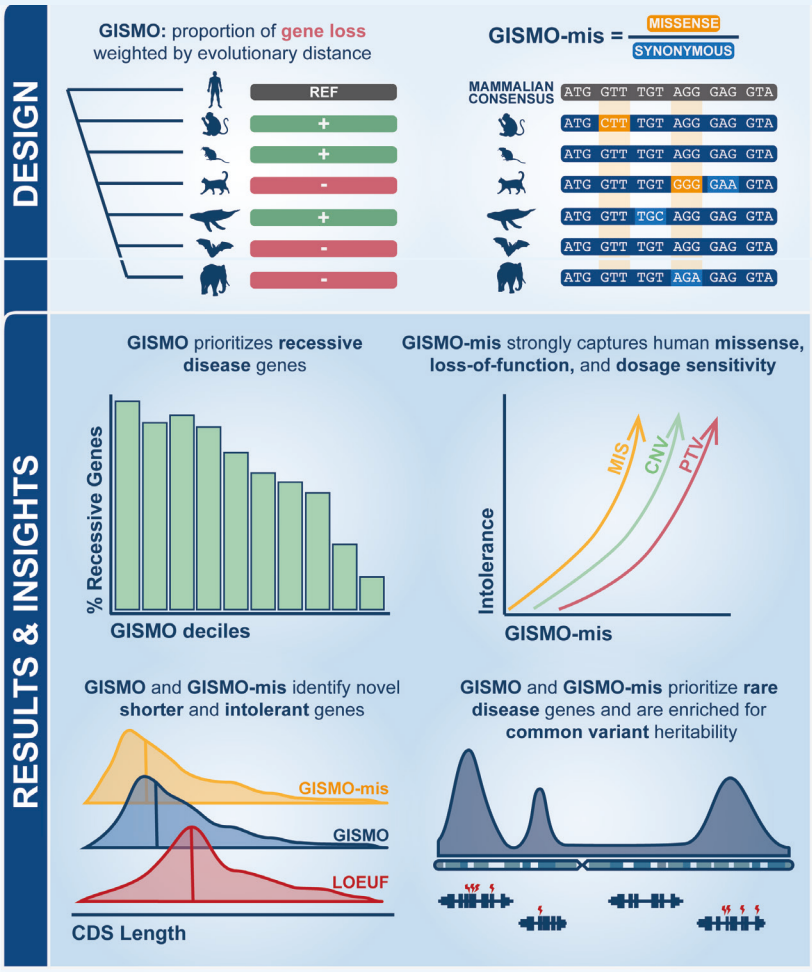
The landscape of gene loss and missense variation across the mammalian tree informs on gene essentiality

Calwing Liao[1,2,3,*], Robert Ye[1,2,3], Franjo Ivankovic[1,2,3], Jack M. Fu[2,3], Raymond Walters[1,2,3], Chelsea Lowther[2,3], Elise Walkanas[2,3], Claire Churchhouse[1,2,3], Kaitlin E. Samocha[1,2,3], Kerstin Lindblad-Toh[2,4], Elinor Karlsson[2,5], Michael Hiller[6,7,8], Michael E. Talkowski[1,2,3], Benjamin M. Neale[1,2,3,*]

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA; [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA; [4]Science for Life Laboratory, Uppsala University, Uppsala, Sweden; [5]Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA; [6]LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany; [7]Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt, Germany; [8]Institute of Cell Biology and Neuroscience, Faculty of Biosciences, Goethe University Frankfurt, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany

ABSTRACT



Background: The degree of gene and sequence preservation across species provides valuable insights into the relative necessity of genes from the perspective of natural selection. Here, we developed novel interspecies metrics across 462 mammalian species, GISMO (Gene identity score of mammalian orthologs) and GISMO-mis (GISMO-missense), to quantify gene loss traversing millions of years of evolution. GISMO is a measure of gene loss across mammals weighed by evolutionary distance relative to humans, whereas GISMO-mis quantifies the ratio of missense to synonymous variants across mammalian species for a given gene.

Rationale: Despite large sample sizes, current human constraint metrics are still not well calibrated for short genes. Traversing over 100 million years of evolution across hundreds of mammals can identify the most essential genes and improve gene-disease association. Beyond human genetics, these metrics provide measures of gene constraint to further enable mammalian genetics research.

Results: Our analyses showed that both metrics are strongly correlated with measures of human gene constraint for loss-of-function, missense, and copy number dosage derived from upwards of a million human samples, which highlight the power of interspecies constraint. Importantly, neither GISMO nor GISMO-mis are strongly correlated with coding sequence length. Therefore both metrics can identify novel constrained genes that were too small for existing human constraint metrics to capture. We also found that GISMO scores capture rare variant association signals across a range of phenotypes associated with decreased fecundity, such as schizophrenia, autism, and neurodevelopmental disorders. Moreover, common variant heritability of disease traits are highly enriched in the most constrained deciles of both metrics, further underscoring the biological relevance of these metrics in identifying functionally important genes. We further showed that both scores have the lowest duplication and deletion rate in the most constrained deciles for copy number variants in the UK Biobank, suggesting that it may be an important metric for dosage sensitivity. We additionally demonstrate that GISMO can improve prioritization of recessive disorder genes and captures homozygous selection.

Conclusions: Overall, we demonstrate that the most constrained genes for gene loss and missense variation capture the largest fraction of heritability, GISMO can help prioritize recessive disorder genes, and identify the most conserved genes across the mammalian tree.

INTRODUCTION

The application of sequencing technologies at scale has enabled the quantification of gene essentiality in humans, particularly for loss-of-function and missense genetic variants[1,2]. From an evolutionary perspective, not all genes are equally impacted by loss-of-function variants, and some genes are under stronger constraint against such variants than others. Genes that are essential for survival, development and reproduction tend to be more constrained, as damaging and loss-of-function variants in these genes lower fitness and often have severe detrimental effects[3,4].

Measuring the selective constraint in humans has been made possible by large-scale genomic datasets such as the Genome Aggregation Database (gnomAD)[2]. The gnomAD consortium provides catalogs of discovered genetic variants across >800,000 human individuals, and metrics such as the loss-of-function observed/expected upper bound fraction (LOEUF) scores and missense equivalent (MOEUF) use estimates of the mutation rate alongside patterns of synonymous mutations to calibrate the expected number of loss-of-function and missense mutations observed in the dataset. Loss-of-function and missense mutations that impact genes with important functions will tend not to propagate across generations, leading to fewer such mutations than expected given the mutation rate[5].

The use of human constraint metrics has already provided valuable insights into the genetic basis of diseases and has been transformative for clinical genetics[6,7]. Genes associated with human genetic disorders that reduce fecundity, such as developmental disorders, are often found to have low LOEUF scores, indicating that these genes carry substantial consequences from a natural selection perspective[8–11].

* Corresponding authors: Calwing Liao (cliao@broadinstitute.org), Benjamin M. Neale (bneale@broadinstitute.org).

One of the main limitations of LOEUF is that genes with shorter sequences have fewer potential loss-of-function mutation sites and attempts to quantify the reduction in such mutations is frustrated by the limited number of potential mutations (i.e., it is difficult to determine whether the number of sites observed is low when the expectation for the number of sites is low itself)[2]. Furthermore, quantifying intraspecies genetic variation requires vastly large sample sizes and can be costly[2,12]. One potential solution is to quantify interspecies constraint across the evolutionary tree.

Orthologs are genes that have evolved from a common ancestral gene through speciation events, typically resulting in similar gene sequences and functions across different species[13]. The identification and characterization of orthologous genes has therefore long had an influential role in understanding the evolutionary relationships and functional conservation of genes across different organisms. This is particularly true in mammals where the study of orthologs has enabled the investigation of the genetic basis of various biological processes and diseases[14]. Mammals are a diverse group of organisms that exhibit a wide range of physiological and anatomical characteristics[15,16]. Despite this diversity, mammals share a common ancestry, and studying the orthologous genes across different mammalian species provides insights into the evolutionary changes that have shaped mammalian biology[17]. Comparative genomics studies have shown that orthologs often retain similar functions, suggesting a strong selective pressure to maintain their essential roles throughout evolution[13,18,19].

Here, we characterized the landscape of gene loss and amino acid (missense) changes across placental mammalian reference genomes, and developed novel metrics for measuring gene loss constraint across 462 mammalian species that cover ~10% of all recognized species in this clade: GISMO (Gene identity score of mammalian orthologs). We additionally developed GISMO-mis (GISMO-missense), which captures the missense to synonymous ratio of a given gene. We find that both metrics capture signals from existing human constraint metrics such as LOEUF/MOEUF, as well as measures of dosage sensitivity for deletion and duplication. GISMO works across longer timescales and therefore has the ability to address issues with short genes in intraspecies metrics such as LOEUF. can help identify shorter constrained genes. We further demonstrate that GISMO can capture common variant heritability and rare variant association to traits with lower fecundity such as neurodevelopmental disorders (NDDs). Finally, we highlight how GISMO can prioritize recessive disease genes and taps into both heterozygous and homozygous selection. Overall, our findings will contribute to our understanding of the intrinsic properties of genes, the maintenance of essential gene functions, and gene prioritization for human disease.

