01×10^{-5}	0.014	depletion	no	MF	nucleic acid binding
GO:0051254	3/669	403/13911	1.06×10^{-5}	0.019	depletion	no	BP	positive regulation of RNA metabolic process

GO category	number LoF genes	number genome-wide genes	raw P	corrected P	direction	significant without ORs*	type	category description
GO:0045893	3/669	402/13911	1.06×10^{-5}	0.019	depletion	no	BP	positive regulation of transcription, DNA-dependent
GO:0010604	11/669	695/13911	1.19×10^{-5}	0.021	depletion	no	BP	positive regulation of macromolecule metabolic process
GO:0044451	5/721	419/15296	1.91×10^{-4}	0.025	depletion	no	CC	nucleoplasm part
GO:0016563	3/731	334/14883	1.28×10^{-4}	0.030	depletion	no	MF	transcription activator activity
GO:0048522	38/669	1470/13911	1.65×10^{-5}	0.030	depletion	no	BP	positive regulation of cellular process
GO:0003723	15/731	718/14883	2.14×10^{-4}	0.049	depletion	no	MF	RNA binding

Table S9. **Gene Ontology (GO) categories significantly enriched or depleted in homozygous LoF-tolerant genes compared to the genome background.** Corrected *P* values were generated by Bonferroni correction for multiple tests.

GO category	number LoF-tolerant genes	number genome-wide genes	raw P	corrected P	direction	category description
GO:0007606	39/179	452/14496	3.54×10^{-21}	6.61×10^{-18}	enrichment	sensory perception of chemical stimulus
GO:0007600	40/179	840/14496	2.61×10^{-13}	4.87×10^{-10}	enrichment	sensory perception
GO:0007186	45/179	1361/14496	9.95×10^{-10}	1.86×10^{-6}	enrichment	G-protein coupled receptor protein signaling pathway
GO:0044249	19/179	4314/14496	1.51×10^{-9}	2.82×10^{-6}	depletion	cellular biosynthetic process
GO:0007165	62/179	2851/14496	2.95×10^{-6}	5.50×10^{-3}	enrichment	signal transduction
GO:0048856	9/179	2432/14496	3.84×10^{-6}	7.15×10^{-3}	depletion	anatomical structure development
GO:0007166	50/179	2165/14496	8.06×10^{-6}	1.50×10^{-2}	enrichment	cell surface receptor linked signaling pathway
GO:0009653	1/179	1104/14496	2.20×10^{-5}	4.09×10^{-2}	depletion	anatomical structure morphogenesis

Table S10. Evidence from frequency spectrum and haplotype-based tests for positive selection on high-confidence LoF variants.
All coordinates relative to the GRCh36 reference build. XP-EHH comparisons between pairs of populations are indicated in the last two columns (note different pair combinations are shown for the three different population groups). Note that the same allele can be shown multiple times if it is significant in multiple populations. For iHS and XP-EHH values, * indicates a value in the extreme 5% of the genome-wide distribution and ** indicates a value in the extreme 1%. iHS and XP-EHH values are not available (NA) for frameshift indels. NS, not significant.

chr	pos	ref	alt	type	gene	derived allele frequency			peak spectrum comb. *P*	|iHS|	XP-EHH	XP-EHH
						CEU	CHBJPT	YRI				
Variants with significant evidence for selection in CEU											**CEU-YRI**	**CEU-CHBJPT**
1	169379114	C	T	nonsense	*FMO6P*	0.51	0.61	0.81	NS	3.258**	0.909	1.481
5	131352149	C	CTG	frameshift	*ACSL6*	0.77	0.44	0.00	6.20×10^{-4}	NA	NA	NA
9	124431062	G	A	nonsense	*OR1B1*	0.31	0.51	0.03	2.33×10^{-8}	0.499	-0.949	0.305
9	124431591	C	CA	frameshift	*OR1B1*	0.53	0.43	0.40	2.33×10^{-8}	NA	NA	NA
11	4747449	CG	C	frameshift	*OR51F1*	0.17	0.00	0.54	6.43×10^{-7}	NA	NA	NA
11	5400712	C	T	nonsense	*OR51Q1*	0.36	0.61	0.25	3.25×10^{-4}	1.679	0.340	0.580
11	60021578	C	T	nonsense	*MS4A12*	0.46	0.43	0.57	4.66×10^{-9}	0.989	-0.618	1.279
11	123561942	T	G	nonsense	*OR10D3P*	0.45	0.60	0.40	NS	1.895*	0.342	0.803
16	79799649	GTT	G	frameshift	*PKD1L2*	0.48	0.80	0.09	2.27×10^{-6}	NA	NA	NA
Variants with significant evidence for selection in CHBJPT											**CHBJPT-YRI**	**CEU-CHBJPT**
9	124431062	G	A	nonsense	*OR1B1*	0.31	0.51	0.03	6.83×10^{-6}	0.430	-1.104	NA
9	124431591	C	CA	frameshift	*OR1B1*	0.53	0.43	0.40	6.83×10^{-6}	NA	NA	NA
9	138754316	G	A	nonsense	*LCN10*	0.13	0.21	0.14	NS	NA	2.242*	-2.624
11	4747449	CG	C	frameshift	*OR51F1*	0.17	0.00	0.54	1.32×10^{-8}	NA	NA	NA
16	79799649	GTT	G	frameshift	*PKD1L2*	0.48	0.80	0.09	1.22×10^{-8}	NA	NA	NA
19	17660246	T	TG	frameshift	*UNC13A*	0.00	0.00	0.53	6.28×10^{-6}	NA	NA	NA
19	56787865	T	A	nonsense	*AC018755.8*	0.08	0.63	0.14	2.52×10^{-6}	NA	-0.686	-0.070

chr	pos	ref	alt	type	gene	derived allele frequency			peak spectrum comb. *P*	|iHS|	XP-EHH	XP-EHH
						CEU	CHBJPT	YRI				
Variants with significant evidence for selection in YRI											**CEU-YRI**	**CHBJPT-YRI**
1	156816116	C	T	nonsense	*OR10X1*	0.47	0.58	0.64	4.54×10^{-5}	1.080	-0.325	-0.308
1	169379114	C	T	nonsense	*FMO6P*	0.51	0.61	0.81	NS	2.104*	0.909	-1.191
3	185236988	G	C	splice	*HTR3D*	0.46	0.56	0.79	NS	NA	-1.947*	-1.467
5	131352149	C	CTG	frameshift	*ACSL6*	0.77	0.44	0.00	6.20×10^{-4}	NA	NA	NA
6	31232828	C	T	nonsense	*CCHCR1*	0.44	0.36	0.52	3.38×10^{-9}	0.275	-0.577	-0.574
9	124431591	C	CA	frameshift	*OR1B1*	0.53	0.43	0.40	1.59×10^{-6}	NA	NA	NA
11	48223312	C	G	nonsense	*OR4X2*	0.12	0.16	0.41	NS	2.089*	-0.336	-0.681
11	55096228	C	T	nonsense	*OR4C16*	0.29	0.34	0.09	NS	0.573	-1.386	-1.990*
11	55162598	C	G	nonsense	*OR4P4*	0.09	0.37	0.14	NS	0.014	-1.612*	-1.975*
19	40410860	C	T	nonsense	*FAM187B*	0.28	0.18	0.27	NS	0.451	-1.805*	-1.855*
22	15849049	C	A	nonsense	*GAB4*	0.29	0.41	0.19	NS	0.427	-3.021**	-3.828**

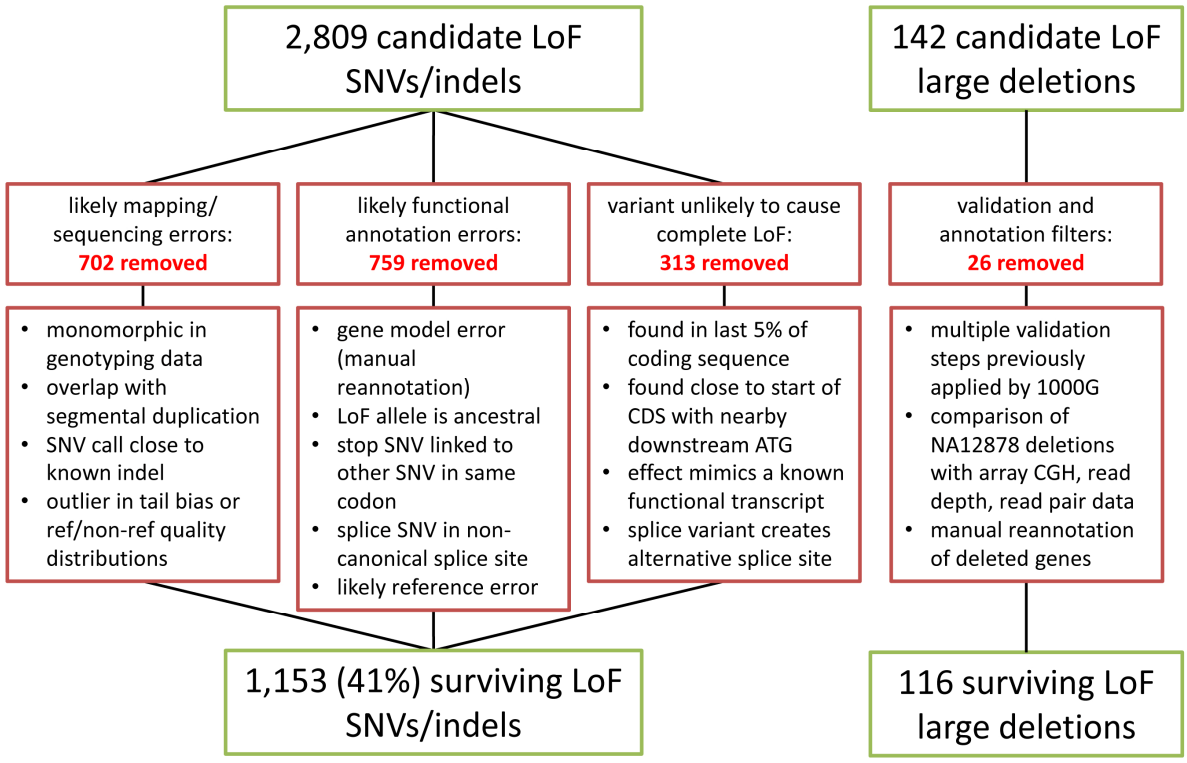


Figure S1. Filtering process for candidate LoF SNVs, indels and large deletions. Details of this process are described in the supplementary text. Note that candidate LoF large deletions had already been subjected to extensive informatic and experimental validation as part of the 1000 Genomes Project pilot analyses.