RESULTS

Characterization of gene loss, missense and synonymous variation across mammals

To characterize the distributions of gene loss, missense and synonymous variation, we initially analyzed a comprehensive ortholog data set compiled for 462 placental mammalian species. Since we aim at better understanding human gene constraint, we used the human gene annotation as a reference and the TOGA method to infer orthologs in other mammals[20]. TOGA also provides codon alignments, which we used to detect synonymous and missense mutations, and detects gene-inactivating mutations (frameshift, stop codon, splice site mutations and larger deletions) that we use to detect potential gene loss events in other mammals. We use the term gene loss throughout the manuscript to indicate the absence of a gene encoding an intact reading frame in other mammals, even though our dataset includes human- or primate-specific genes that never existed outside of primates.

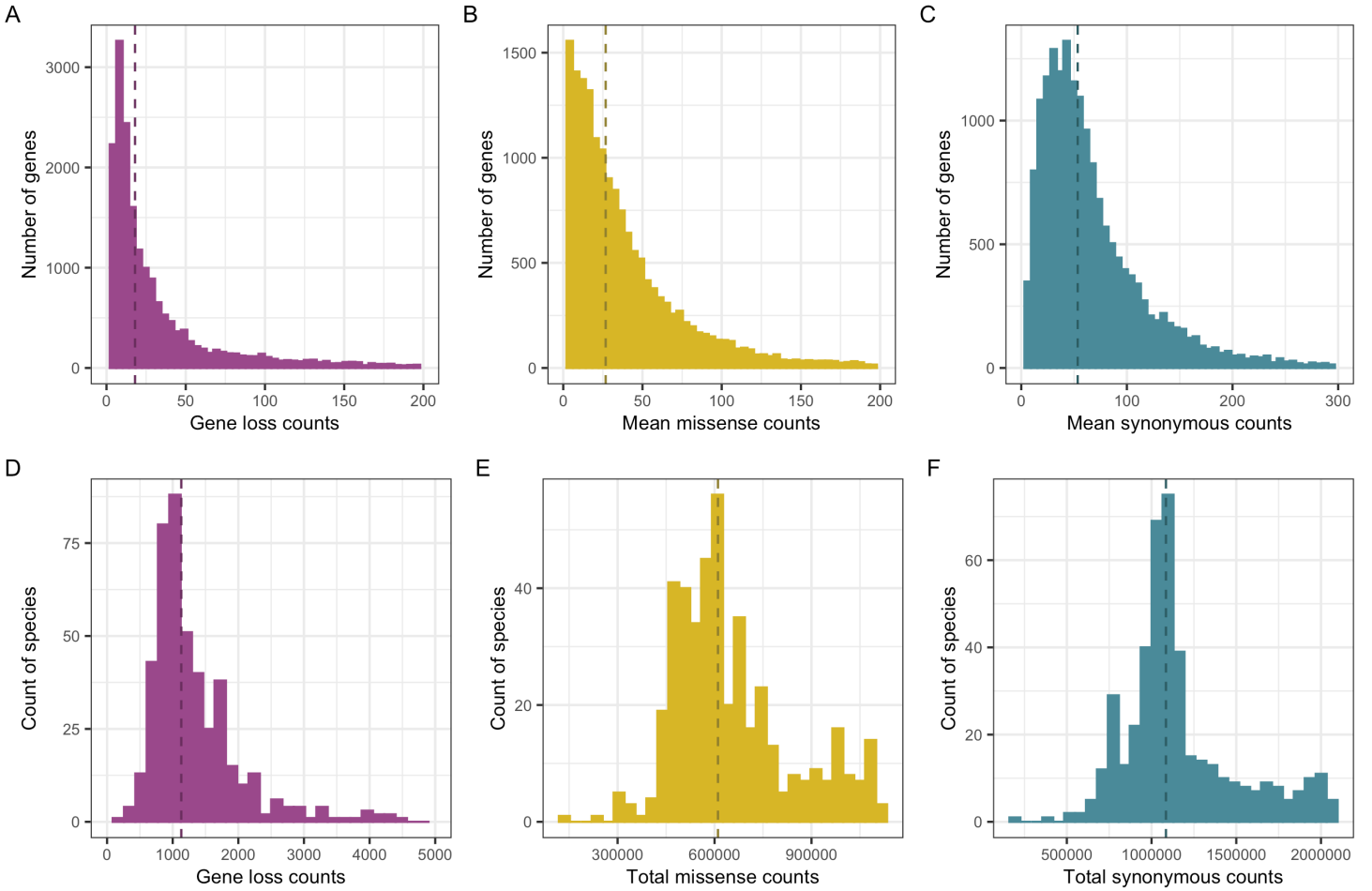


Fig. 1 | Characterization of gene loss, missense and synonymous counts across 462 placental mammals.
(A) Distribution of gene loss. (B) Mean missense variant counts across genes relative to the mammalian consensus. (C) Mean synonymous variant counts across genes relative to the mammalian consensus. (D) Frequency of gene loss per species. (E) Total missense variant counts per species. (F) Total synonymous variant counts per species.

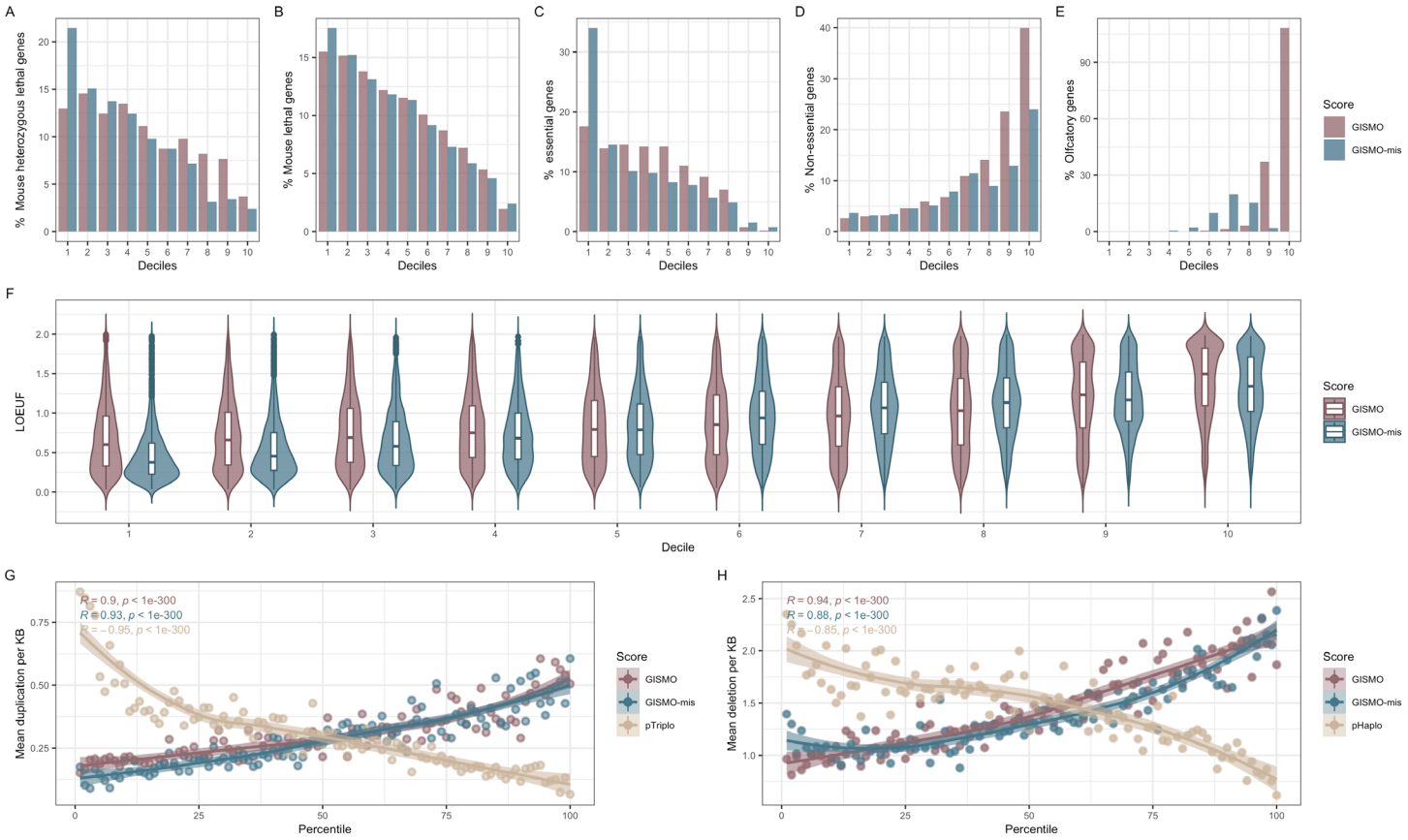


Fig. 2 | The biological and functional spectrum of gene loss and fixed missense differences.
(A-E) The percentage of genes present in each decile of GISMO and GISMO-mis across different genesets (mouse heterozygous lethal, mouse lethal, essential, non-essential, and olfactory genes). The first decile indicates the most constrained. (F) Comparison of GISMO and GISMO-mis to human loss-of-function constraint. (G-H) Comparison of GISMO and GISMO-mis to copy number variants in the UK Biobank (G: duplication, H: deletions). For pHaplo and pTriplo, a higher percentile indicates more constrained whereas for GISMO and GISMO-mis a higher percentile indicates less constrained.

For example, 12 genes have been lost across all species except humans, which were typically pseudogenes or single exon genes (**Supplementary Table 1**).

We found that each gene had a median of 18 gene loss events (**Figure 1A**), whereas across the species, we find a median of 1130 likely gene loss events per species (**Figure 1D**). Here, we define gene loss as disruption of the open reading frame. We identified only six genes with no instances of gene loss across the 462 mammalian species: nucleosome assembly protein 1 like 1 (*NAP1L1*), zinc finger CCCH-type containing 14 (*ZC3H14*), Meis homeobox 2 (*MEIS2*), striatin interacting protein 1 (*STRIP1*), chaperonin containing TCP1 subunit 5 (*CCT5*), and ATPase H+ transporting V0 subunit d1 (*ATP6V0D1*).

For each gene, we calculated the mean missense and synonymous count across the mammals. The median missense average was 27.7 (**Figure 1B**), whereas the median synonymous average was 53.4 (**Figure 1C**). For each species, there was a median of 610,603 missense variants (**Figure 1E**) and 1,084,983 synonymous variants (**Figure 1F**) across all genes.

Defining constraint

To capture human gene essentiality, we developed the novel metric, GISMO (Gene Identity Score of Mammalian Orthologs), which estimates the proportion of gene loss that occurs across the 462 mammals surveyed, weighted by evolutionary distance relative to humans. GISMO is a quantitative metric, where a smaller value indicates less frequent gene loss across mammalia (**Supplementary Table 2**). Next, we sought to quantify the missense differences across mammalia relative to a mammalian consensus sequence, a proxy for molecular adaptation and selection of genes. We calculated a second metric, GISMO-mis, which measures the mean missense to synonymous ratio across mammalia (**Supplementary Table 3**).

We evaluated both metrics against other measures of biological and functional importance of genes. The most constrained GISMO and GISMO-mis deciles had the highest enrichment of genes that cause lethality in mouse knockouts and considered essential in cell lines (**Figures 2A-C**). The least constrained deciles have the highest proportion of non-essential and olfactory genes (**Figure 2 D-E**). We also found that genes that were frequently lost across the evolutionary tree were expressed in less tissues across GTEx (**Supplementary Figure 1**).

We compared both metrics against human constraint metrics and found that GISMO (R = 0.38, P < 1E-300) and GISMO-mis (R = 0.56, P<1E-300) were both strongly associated with LOEUF (**Figure 2F**, **Supplementary Figure 2**), with GISMO-mis having a stronger association. Both metrics also had strong correlations with missense constraint (MOEUF), which was comparable to LOEUF for GISMO-mis (R = 0.61, P < 1E-300) but lower for GISMO (R = 0.33, P < 1E-300). We additionally found that both metrics were strongly associated with dosage sensitivity metrics to predict the strength of selection against loss of a gene copy (pHaplo; R GISMO = -0.32. , P < 1E-300; R GISMO-mis = -0.44, P < 1E-300) and duplication of a gene (pTriplo; RGISMO = -0.32, P < 1E-300; R GISMO-mis = -0.50, P < 1E-300). We additionally compared GISMO-mis with the new gene-level AlphaMissense metric and found a stronger correlation to this metric (R = 0.84, P < 1E-300) than LOEUF or missense constraint.

Next, to assess the relationship between copy number variants (CNVs) and mammalian gene constraint, we leveraged a CNV dataset generated from exome sequencing in the UK Biobank[21] (N = 197,306) that enabled detection of individual gene level variants. The most constrained deciles of both GISMO and GISMO-mis had the lowest number of deletions and duplications. We found that GISMO had an even stronger correlation in the UK Biobank for deletions compared to pHaplo, a dosage-sensitivity metric derived from CNVs aggregated from lower-resolution microarray data across ~1 million human samples[22] (**Figure 2G-H**).

Leveraging GISMO to identify shorter intolerant genes

One pitfall of existing human constraint metrics are the biases with gene length. Particularly, shorter genes have a smaller mutational target and are not well calibrated in many human constraint models. The correlation of GISMO and GISMO-mis with coding sequence (CDS) length was -0.12, P = 3.8E-52 for GISMO, -0.12, P = 4.0E-52 for GISMO-mis, and was several magnitudes lower compared to LOEUF (-0.50, P < 1E-300) (**Figure 3A**). Given that LOEUF may not be well calibrated for many short genes, we sought to assess whether GISMO and GISMO-mis can identify shorter intolerant genes. We defined genes to be GISMO and GISMO intolerance if the genes fell within the top 15% most constrained genes (lowest scores). Both GISMO and GISMO-mis identify much shorter intolerant genes (GISMO median CDS = 1502 base pairs for 2897 genes; GISMO-mis median CDS = 1302 base pairs for 2,502 genes) than LOEUF intolerant genes (median CDS = 2267 base pairs for 2,968 genes). We additionally find that genes predicted to be intolerant by GISMO and GISMO-mis but not LOEUF tend to be much shorter compared to genes that are intolerant to LOEUF and one of the GISMO metrics (**Figure 3B**).

The genes that were considered tolerant for LOEUF (>0.35) and GISMO intolerant were enriched for genes involved in the following pathways/ gene sets in mice: sperm immotility (Bonferroni-adjusted P = 2.17E-03, [23/46 genes]), absent acrosome (Bonferroni-adjusted P, 4.86E-2, [16/30 genes]), abnormal actin cytoskeleton morphology (Bonferroni-adjusted P = 4.86E-02, [15/27 genes]) (**Supplementary Table 4**). Intriguingly, when we explored enrichment of these genes amongst human traits, we found phenotypes related to later onset of disease age, which may reflect changes in fecundity and recent selective pressures due to modern medicine, such as abnormal posterior eye segment morphology (Bonferroni-adjusted P = 8.3E-02, [328/1439 genes]) and lipid accumulation in hepatocytes (Bonferroni-adjusted P = 1.0E-01, 50/149).

Mammalian gene loss and fixed missense differences capture human disease signals

Next, we evaluated whether mammalian gene loss and fixed missense differences could help prioritize genes relevant to human diseases. Given that many disease models are mammals (i.e. mus musculus), assessing whether disease genes are often strongly conserved may help pinpoint what human genes may be reasonably modeled in mammals. First, we assessed common variant heritability. For both metrics, we partitioned heritability across 276 independent traits from the UK Biobank and across several disease traits not well captured in the biobank. We found that there was a strong linear enrichment of heritability across the deciles of both metrics, with lowest enrichment in the least constrained deciles (**Figure 4A**). Second, we looked at rare variant association studies of traits with decreased fecundity, such as autism, schizophrenia, and neurodevelopmental disorders. We found that both GISMO and GISMO-mis constrained genes had a higher association to neurodevelopmental disorders and the metrics were strongly correlated with the association strength (**Figure 4B**, **Supplementary Figure 3-4**).

Improving recessive disease gene prioritization

Given that gene loss in a species can be reflective of selection on homozygous loss-of-function carriers, we hypothesized that GISMO could be used to prioritize recessive genes associated to disease (i.e. genes under homozygous selection should be more constrained by GISMO). To assess this, we assembled a list of 1,183 recessive genes and found that there was enrichment amongst the most constrained deciles (**Figure 4C**), whereas GISMO-mis expectedly did not have strong enrichment. Next, we sought to test whether GISMO may help with prioritization of recessive inheritance candidate genes in the Deciphering Developmental Disorder