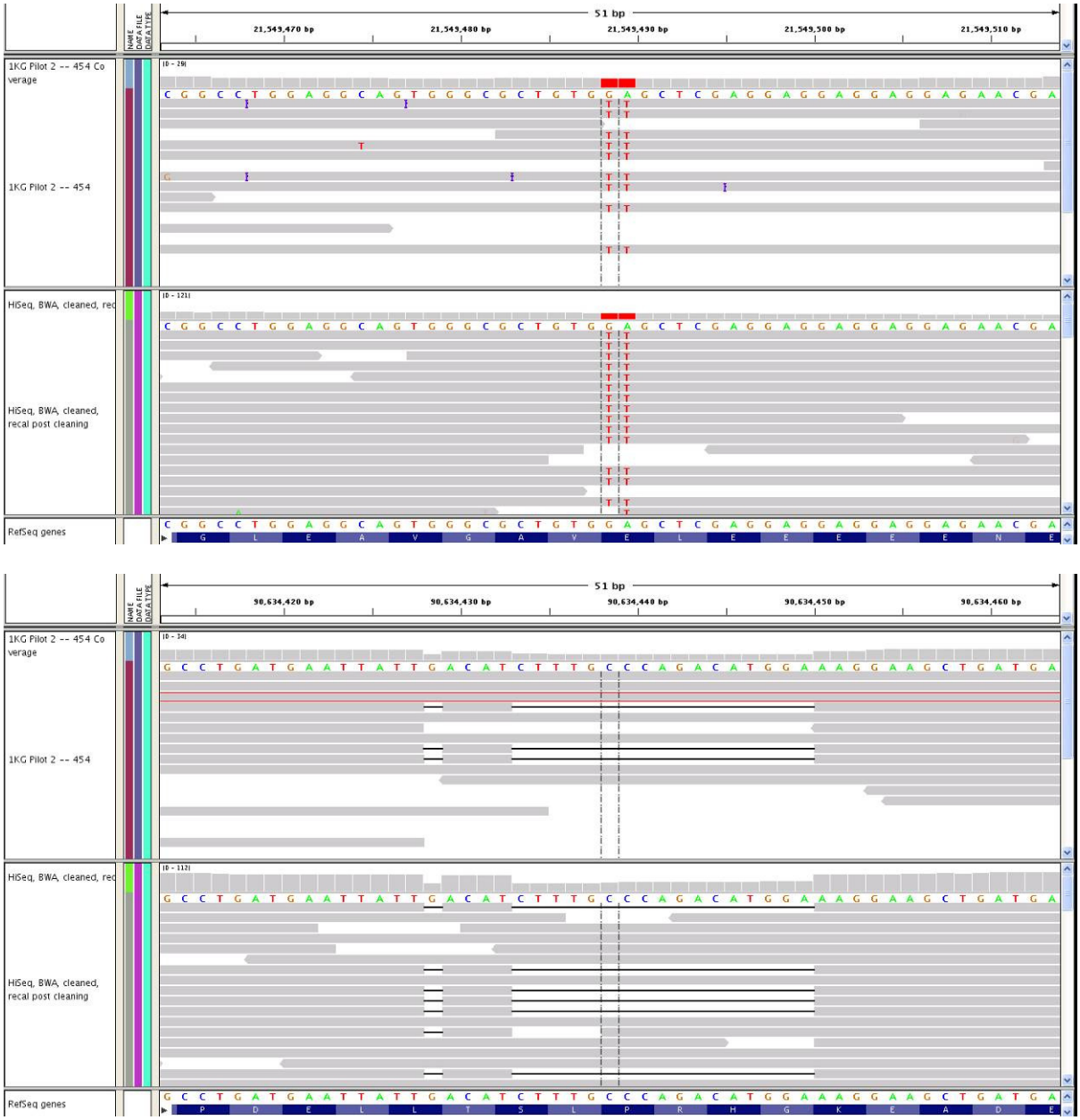


Figure S2. Accurate functional interpretation requires integrating multiple variants on the same haplotype. A. A homozygous SNV annotated as a nonsense (GAG>TAG) polymorphism in the *DNAH11* gene is in fact part of a two-base substitution resulting in a missense change (GAG>TTG; Glu>Leu). **B.** Two apparent heterozygous frameshift coding deletions (1 bp and 17 bp long) are in fact present on the same haplotype, with the combined effect being an in-frame deletion of six amino acids. Both screenshots are taken from analysis of sequence data from NA12878 using Integrative Genomics Viewer; in each case the top panel shows 454 reads, while the bottom panel shows reads from the HiSeq 2000.

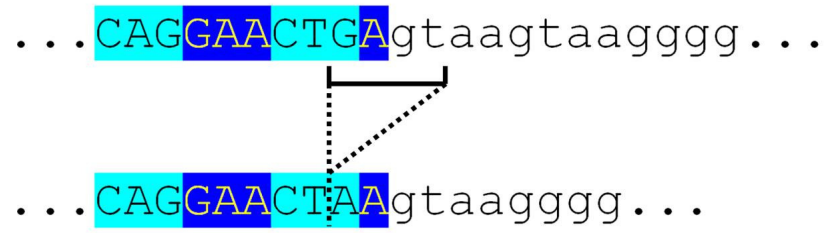


Figure S3. Putative frameshift indels close to or spanning exon splice sites can be rescued by alternative splice sites. A 4 bp deletion spanning a splice site in the *CHIT1* gene creates an alternative splice site that maintains the reading frame and results in a synonymous (Leu>Leu) substitution. Top line: reference allele, with exonic bases in capitals and alternating codons indicated in dark and light blue. Deleted region is indicated with a horizontal line. Final effect of the deletion (including the restored reading frame) is shown on the bottom line.