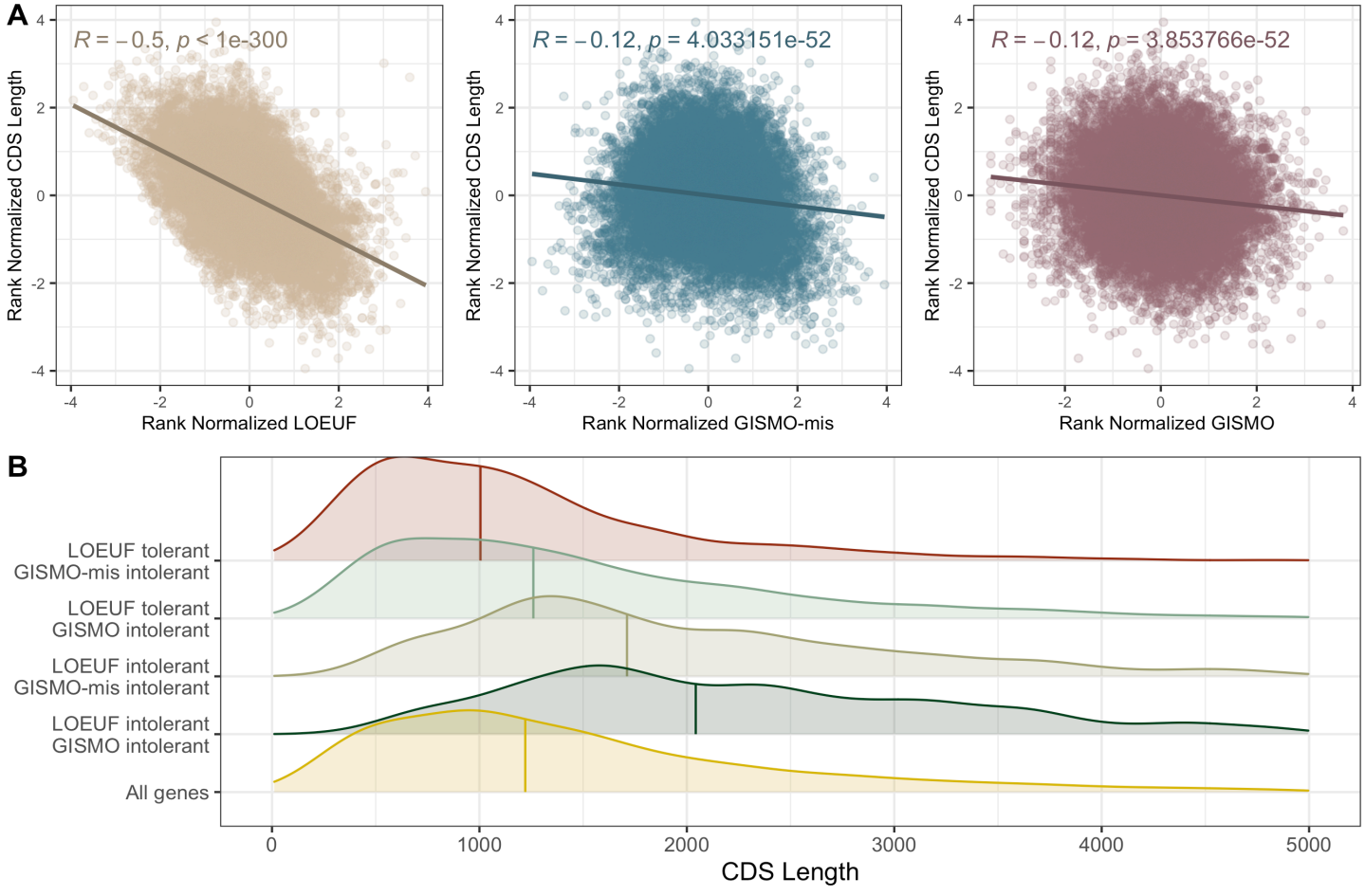


Fig. 3 | GISMO and GISMO-mis capture shorter genes relative to human constraint metrics.
(A) Correlation of GISMO, GISMO-mis and LOEUF against CDS length. The CDS, LOEUF, GISMO and GISMO-mis were inverse rank normalized. A Pearson's correlation was done. (B) Short constrained GISMO and GISMO-mis genes are considered tolerant for LOEUF. Genes were split into categories of GISMO / GISMO-mis intolerant or LOEUF intolerant. We define intolerance for LOEUF as <0.35 and GISMO/GISMO-mis as the first two most constrained deciles respectively.

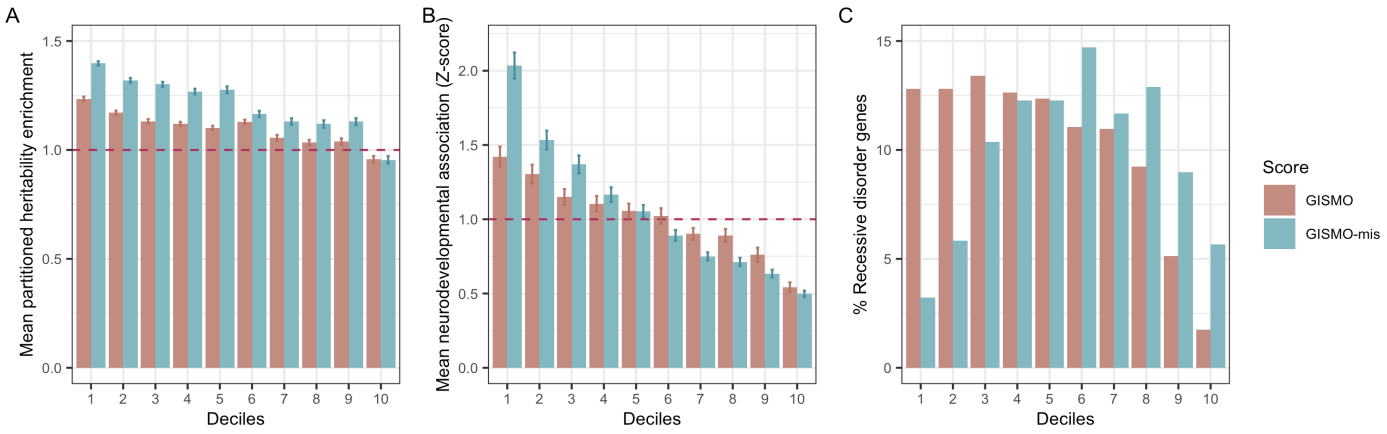


Fig. 4 | GISMO and GISMO-mis capture common variant heritability and rare variant association signals.
(A) Mean enrichment of per-SNV partitioned heritability of XX traits explained by common variants within 100-kb of genes for GISMO and GISMO-mis deciles. (B) Mean neurodevelopmental disorder association across GISMO and GISMO-mis deciles. The association reflects P-values from NDD rare variant associations that were converted to absolute Z-scores. (C) The percentage of genes present in each decile of GISMO and GISMO-mis for 1,183 recessive disorder genes aggregated from OMIM.

Study consisting of 14,000 trios. We found that the GISMO distribution for recessive developmental disorder genes with a high confidence (confident or strong) followed similar distributions as Mendelian recessive genes, whereas the recessive genes with limited and moderate evidence were unclear (**Supplementary Figure 5**). We additionally found that the limited and moderate evidence categories prioritized genes with high constraint for LOEUF, a reflection of strong heterozygous selection, which are unlikely gene candidates for recessive disorders.

DISCUSSION

Evolution and selection are powerful mechanisms that provide insights into essential biology. Comparative genomics leverages millions of years of evolution and selection to understand the importance of genes across various environmental influences. Here, we characterized gene loss and fixed missense differences across the placental mammalian tree, including genomes that cover ~10% of all recognized species. We found that as expected, most genes are present in the majority of mammals, yet any given species may have upwards of over a thousand genes lost. This suggests that certain gene loss is concentrated in a highly select few genes that are tolerant to this loss. Similarly, missense fixation events tend to be low for most genes, and expectedly fixation differences occur more frequently in genes that are less important for biology. We developed novel metrics, GISMO and GISMO-mis, based on the current largest and most comprehensive set of mammalian species representing roughly 10% of mammals to quantify these selection measures.

GISMO is the first metric, to our knowledge, to quantify gene loss across the mammalian tree. We found that GISMO is well saturated, given that only 6 genes did not have any gene loss events. Amongst these 6 conserved genes, we found that they were related to DNA replication, polyA binding for RNA, and development. In contrast to human-based metrics from gnomAD such as LOEUF and pLI, where over 1000 genes do not have an observed loss of function event across 141,456 individuals in gnomAD v2.1 and 832 genes across 807,162 individuals in gnomAD v4.0. We found that most genes that were likely to be human-specific tended to be enriched for single exon and pseudogenes, likely reflecting potential noise or events that are not real or biologically meaningful. We benchmarked these metrics against known gene sets and model organism data and highlighted the potential for identifying disease genes and function.

There are several advantages of harnessing mammalian biology in the form of these metrics. We demonstrated that both GISMO and GISMO-mis can identify shorter constrained genes under selection, which are much shorter than LOEUF constrained genes. In biology, gene essentiality does not depend on gene length; new metrics that capture gene essentiality without biases from gene length are quintessential for understanding the biological and functional spectrum. We also show that GISMO is able to help prioritize recessive disorder genes, which most constraint metrics are not well calibrated to do. Prior studies have highlighted how recessive metrics such as pRec, which measures probability of a gene being recessive, cannot differentiate weak selection on heterozygotes from homozygotes, a limitation of human genetic data[23]. We further show that combining GISMO and heterozygous selection metrics such as LOEUF can help improve clinical prioritization of genes. For instance, a gene candidate is much more likely to be recessive if the GISMO score is low and the LOEUF score is high. We additionally find that GISMO intolerant and LOEUF tolerant genes are important for fertility and reproductive success, with pathway enrichments such as immotile sperm in mice.

We further emphasize that the strong correlations between GISMO and human constraint metrics highlight the power of interspecies over intraspecies metrics. Particularly, the high costs and sample sizes required to generate calibrated intraspecies human constraint metrics such as LOEUF (>800,000 individuals) and pHaplo/pTriplo (>1 million individuals) in contrast to 462 mammals for GISMO.

Moreover, in human genetics, constraint metrics have been transformative for many different subfields. In particular, the probability of loss-of-function intolerance (pLI) and subsequent loss-of-functional observed/expected upper bound fraction (LOEUF) metrics, derived from large-scale sequencing studies, have become standard parts of genetic and genomic analysis pipelines. These metrics have revolutionized a number of workflows in human genetics, including filtering and prioritization of genes and how likely they are to have phenotypic impact for association studies, identification of which genes or variants are likely under heterozygous loss-of-function selection, as well as clinical interpretation of patient variation. Importantly, the GISMO metrics will allow non-human mammalian researchers to have constraint metrics to further enable and advance the analytical framework in the large field.

An open question from these analyses was whether mammalian gene loss and fixed missense differences could be leveraged to capture important biological insights into human diseases and traits. To explore this question, we analyzed an array of quantitative traits, human disease phenotypes, and gene sets robustly associated with rare and common diseases. We found significant common variant heritability enrichment across a vast number of traits in the most constrained deciles of both GISMO and GISMO-mis. We similarly find both metrics capture rare variant association to traits with decreased fecundity. We posit that GISMO and GISMO-mis can provide orthogonal levels of disease evidence and may help with increasing disease association power. Moreover, we show that the majority of rare and common variant heritability is concentrated in genes that are strongly conserved across species. This reinforces the potential utility of mammalian models for dissection of heritable quantitative traits in humans.

Despite the increased power and clear value demonstrated herein for derived metrics based on mammalian orthologs such as GISMO and GISMO-mis, there remains several limitations in the derivation of these metrics. Mammalian gene conservation may not reflect the human selective pressure and recent innovation that has affected the selection regime of certain traits. It is also important to understand the practical use of GISMO

given it taps into both homozygous and heterozygous selection; pairing with additional constraint metrics such as LOEUF and GISMO-mis can help tease apart these relationships. Moreover, the species included may be subjected to survivorship bias and accessibility bias, so further sampling of a broader spectrum of mammals will improve metrics like GISMO and GISMO-mis.

Overall, we demonstrate that mammalian gene loss and missense fixation are important measures of selection. We developed several powerful metrics to quantify evolutionary constraints for gene loss (GISMO) and molecular adaptation (GISMO-mis). Our estimates provide informative ranking of gene importance, which ultimately allow us to better understand gene essentiality and disease association.

METHODS

Development of the GISMO metric

To infer orthologs, the Tool to infer Orthologs from Genome Alignments (TOGA)(20) was used with the human GENCODE 38 annotation[24] across 462 mammals. Since for some species TOGA data was available for multiple assemblies, we selected only the assembly with the highest contig N50. An orthologous gene is considered lost when and only when all transcripts of the respective gene are categorized as lost. To assess the type of orthology, TOGA subsequently evaluates, for each human reference gene, the classification of all its corresponding orthologous loci and which reference genes were annotated.

GISMO was calculated with the following formula:

$$GISMO_{Gene} = \frac{\sum_i^n Gene\ Loss_i Evolutionary\ Distance_i}{\sum_i^n Evolutionary\ Distance_i}$$

A 95% confidence interval was simulated using a binomial distribution and the upper 95% confidence interval was used. Briefly, the binomial probability was estimated for each phylogenetic order and counts were simulated 10,000 times for the 462 species. Each count was subsequently weighted by the evolutionary distance relative to humans, where 1 = gene loss. Evolutionary distance was standardized by dividing by the maximum evolutionary distance amongst the mammals used to generate GISMO.

Development of the GISMO-mis

To calculate GISMO-mis, multiple codon alignments including up to 462 mammals have been generated using MACSE v2[25] and were downloaded from http://genome.senckenberg.de/download/TOGA/. To generate codon alignments, the following selection procedures were performed: 1) Human transcript with the longest coding sequence length, 2) Orthologs were considered if they were classified as intact, partial intact, or uncertain loss. To ensure alignments were mostly 1 to 1 orthologs, if a query species has more than four predicted orthologs, this species was not included in the multiple codon alignment. Additionally, if the gene did not have a single ortholog for at least 75% of all query species, a multiple codon alignment was not computed for this gene. Subsequently for each gene, a mammalian consensus transcript sequence across all species was generated and used as a reference. To determine each consensus transcript sequence, all mammalian sequences for that gene were considered at a codon-level resolution and the most represented codon(s) across all queried species was chosen. In the event of tied codon counts, the consensus sequence permits multiple possible reference codons. Next, missense and synonymous counts across each species were quantified against the mammalian consensus sequence for each transcript. Any codon comparisons involving gaps in either reference or queried codons were not considered. In the case of multiple reference codons at a given position, missense and synonymous counts were conservatively generated by only classifying a queried codon as missense if none of the reference codons were synonymous with the queried codon. Similarly in the case of a codon with ambiguous nucleotides, each ambiguous nucleotide was exhaustively replaced until either 1) a synonymous result was achieved, after which the query would be classified as synonymous, or 2) no synonymous result was achieved across all potential substitutions, after which the query would be classified as missense. Finally for each gene, the mean missense to synonymous ratio across all queried species was calculated.

Benchmarking against gene sets and pathways

To benchmark GISMO, we compared GISMO against several independent gene sets: lethal mouse, olfactory, essential genes, and non-essential genes. These were the same genesets previously used in gnomAD benchmarking to have a reasonable comparison[2]. Both GISMO and GISMO-mis were split into deciles, where the lowest decile (1st) represented the most constrained. Additionally, data from GTEx 53 was used to assess how expressed genes are across the different number of tissues. Genes were considered expressed in a tissue with a TPM > 0.3. Pathway enrichment was done using https://toppgene.cchmc.org/. A recessive gene set was curated from OMIM[26], which included a total of 1,183 autosomal recessive genes.

Correlation between GISMO and GISMO-mis against independent constraint metrics

To assess whether GISMO and GISMO-mis are associated with different constraint scores, both metrics were correlated against other gene-level metrics. First, the GISMO metrics were compared against two human constraint metrics derived from gnomAD for loss-of-function constraint (LOEUF) and missense constraint (MOEUF). Next, to test whether GISMO and GISMO-mis can capture copy number variant dosage sensitivity, the scores were benchmarked against pTriplo (triplosensitivity) and pHaplo (haploinsufficiency)[22]. We additionally benchmarked GISMO and GISMO-mis against gene-level AlphaMissense scores[27], which measures the effects of missense variation on predicted structural context from AlphaFold. A Spearman's correlation was used for all correlations.

Partitioned heritability across independent traits

To assess the distribution of heritability enrichment across GISMO and GISMO-mis, LD score regression (LDSC) was used to partition heritability across gene deciles of both metrics[28–30]. For both metrics, we included a 100kb flanking region both up and downstream for each gene and, in conjunction with genotype data from the 1000 Genomes Project. The SNPs were restricted to HapMap3 SNPs with an estimated annotation-specific LD scores using a 1cM window. Next, partitioned heritability was applied to 276 independent traits from the UK Biobank (https://www.nealelab.is/uk-biobank/), as well as additional disease traits such as schizophrenia, bipolar disorder, autism spectrum disorder, attention-deficit hyperactivity disorder, and coronary artery disease[31–35]. Phenotypes were selected from the UK Biobank based on having a significant p-value ($P < 0.05$) after Bonferroni correction. Additionally, phenotypes were assessed for phenotypic correlation and only independent traits were included. The baseline model (v2.2), which includes 74 annotations to capture genomic properties, was included alongside the estimated LD scores as a predictor in the LD Score regression. HapMap3 SNPs, excluding the HLA region, were used as default regression SNPs.

Assess GISMO and GISMO-mis coding sequence biases

To assess whether GISMO and GISMO-mis are biased by coding sequence length, we correlated against the coding sequencing length (CDS) of each gene in both metrics. We compared against the gold standard for loss-of-function constraint, LOEUF from gnomAD. Given that LOEUF had a significantly stronger correlation with CDS relative to GISMO and GISMO-mis, we hypothesized that the GISMO metrics may help prioritize short essential genes. We considered genes with a LOEUF score <0.35 as LOEUF-constrained and intolerant, representing roughly the top 15% most constrained genes. Similarly, for both GISMO and GISMO-mis, we took the top 15% most constrained genes, which we considered intolerant. We compared the CDS length distribution for GISMO and GISMO-mis intolerant and LOEUF tolerant.

ACKNOWLEDGEMENTS

C.L. is funded by the CIHR Banting Fellowship. K.L.T is a Distinguished professor funded by the Swedish Research Council. M.H was supported by the LOEWE-Centre for Translational Biodiversity Genomics (TBG) funded by the Hessen State Ministry of Higher Education, Research and the Arts (HMWK).

AUTHOR CONTRIBUTIONS

C.L. and B.M.N. came up with the project. C.L. and R.Y. conducted all analyses and wrote the manuscript. F.I., J.M.F., C.L, E.V. provided analytical and scientific assistance. B.M.N., M.E.T. supervised the project. K.E.S., K.L.T., E.K., M.H., C.C., R.W. provided additional supervision and edits to the manuscript.

CONFLICTS OF INTEREST

B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora. K.E.S. has received support from Microsoft for work related to rare disease diagnostics. All other authors report no relevant conflicts of interest.

CODE AND DATA AVAILABILITY

All code and relevant data are available on GitHub: https://github.com/cliao5/GISMO/.

REFERENCES

1. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

2. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 590, E53 (2021).

3. J. Kaplanis, K. E. Samocha, L. Wiel, Z. Zhang, K. J. Arvai, R. Y. Eberhardt, G. Gallone, S. H. Lelieveld, H. C. Martin, J. F. McRae, P. J. Short, R. I. Torene, E. de Boer, P. Danecek, E. J. Gardner, N. Huang, J. Lord, I. Martincorena, R. Pfundt, M. R. F. Reijnders, A. Yeung, H. G. Yntema, Deciphering Developmental Disorders Study, L. E. L. M. Vissers, J. Juusola, C. F. Wright, H. G. Brunner, H. V. Firth, D. R. FitzPatrick, J. C. Barrett, M. E. Hurles, C. Gilissen, K. Retterer, Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature 586, 757–762 (2020).

4. I. Agarwal, Z. L. Fuller, S. R. Myers, M. Przeworski, Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. Elife 12 (2023).

5. O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, E. S. Lander, Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. U. S. A. 111, E455–64 (2014).