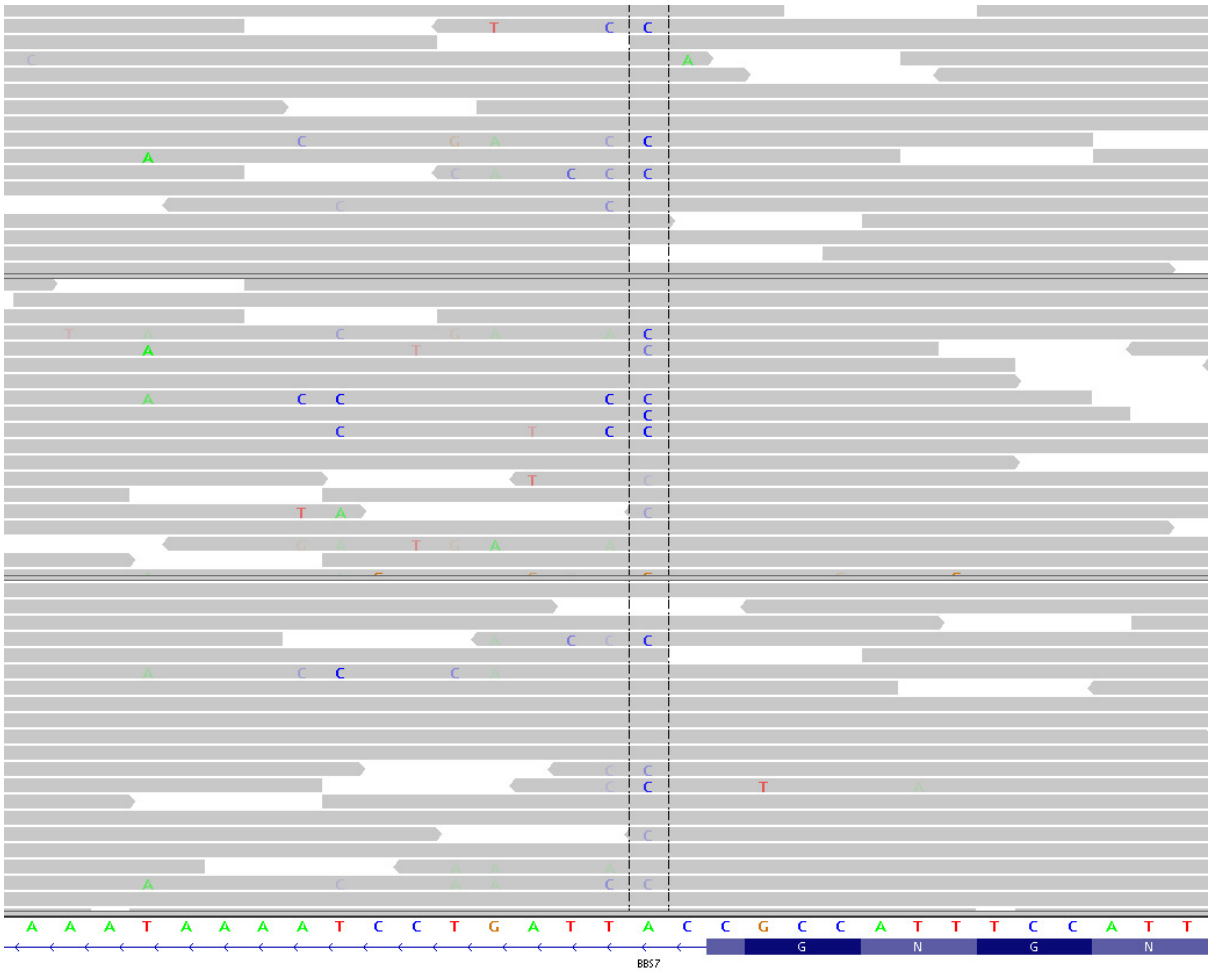
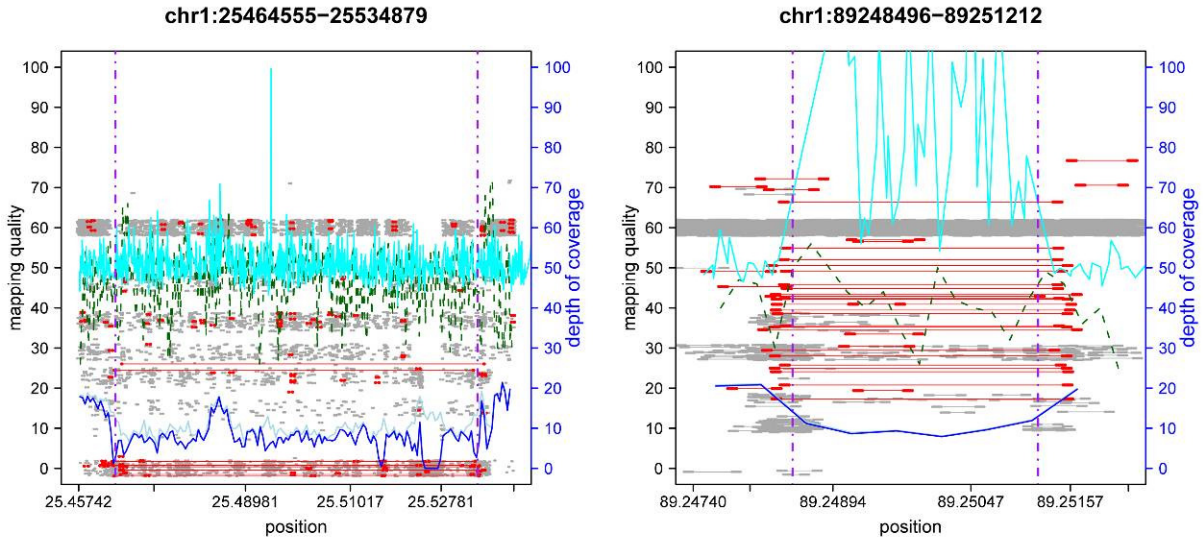
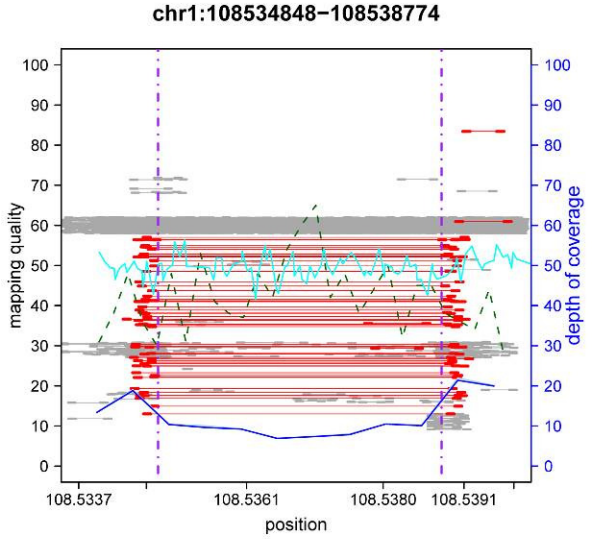
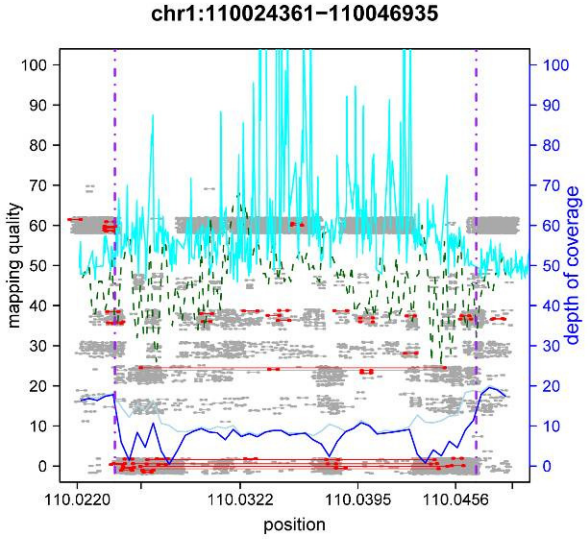