6. L. C. Walker, M. de la Hoya, G. A. R. Wiggins, A. Lindy, L. M. Vincent, M. T. Parsons, D. M. Canson, D. Bis-Brewer, A. Cass, A. Tchourbanov, H. Zimmermann, A. B. Byrne, T. Pesaran, R. Karam, S. M. Harrison, A. B. Spurdle, L. G. Biesecker, S. M. Harrison, A. A. Tayoun, J. S. Berg, S. E. Brenner, G. R. Cutting, S. Ellard, M. S. Greenblatt, P. Kang, I. Karbassi, R. Karchin, J. Mester, A. O'Donnell-Luria, T. Pesaran, S. E. Plon, H. L. Rehm, N. T. Strande, S. V. Tavtigian, S. Topper, Using the ACMG/AMP framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the ClinGen SVI Splicing Subgroup. Am. J. Hum. Genet. 110, 1046–1067 (2023).

7. S. Gudmundsson, M. Singer-Berk, N. A. Watts, W. Phu, J. K. Goodrich, M. Solomonson, Genome Aggregation Database Consortium, H. L. Rehm, D. G. MacArthur, A. O'Donnell-Luria, Variant interpretation using population databases: Lessons from gnomAD. Hum. Mutat. 43, 1012–1030 (2022).

8. J. M. Fu, F. K. Satterstrom, M. Peng, H. Brand, R. L. Collins, S. Dong, B. Wamsley, L. Klei, L. Wang, S. P. Hao, C. R. Stevens, C. Cusick, M. Babadi, E. Banks, B. Collins, S. Dodge, S. B. Gabriel, L. Gauthier, S. K. Lee, L. Liang, A. Ljungdahl, B. Mahjani, L. Sloofman, A. N. Smirnov, M. Barbosa, C. Betancur, A. Brusco, B. H. Y. Chung, E. H. Cook, M. L. Cuccaro, E. Domenici, G. B. Ferrero, J. J. Gargus, G. E. Herman, I. Hertz-Picciotto, P. Maciel, D. S. Manoach, M. R. Passos-Bueno, A. M. Persico, A. Renieri, J. S. Sutcliffe, F. Tassone, E. Trabetti, G. Campos, S. Cardaropoli, D. Carli, M. C. Y. Chan, C. Fallerini, E. Giorgio, A. C. Girardi, E. Hansen-Kiss, S. L. Lee, C. Lintas, Y. Ludena, R. Nguyen, L. Pavinato, M. Pericak-Vance, I. N. Pessah, R. J. Schmidt, M. Smith, C. I. S. Costa, S. Trajkova, J. Y. T. Wang, M. H. C. Yu, D. J. Cutler, S. De Rubeis, J. Buxbaum, M. J. Daly, B. Devlin, K. Roeder, S. J. Sanders, M. E. Talkowski, Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. 54, 1320–1331 (2022).

9. T. Singh, T. Poterba, D. Curtis, H. Akil, M. Al Eissa, J. D. Barchas, N. Bass, T. B. Bigdeli, G. Breen, E. J. Bromet, P. F. Buckley, W. E. Bunney, J. Bybjerg-Grauholm, W. F. Byerley, S. B. Chapman, W. J. Chen, C. Churchhouse, N. Craddock, C. M. Cusick, L. DeLisi, S. Dodge, M. A. Escamilla, S. Eskelinen, A. H. Fanous, S. V. Faraone, A. Fiorentino, L. Francioli, S. B. Gabriel, D. Gage, S. A. Gagliano Taliun, A. Ganna, G. Genovese, D. C. Glahn, J. Grove, M.-H. Hall, E. Hämäläinen, H. O. Heyne, M. Holi, D. M. Hougaard, D. P. Howrigan, H. Huang, H.-G. Hwu, R. S. Kahn, H. M. Kang, K. J. Karczewski, G. Kirov, J. A. Knowles, F. S. Lee, D. S. Lehrer, F. Lescai, D. Malaspina, S. R. Marder, S. A. McCarroll, A. M. McIntosh, H. Medeiros, L. Milani, C. P. Morley, D. W. Morris, P. B. Mortensen, R. M. Myers, M. Nordentoft, N. L. O'Brien, A. M. Olivares, D. Ongur, W. H. Ouwehand, D. S. Palmer, T. Paunio, D. Quested, M. H. Rapaport, E. Rees, B. Rollins, F. K. Satterstrom, A. Schatzberg, E. Scolnick, L. J. Scott, S. I. Sharp, P. Sklar, J. W. Smoller, J. L. Sobell, M. Solomonson, E. A. Stahl, C. R. Stevens, J. Suvisaari, G. Tiao, S. J. Watson, N. A. Watts, D. H. Blackwood, A. D. Børglum, B. M. Cohen, A. P. Corvin, T. Esko, N. B. Freimer, S. J. Glatt, C. M. Hultman, A. McQuillin, A. Palotie, C. N. Pato, M. T. Pato, A. E. Pulver, D. St. Clair, M. T. Tsuang, M. P. Vawter, J. T. Walters, T. M. Werge, R. A. Ophoff, P. F. Sullivan, M. J. Owen, M. Boehnke, M. C. O'Donovan, B. M. Neale, M. J. Daly, Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 604, 509–516 (2022).

10. D. S. Palmer, D. P. Howrigan, S. B. Chapman, R. Adolfsson, N. Bass, D. Blackwood, M. P. M. Boks, C.-Y. Chen, C. Churchhouse, A. P. Corvin, N. Craddock, D. Curtis, A. Di Florio, F. Dickerson, N. B. Freimer, F. S. Goes, X. Jia, I. Jones, L. Jones, L. Jonsson, R. S. Kahn, M. Landén, A. E. Locke, A. M. McIntosh, A. McQuillin, D. W. Morris, M. C. O'Donovan, R. A. Ophoff, M. J. Owen, N. L. Pedersen, A. Reif, N. Risch, C. Schaefer, L. Scott, T. Singh, J. W. Smoller, M. Solomonson, D. S. Clair, E. A. Stahl, A. Vreeker, J. T. R. Walters, W. Wang, N. A. Watts, R. Yolken, P. P. Zandi, B. M. Neale, Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. Nat. Genet. 54, 541–547 (2022).

11. Epi25 Collaborative, S. Chen, B. M. Neale, S. F. Berkovic, Shared and distinct ultra-rare genetic risk for diverse epilepsies: A whole-exome sequencing study of 54,423 individuals across multiple genetic ancestries. medRxiv, doi: 10.1101/2023.02.22.23286310 (2023).

12. E. V. Minikel, K. J. Karczewski, H. C. Martin, B. B. Cummings, N. Whiffin, D. Rhodes, J. Alföldi, R. C. Trembath, D. A. van Heel, M. J. Daly, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, S. L. Schreiber, D. G. MacArthur, Author Correction: Evaluating drug targets through human loss-of-function genetic variation. Nature 590, E56 (2021).

13. E. V. Koonin, Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39, 309–338 (2005).

14. C. P. Austin, J. F. Battey, A. Bradley, M. Bucan, M. Capecchi, F. S. Collins, W. F. Dove, G. Duyk, S. Dymecki, J. T. Eppig, F. B. Grieder, N. Heintz, G. Hicks, T. R. Insel, A. Joyner, B. H. Koller, K. C. K. Lloyd, T. Magnuson, M. W. Moore, A. Nagy, J. D. Pollock, A. D. Roses, A. T. Sands, B. Seed, W. C. Skarnes, J. Snoddy, P. Soriano, D. J. Stewart, F. Stewart, B. Stillman, H. Varmus, L. Varticovski, I. M. Verma, T. F. Vogt, H. von Melchner, J. Witkowski, R. P. Woychik, W. Wurst, G. D. Yancopoulos, S. G. Young, B. Zambrowicz, The knockout mouse project. Nat. Genet. 36, 921–924 (2004).

15. E. A. Susaki, H. Ukai, H. R. Ueda, Next-generation mammalian genetics toward organism-level systems biology. npj Systems Biology and Applications 3, 1–11 (2017).

16. K. E. Jones, K. Safi, Ecology and evolution of mammalian biodiversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. 366, 2451–2461 (2011).

17. M. J. Christmas, I. M. Kaplow, D. P. Genereux, M. X. Dong, G. M. Hughes, X. Li, P. F. Sullivan, A. G. Hindle, G. Andrews, J. C. Armstrong, M. Bianchi, A. M. Breit, M. Diekhans, C. Fanter, N. M. Foley, D. B. Goodman, L. Goodman,

K. C. Keough, B. Kirilenko, A. Kowalczyk, C. Lawless, A. L. Lind, J. R. S. Meadows, L. R. Moreira, R. W. Redlich, L. Ryan, R. Swofford, A. Valenzuela, F. Wagner, O. Wallerman, A. R. Brown, J. Damas, K. Fan, J. Gatesy, J. Grimshaw, J. Johnson, S. V. Kozyrev, A. J. Lawler, V. D. Marinescu, K. M. Morrill, A. Osmanski, N. S. Paulat, B. N. Phan, S. K. Reilly, D. E. Schäffer, C. Steiner, M. A. Supple, A. P. Wilder, M. E. Wirthlin, J. R. Xue, Zoonomia Consortium§, B. W. Birren, S. Gazal, R. M. Hubley, K.-P. Koepfli, T. Marques-Bonet, W. K. Meyer, M. Nweeia, P. C. Sabeti, B. Shapiro, A. F. A. Smit, M. S. Springer, E. C. Teeling, Z. Weng, M. Hiller, D. L. Levesque, H. A. Lewin, W. J. Murphy, A. Navarro, B. Paten, K. S. Pollard, D. A. Ray, I. Ruf, O. A. Ryder, A. R. Pfenning, K. Lindblad-Toh, E. K. Karlsson, Evolutionary constraint and innovation across hundreds of placental mammals. Science 380, eabn3943 (2023).