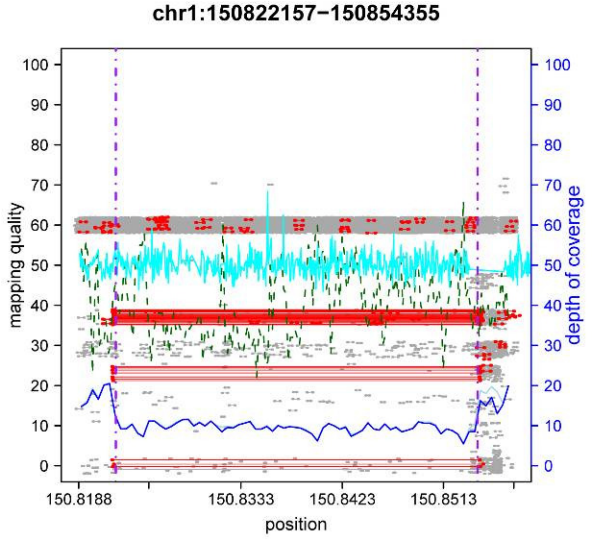
Figure S4. Systematic sequencing error at the site of a reported disease-causing mutation in the *BBS7* gene. An A>G splice mutation at this location has previously been reported as disease-causing in an Italian family (*71*). The 1000 Genomes low-coverage pilot called a A>C substitution at this location in 30 CHB+JPT individuals, which failed to validate in two separate genotyping assays and also revealed an excess of low-quality base calls for the alternative allele ($P = 2.7 \times 10^{-10}$). Figure shows an IGV image of reads spanning this location in CHB (top), JPT (middle) and CEU (bottom), which supports widespread systematic error at this site in Illumina sequencing data.

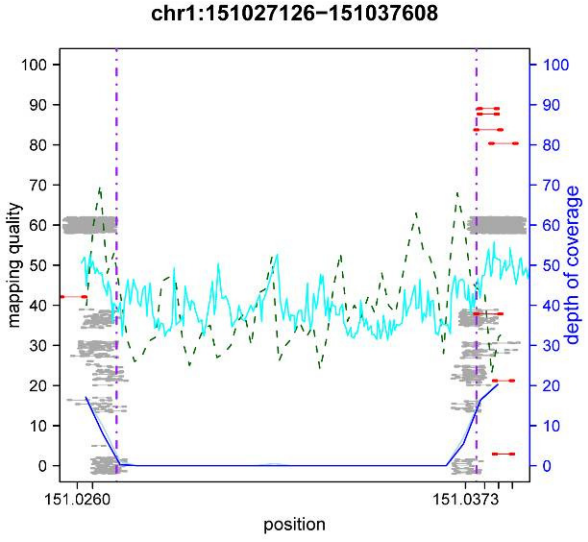

chr1:25464555−25534879
mapping quality
depth of coverage
position
25.45742
25.48981
25.51017
25.52781

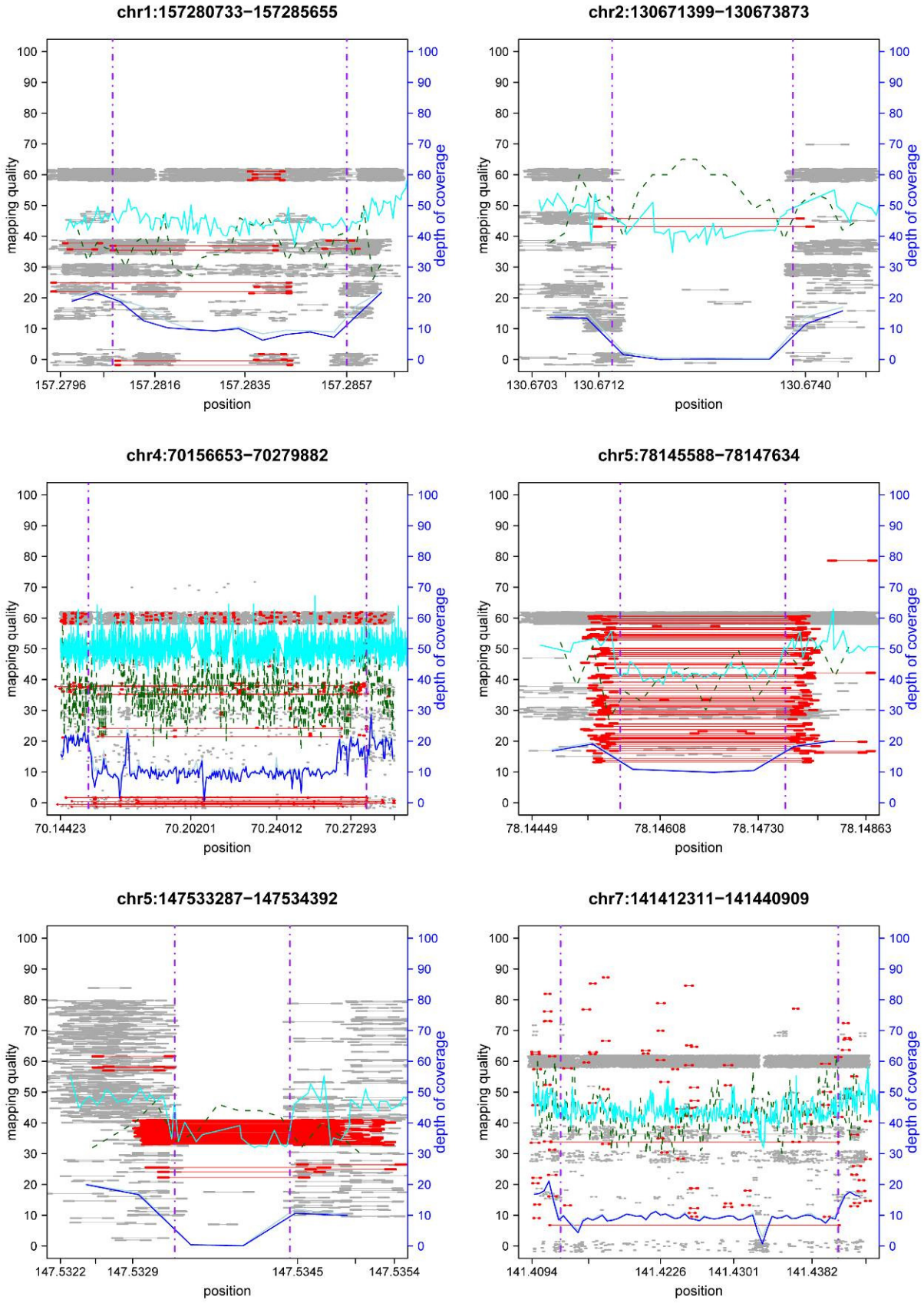
chr1:89248496−89251212
mapping quality
depth of coverage
position
89.24740
89.24894
89.25047
89.25157

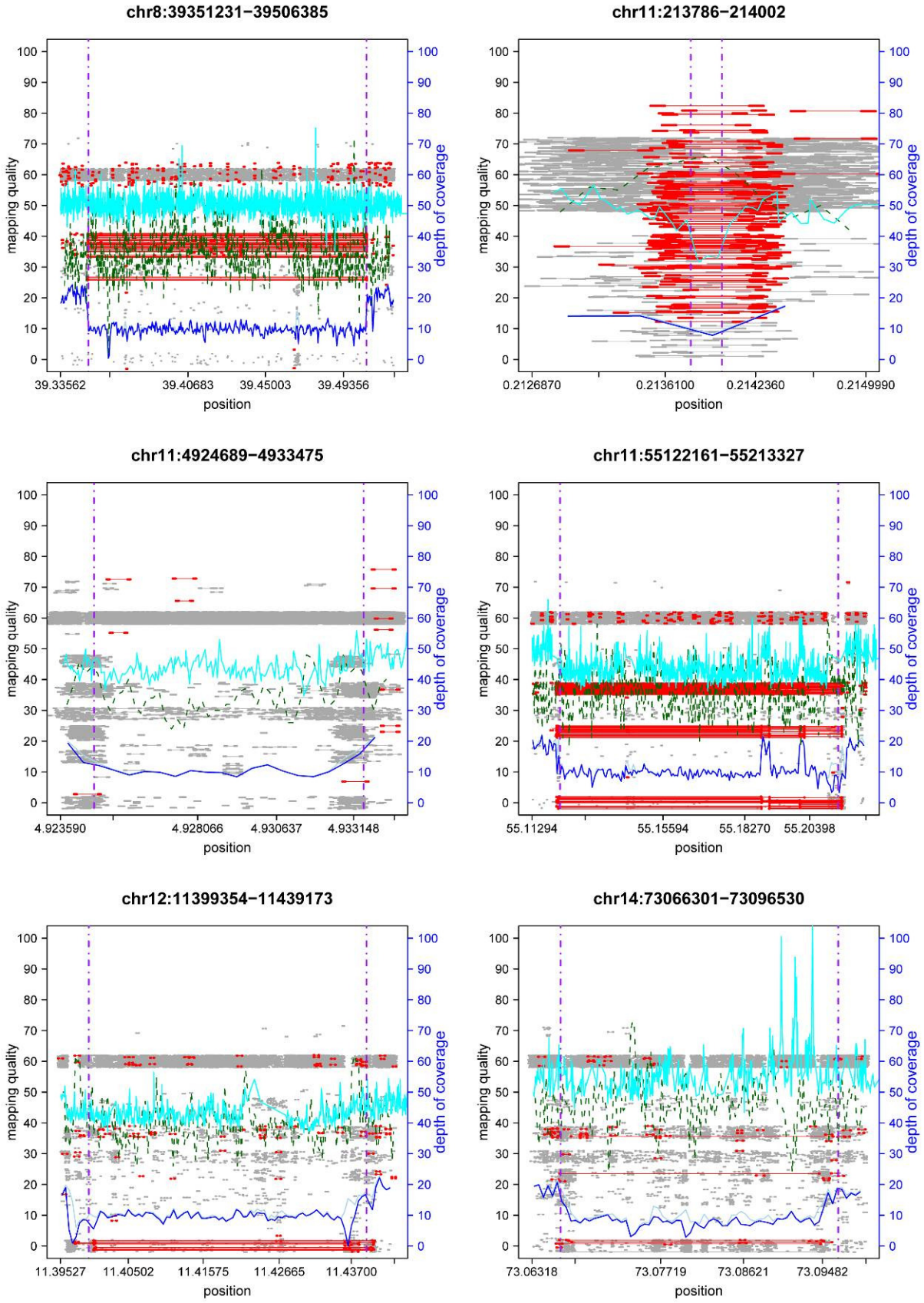

chr1:108534848−108538774
mapping quality
depth of coverage
position
108.5337
108.5361
108.5380
108.5391


chr1:110024361−110046935
mapping quality
depth of coverage
position
110.0220
110.0322
110.0395
110.0456


chr1:150822157−150854355
mapping quality
depth of coverage
position
150.8188
150.8333
150.8423
150.8513


chr1:151027126−151037608
mapping quality
depth of coverage
position
151.0260
151.0373


chr1:157280733−157285655
chr2:130671399−130673873
chr4:70156653−70279882
chr5:78145588−78147634
chr5:147533287−147534392
chr7:141412311−141440909
mapping quality
depth of coverage
position
157.2796
157.2816
157.2835
157.2857
130.6703
130.6712
130.6740
70.14423
70.20201
70.24012
70.27293
78.14449
78.14608
78.14730
78.14863
147.5322
147.5329
147.5345
147.5354
141.4094
141.4226
141.4301
141.4382


chr8:39351231−39506385
chr11:213786−214002
chr11:4924689−4933475
chr11:55122161−55213327
chr12:11399354−11439173
chr14:73066301−73096530
mapping quality
depth of coverage
position
39.33562
39.40683
39.45003
39.49356
0.2126870
0.2136100
0.2142360
0.2149990
4.923590
4.928066
4.930637
4.933148
55.11294
55.15594
55.18270
55.20398
11.39527
11.40502
11.41575
11.42665
11.43700
73.06318
73.07719
73.08621
73.09482