18. N. L. Nehrt, W. T. Clark, P. Radivojac, M. W. Hahn, Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput. Biol. 7, e1002073 (2011).

19. M. E. Peterson, F. Chen, J. G. Saven, D. S. Roos, P. C. Babbitt, A. Sali, Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci. 18, 1306–1315 (2009).

20. B. M. Kirilenko, C. Munegowda, E. Osipova, D. Jebb, V. Sharma, M. Blumer, A. E. Morales, A.-W. Ahmed, D.-G. Kontopoulos, L. Hilgers, K. Lindblad-Toh, E. K. Karlsson, Zoonomia Consortium‡, M. Hiller, Integrating gene annotation with orthology inference at scale. Science 380, eabn3107 (2023).

21. M. Babadi, J. M. Fu, S. K. Lee, A. N. Smirnov, L. D. Gauthier, M. Walker, D. I. Benjamin, X. Zhao, K. J. Karczewski, I. Wong, R. L. Collins, A. Sanchis-Juan, H. Brand, E. Banks, M. E. Talkowski, GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. Nat. Genet. 55, 1589–1597 (2023).

22. R. L. Collins, J. T. Glessner, E. Porcu, M. Lepamets, R. Brandon, C. Lauricella, L. Han, T. Morley, L.-M. Niestroj, J. Ulirsch, S. Everett, D. P. Howrigan, P. M. Boone, J. Fu, K. J. Karczewski, G. Kellaris, C. Lowther, D. Lucente, K. Mohajeri, M. Nõukas, X. Nuttle, K. E. Samocha, M. Trinh, F. Ullah, U. Võsa, Epi25 Consortium, Estonian Biobank Research Team, M. E. Hurles, S. Aradhya, E. E. Davis, H. Finucane, J. F. Gusella, A. Janze, N. Katsanis, L. Matyakhina, B. M. Neale, D. Sanders, S. Warren, J. C. Hodge, D. Lal, D. M. Ruderfer, J. Meck, R. Mägi, T. Esko, A. Reymond, Z. Kutalik, H. Hakonarson, S. Sunyaev, H. Brand, M. E. Talkowski, A cross-disorder dosage sensitivity map of the human genome. Cell 185, 3041–3055.e25 (2022).

23. Z. L. Fuller, J. J. Berg, H. Mostafavi, G. Sella, M. Przeworski, Measuring intolerance to mutation in human genetics. Nat. Genet. 51, 772–776 (2019).

24. A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress, P. Flicek, GENCODE 2021. Nucleic Acids Res. 49, D916–D923 (2021).

25. V. Ranwez, E. J. P. Douzery, C. Cambon, N. Chantret, F. Delsuc, MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol. Biol. Evol. 35, 2582–2584 (2018).

26. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, V. A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33, D514–7 (2005).

27. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381, eadg7492 (2023).

28. H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, S. Ripke, F. R. Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, S. Purcell, E. Stahl, S. Lindstrom, J. R. B. Perry, Y. Okada, S. Raychaudhuri, M. J. Daly, N. Patterson, B. M. Neale, A. L. Price, Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).

29. H. K. Finucane, Y. A. Reshef, V. Anttila, K. Slowikowski, A. Gusev, A. Byrnes, S. Gazal, P.-R. Loh, C. Lareau, N. Shoresh, G. Genovese, A. Saunders, E. Macosko, S. Pollack, Brainstorm Consortium, J. R. B. Perry, J. D. Buenrostro, B. E. Bernstein, S. Raychaudhuri, S. McCarroll, B. M. Neale, A. L. Price, Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. 50, 621–629 (2018).

30. B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, B. M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295 (2015).

31. J. Grove, S. Ripke, T. D. Als, M. Mattheisen, R. K. Walters, H. Won, J. Pallesen, E. Agerbo, O. A. Andreassen, R. Anney, S. Awashti, R. Belliveau, F. Bettella, J. Buxbaum, J. Bybjerg-Grauholm, M. Bækvad-Hansen, F. Cerrato, K. Chambert, J. H. Christensen, C. Churchhouse, K. Dellenvall, D. Demontis, S. De Rubeis, B. Devlin, S. Djurovic, A. L. Dumont, J. I. Goldstein, C. S. Hansen, M. E. Hauberg, M. V. Hollegaard, S. Hope, D. P. Howrigan, H. Huang, C. M. Hultman, L. Klei, J. Maller, J. Martin, A. R. Martin, J. L. Moran, M. Nyegaard, T. Nærland, D. S. Palmer, A. Palotie, C. B. Pedersen, M. G. Pedersen, T. dPoterba, J. B. Poulsen, B. S. Pourcain, P. Qvist, K. Rehnström, A. Reichenberg, J. Reichert, E. B. Robinson, K. Roeder, P. Roussos, E. Saemundsen, S. Sandin, F. K. Satterstrom, G. Davey Smith, H. Stefansson, S. Steinberg, C. R. Stevens, P. F. Sullivan, P. Turley, G. B. Walters, X. Xu, K. Stefansson, D. H. Geschwind, M. Nordentoft, D. M. Hougaard, T. Werge, O. Mors, P. B. Mortensen, B. M. Neale, M. J. Daly, A. D. Børglum, Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. 51, 431–444 (2019).

32. N. Mullins, A. J. Forstner, K. S. O'Connell, B. Coombes, J. R. I. Coleman, Z. Qiao, T. D. Als, T. B. Bigdeli, S. Børte, J. Bryois, A. W. Charney, O. K. Drange, M. J. Gandal, S. P. Hagenaars, M. Ikeda, N. Kamitaki, M. Kim, K. Krebs, G. Panagiotaropoulou, B. M. Schilder, L. G. Sloofman, S. Steinberg, V. Trubetskoy, B. S. Winsvold, H.-H. Won, L. Abramova, K. Adorjan, E. Agerbo, M. Al Eissa, D. Albani, N. Alliey-Rodriguez, A. Anjorin, V. Antilla, A. Antoniou, S. Awasthi, J. H. Baek, M. Bækvad-Hansen, N. Bass, M. Bauer, E. C. Beins, S. E. Bergen, A. Birner, C. Bøcker Pedersen, E. Bøen, M. P. Boks, R. Bosch, M. Brum, B. M. Brumpton, N. Brunkhorst-Kanaan, M. Budde, J. Bybjerg-Grauholm, W. Byerley, M. Cairns, M. Casas, P. Cervantes, T.-K. Clarke, C. Cruceanu, A. Cuellar-Barboza, J. Cunningham, D. Curtis, P. M. Czerski, A. M. Dale, N. Dalkner, F. S. David, F. Degenhardt, S. Djurovic, A. L. Dobbyn, A. Douzenis, T. Elvsåshagen, V. Escott-Price, I. N. Ferrier, A. Fiorentino, T. M. Foroud, L. Forty, J. Frank, O. Frei, N. B. Freimer, L. Frisén, K. Gade, J. Garnham, J. Gelernter, M. Giørtz Pedersen, I. R. Gizer, S. D. Gordon, K. Gordon-Smith, T. A. Greenwood, J. Grove, J. Guzman-Parra, K. Ha, M. Haraldsson, M. Hautzinger, U. Heilbronner, D. Hellgren, S. Herms, P. Hoffmann, P. A. Holmans, L. Huckins, S. Jamain, J. S. Johnson, J. L. Kalman, Y. Kamatani, J. L. Kennedy, S. Kittel-Schneider, J. A. Knowles, M. Kogevinas, M. Koromina, T. M. Kranz, H. R. Kranzler, M. Kubo, R. Kupka, S. A. Kushner, C. Lavebratt, J. Lawrence, M. Leber, H.-J. Lee, P. H. Lee, S. E. Levy, C. Lewis, C. Liao, S. Lucae, M. Lundberg, D. J. MacIntyre, S. H. Magnusson, W. Maier, A. Maihofer, D. Malaspina, E. Maratou, L. Martinsson, M. Mattheisen, S. A. McCarroll, N. W. McGregor, P. McGuffin, J. D. McKay, H. Medeiros, S. E. Medland, V. Millischer, G. W. Montgomery, J. L. Moran, D. W. Morris, T. W. Mühleisen, N. O'Brien, C. O'Donovan, L. M. Olde Loohuis, L. Oruc, S. Papiol, A. F. Pardiñas, A. Perry, A. Pfennig, E. Porichi, J. B. Potash, D. Quested, T. Raj, M. H. Rapaport, J. R. DePaulo, E. J. Regeer, J. P. Rice, F. Rivas, M. Rivera, J. Roth, P. Roussos, D. M. Ruderfer, C. Sánchez-Mora, E. C. Schulte, F. Senner, S. Sharp, P. D. Shilling, E. Sigurdsson, L. Sirignano, C. Slaney, O. B. Smeland, D. J. Smith, J. L. Sobell, C. Søholm Hansen, M. Soler Artigas, A. T. Spijker, D. J. Stein, J. S. Strauss, B. Świątkowska, C. Terao, T. E. Thorgeirsson, C. Toma, P. Tooney, E.-E. Tsermpini, M. P. Vawter, H. Vedder, J. T. R. Walters, S. H. Witt, S. Xi, W. Xu, J. M. K. Yang, A. H. Young, H. Young, P. P. Zandi, H. Zhou, L. Zillich, R. Adolfsson, I. Agartz, M. Alda, L. Alfredsson, G. Babadjanova, L. Backlund, B. T. Baune, F. Bellivier, S. Bengesser, W. H. Berrettini, D. H. R. Blackwood, M. Boehnke, A. D. Børglum, G. Breen, V. J. Carr, S. Catts, A. Corvin, N. Craddock, U. Dannlowski, D. Dikeos, T. Esko, B. Etain, P. Ferentinos, M. Frye, J. M. Fullerton, M. Gawlik, E. S. Gershon, F. S. Goes, M. J. Green, M. Grigoroiu-Serbanescu, J. Hauser, F. Henskens, J. Hillert, K. S. Hong, D. M. Hougaard, C. M. Hultman, K. Hveem, N. Iwata, A. V. Jablensky, I. Jones, L. A. Jones, R. S. Kahn, J. R. Kelsoe, G. Kirov, M. Landén, M. Leboyer, C. M. Lewis, Q. S. Li, J. Lissowska, C. Lochner, C. Loughland, N. G. Martin, C. A. Mathews, F. Mayoral, S. L. McElroy, A. M. McIntosh, F. J. McMahon, I. Melle, P. Michie, L. Milani, P. B. Mitchell, G. Morken, O. Mors, P. B. Mortensen, B. Mowry, B. Müller-Myhsok, R. M. Myers, B. M. Neale, C. M. Nievergelt, M. Nordentoft, M. M. Nöthen, M. C. O'Donovan, K. J. Oedegaard, T. Olsson, M. J. Owen, S. A. Paciga, C. Pantelis, C. Pato, M. T. Pato, G. P. Patrinos, R. H. Perlis, J. A. Ramos-Quiroga, A. Reif, E. Z. Reininghaus, M. Ribasés, M. Rietschel, S. Ripke, G. A. Rouleau, T. Saito, U. Schall, M. Schalling, P. R. Schofield, T. G. Schulze, L. J. Scott, R. J. Scott, A.