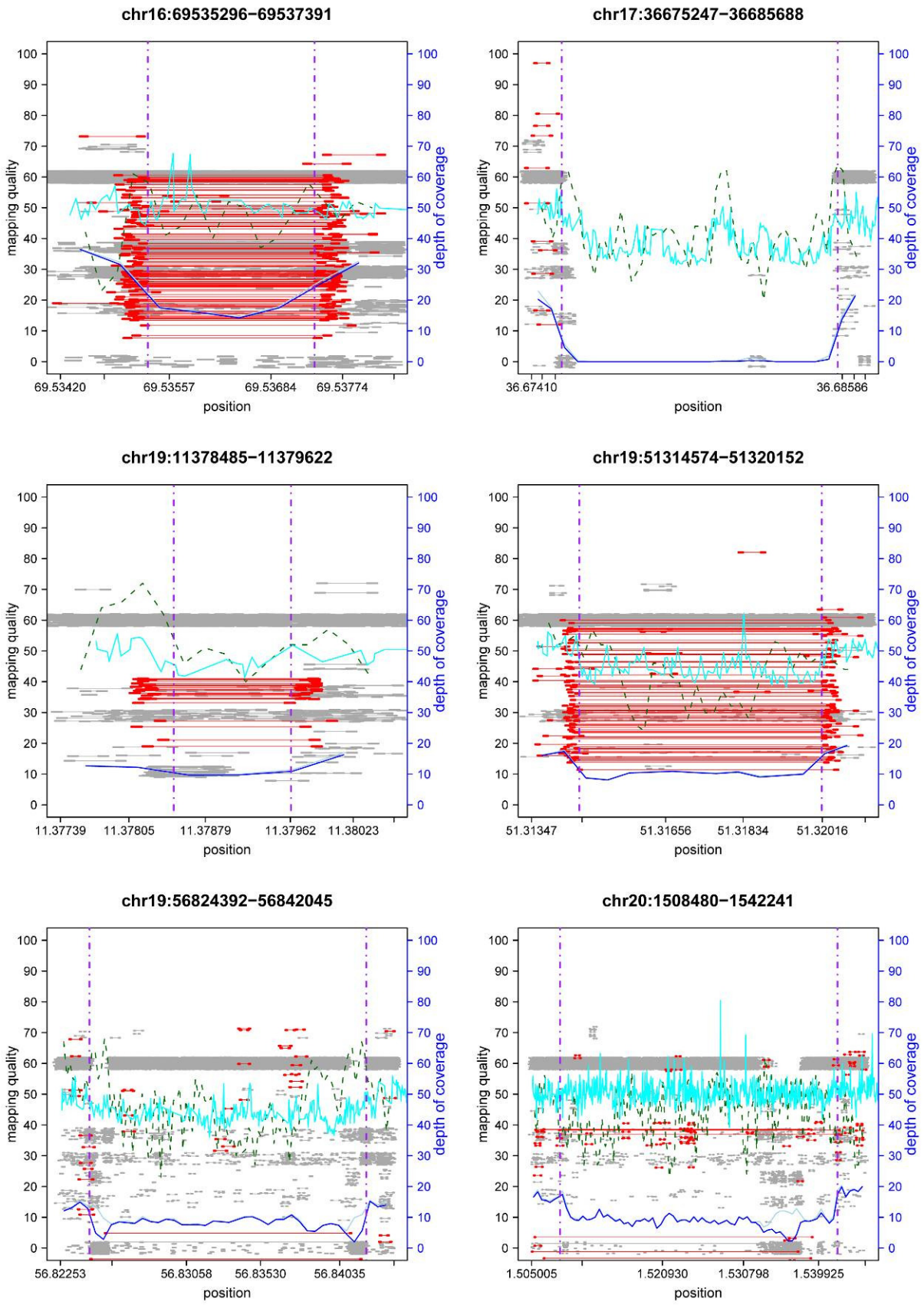


Figure S5. Plots showing evidence for the LoF deletions identified in NA12878. Purple dashed vertical lines indicate the predicted breakpoints of the deletion. Grey and red lines indicate mapped reads from the NA12878 HiSeq data, with the position on the Y axis indicating mapping quality; red lines indicate

mate pairs mapped with an anomalously large insert size, suggestive of a mate pair spanning a deletion. Blue lines towards the bottom of each plot show read depth in 500 bp windows, with the depth of reads with a mapping quality of 0 indicated in light blue, and non-zero quality mapped reads in dark blue. The aqua line across the middle of the plot shows the average intensity of the signal for high-resolution array CGH (42 million probes) analysis of NA12878 performed by Conrad et al. (2010). (Note that as this experiment involved comparative array hybridization with a reference sample, it will not provide support for deletions that are also present in the reference.) Green dashed line shows average GC content (in 200 bp windows) across the region.

References and notes

1. P. C. Ng *et al.*, *PLoS Genet.* **4**, e1000160 (2008).
2. 1000 Genomes Project Consortium, *Nature* **467**, 1061 (2010).
3. K. Pelak *et al.*, *PLoS Genet.* **6**, e1001111 (2010).
4. D. G. MacArthur, C. Tyler-Smith, *Hum. Mol. Genet.* **19**, R125 (2010).
5. S. Balasubramanian *et al.*, *Genes Dev.* **25**, 1 (2011).
6. M. A. DePristo *et al.*, *Nat. Genet.* **43**, 491 (2011).
7. J. Harrow *et al.*, *Genome Biol.* **7 Suppl 1**, 1 (2006).
8. http://vat.gersteinlab.org/
9. See supporting material on *Science* online.
10. R. E. Mills *et al.*, *Nature* **470**, 59 (2011).
11. A. Uzumcu *et al.*, *J. Med. Genet.* **43**, e5 (2006).
12. D. G. MacArthur *et al.*, *Nat. Genet.* **39**, 1261 (2007).
13. Y. Xue *et al.*, *Am. J. Hum. Genet.* **78**, 659 (2006).
14. Z. D. Zhang, A. Frankish, T. Hunt, J. Harrow, M. Gerstein, *Genome Biol.* **11**, R26 (2010).
15. B. Yngvadottir *et al.*, *Am. J. Hum. Genet.* **84**, 224 (2009).
16. J. R. Lupski *et al.*, *N. Engl. J. Med.* **362**, 1181 (2010).
17. J. M. Chen, D. N. Cooper, N. Chuzhanova, C. Ferec, G. P. Patrinos, *Nature Reviews Genetics* **8**, 762 (2007).
18. C. Casola, U. Zekonyte, A. D. Phillips, D. N. Cooper, M. W. Hahn, *Genome Res.* **22**, advance online (2011).
19. N. Huang, I. Lee, E. M. Marcotte, M. E. Hurles, *PLoS Genet.* **6**, e1001154 (2010).
20. Y. Ishimaru *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12569 (2006).
21. A. L. Huang *et al.*, *Nature* **442**, 934 (2006).
22. Wellcome Trust Case Control Consortium, *Nature* **447**, 661 (2007).
23. D. F. Conrad *et al.*, *Nature* **464**, 704 (2010).
24. S. B. Montgomery *et al.*, *Nature* **464**, 773 (2010).
25. J. K. Pickrell *et al.*, *Nature* **464**, 768 (2010).
26. E. Nagy, L. E. Maquat, *Trends Biochem. Sci.* **23**, 198 (1998).
27. M. V. Olson, *Am. J. Hum. Genet.* **64**, 18 (1999).
28. A. E. Fry *et al.*, *Hum. Mol. Genet.* **18**, 2683 (2009).
29. A. H. Bittles, J. V. Neel, *Nat. Genet.* **8**, 117 (1994).
30. A. R. McCune *et al.*, *Science* **296**, 2398 (2002).
31. C. A. Albers *et al.*, *Genome Res.* **21**, 961 (2011).
32. S. A. McCarroll *et al.*, *Nat. Genet.* **40**, 1166 (2008).
33. J. M. Kidd *et al.*, *Genome Res.* **18**, 2016 (2008).
34. A. McKenna *et al.*, *Genome Res.* **20**, 1297 (2010).
35. K. Wang, M. Li, H. Hakonarson, *Nucleic Acids Res.* **38**, e164 (2010).
36. J. M. Kidd *et al.*, *Nature* **453**, 56 (2008).
37. K. Chen *et al.*, *Nat. Methods* **6**, 677 (2009).

38. K. Ye, M. H. Schulz, Q. Long, R. Apweiler, Z. Ning, *Bioinformatics* **25**, 2865 (2009).
39. J. A. Morris, J. C. Randall, J. B. Maller, J. C. Barrett, *Bioinformatics* **26**, 1786 (2010).
40. J. T. Robinson *et al.*, *Nat. Biotechnol.* **29**, 24 (2011).
41. B. Paten, J. Herrero, K. Beal, S. Fitzgerald, E. Birney, *Genome Res.* **18**, 1814 (2008).
42. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
43. R. Mott, *Comput. Appl. Biosci.* **13**, 477 (1997).
44. S. M. Searle, J. Gilbert, V. Iyer, M. Clamp, *Genome Res.* **14**, 963 (2004).
45. E. L. Sonnhammer, J. C. Wootton, *Proteins* **45**, 262 (2001).
46. A. Bateman *et al.*, *Nucleic Acids Res.* **32**, D138 (2004).
47. J. Zhang, L. E. Maquat, *RNA* **2**, 235 (1996).
48. P. Flicek *et al.*, *Nucleic Acids Res.* **39**, D800 (2011).
49. K. D. Pruitt *et al.*, *Genome Res.* **19**, 1316 (2009).
50. S. Katada, M. Tanaka, K. Touhara, *J. Neurochem.* **90**, 1453 (2004).
51. M. Krawczak, J. Reiss, D. N. Cooper, *Hum. Genet.* **90**, 41 (1992).
52. M. Chillon *et al.*, *Am. J. Hum. Genet.* **56**, 623 (1995).
53. D. Baralle, M. Baralle, *J. Med. Genet.* **42**, 737 (2005).
54. K. L. Chambliss *et al.*, *Am. J. Hum. Genet.* **63**, 399 (1998).
55. E. B. Wilson, *J. Amer. Statistical Assoc.* **22**, 209 (1927).
56. J. C. Barrett *et al.*, *Nat. Genet.* **41**, 703 (2009).
57. F. Tajima, *Genetics* **123**, 585 (1989).
58. J. C. Fay, C. I. Wu, *Genetics* **155**, 1405 (2000).
59. R. Nielsen *et al.*, *Genome Res.* **15**, 1566 (2005).
60. S. F. Schaffner *et al.*, *Genome Res.* **15**, 1576 (2005).
61. P. C. Sabeti *et al.*, *Nature* **449**, 913 (2007).
62. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, *PLoS Biol.* **4**, e72 (2006).
63. G. M. Cooper *et al.*, *Genome Res.* **15**, 901 (2005).
64. K. R. Brown, I. Jurisica, *Bioinformatics* **21**, 2076 (2005).
65. A. Ceol *et al.*, *Nucleic Acids Res.* **38**, D532 (2010).
66. T. S. Keshava Prasad *et al.*, *Nucleic Acids Res.* **37**, D767 (2009).
67. J. F. Rual *et al.*, *Nature* **437**, 1173 (2005).
68. I. Lee, Z. Li, E. M. Marcotte, *PLoS ONE* **2**, e988 (2007).
69. I. Lee *et al.*, *Nat. Genet.* **40**, 181 (2008).
70. S. van Dongen, *SIAM J. Matrix Anal. Appl.* **30**, 121 (2008).
71. J. Bin *et al.*, *Hum. Mutat.* **30**, E737 (2009).