Serretti, C. Shannon Weickert, J. W. Smoller, H. Stefansson, K. Stefansson, E. Stordal, F. Streit, P. F. Sullivan, G. Turecki, A. E. Vaaler, E. Vieta, J. B. Vincent, I. D. Waldman, T. W. Weickert, T. Werge, N. R. Wray, J.-A. Zwart, J. M. Biernacka, J. I. Nurnberger, S. Cichon, H. J. Edenberg, E. A. Stahl, A. McQuillin, A. Di Florio, R. A. Ophoff, O. A. Andreassen, Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat. Genet. 53, 817–829 (2021).

33. D. Demontis, G. B. Walters, G. Athanasiadis, R. Walters, K. Therrien, T. T. Nielsen, L. Farajzadeh, G. Voloudakis, J. Bendl, B. Zeng, W. Zhang, J. Grove, T. D. Als, J. Duan, F. K. Satterstrom, J. Bybjerg-Grauholm, M. Bækved-Hansen, O. O. Gudmundsson, S. H. Magnusson, G. Baldursson, K. Davidsdottir, G. S. Haraldsdottir, E. Agerbo, G. E. Hoffman, S. Dalsgaard, J. Martin, M. Ribasés, D. I. Boomsma, M. Soler Artigas, N. Roth Mota, D. Howrigan, S. E. Medland, T. Zayats, V. M. Rajagopal, M. Nordentoft, O. Mors, D. M. Hougaard, P. B. Mortensen, M. J. Daly, S. V. Faraone, H. Stefansson, P. Roussos, B. Franke, T. Werge, B. M. Neale, K. Stefansson, A. D. Børglum, Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. Nat. Genet. 55, 198–208 (2023).

34. C. Tcheandjieu, X. Zhu, A. T. Hilliard, S. L. Clarke, V. Napolioni, S. Ma, K. M. Lee, H. Fang, F. Chen, Y. Lu, N. L. Tsao, S. Raghavan, S. Koyama, B. R. Gorman, M. Vujkovic, D. Klarin, M. G. Levin, N. Sinnott-Armstrong, G. L. Wojcik, M. E. Plomondon, T. M. Maddox, S. W. Waldo, A. G. Bick, S. Pyarajan, J. Huang, R. Song, Y.-L. Ho, S. Buyske, C. Kooperberg, J. Haessler, R. J. F. Loos, R. Do, M. Verbanck, K. Chaudhary, K. E. North, C. L. Avery, M. Graff, C. A. Haiman, L. Le Marchand, L. R. Wilkens, J. C. Bis, H. Leonard, B. Shen, L. A. Lange, A. Giri, O. Dikilitas, I. J. Kullo, I. B. Stanaway, G. P. Jarvik, A. S. Gordon, S. Hebbring, B. Namjou, K. M. Kaufman, K. Ito, K. Ishigaki, Y. Kamatani, S. S. Verma, M. D. Ritchie, R. L. Kember, A. Baras, L. A. Lotta, S. Kathiresan, E. R. Hauser, D. R. Miller, J. S. Lee, D. Saleheen, P. D. Reaven, K. Cho, J. M. Gaziano, P. Natarajan, J. E. Huffman, B. F. Voight, D. J. Rader, K.-M. Chang, J. A. Lynch, S. M. Damrauer, P. W. F. Wilson, H. Tang, Y. V. Sun, P. S. Tsao, C. J. O'Donnell, T. L. Assimes, Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. Nat. Med. 28, 1679–1692 (2022).

35. V. Trubetskoy, A. F. Pardiñas, T. Qi, G. Panagiotaropoulou, S. Awasthi, T. B. Bigdeli, J. Bryois, C.-Y. Chen, C. A. Dennison, L. S. Hall, M. Lam, K. Watanabe, O. Frei, T. Ge, J. C. Harwood, F. Koopmans, S. Magnusson, A. L. Richards, J. Sidorenko, Y. Wu, J. Zeng, J. Grove, M. Kim, Z. Li, G. Voloudakis, W. Zhang, M. Adams, I. Agartz, E. G. Atkinson, E. Agerbo, M. Al Eissa, M. Albus, M. Alexander, B. Z. Alizadeh, K. Alptekin, T. D. Als, F. Amin, V. Arolt, M. Arrojo, L. Athanasiu, M. H. Azevedo, S. A. Bacanu, N. J. Bass, M. Begemann, R. A. Belliveau, J. Bene, B. Benyamin, S. E. Bergen, G. Blasi, J. Bobes, S. Bonassi, A. Braun, R. A. Bressan, E. J. Bromet, R. Bruggeman, P. F. Buckley, R. L. Buckner, J. Bybjerg-Grauholm, W. Cahn, M. J. Cairns, M. E. Calkins, V. J. Carr, D. Castle, S. V. Catts, K. D. Chambert, R. C. K. Chan, B. Chaumette, W. Cheng, E. F. C. Cheung, S. A. Chong, D. Cohen, A. Consoli, Q. Cordeiro, J. Costas, C. Curtis, M. Davidson, K. L. Davis, L. de Haan, F. Degenhardt, L. E. DeLisi, D. Demontis, F. Dickerson, D. Dikeos, T. Dinan, S. Djurovic, J. Duan, G. Ducci, F. Dudbridge, J. G. Eriksson, L. Fañanás, S. V. Faraone, A. Fiorentino, A. Forstner, J. Frank, N. B. Freimer, M. Fromer, A. Frustaci, A. Gadelha, G. Genovese, E. S. Gershon, M. Giannitelli, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, J. González Peñas, A. González-Pinto, S. Gopal, J. Gratten, M. F. Green, T. A. Greenwood, O. Guillin, S. Gülöksüz, R. E. Gur, R. C. Gur, B. Gutiérrez, E. Hahn, H. Hakonarson, V. Haroutunian, A. M. Hartmann, C. Harvey, C. Hayward, F. A. Henskens, S. Herms, P. Hoffmann, D. P. Howrigan, M. Ikeda, C. Iyegbe, I. Joa, A. Julià, A. K. Kähler, T. Kam-Thong, Y. Kamatani, S. Karachanak-Yankova, O. Kebir, M. C. Keller, B. J. Kelly, A. Khrunin, S.-W. Kim, J. Klovins, N. Kondratiev, B. Konte, J. Kraft, M. Kubo, V. Kučinskas, Z. A. Kučinskiene, A. Kusumawardhani, H. Kuzelova-Ptackova, S. Landi, L. C. Lazzeroni, P. H. Lee, S. E. Legge, D. S. Lehrer, R. Lencer, B. Lerer, M. Li, J. Lieberman, G. A. Light, S. Limborska, C.-M. Liu, J. Lönnqvist, C. M. Loughland, J. Lubinski, J. J. Luykx, A. Lynham, M. Macek, A. Mackinnon, P. K. E. Magnusson, B. S. Maher, W. Maier, D. Malaspina, J. Mallet, S. R. Marder, S. Marsal, A. R. Martin, L. Martorell, M. Mattheisen, R. W. McCarley, C. McDonald, J. J. McGrath, H. Medeiros, S. Meier, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, M. Mitjans, E. Molden, E. Molina, M. D. Molto, V. Mondelli, C. Moreno, C. P. Morley, G. Muntané, K. C. Murphy, I. Myin-Germeys, I. Nenadić, G. Nestadt, L. Nikitina-Zake, C. Noto, K. H. Nuechterlein, N. L. O'Brien, F. A. O'Neill, S.-Y. Oh, A. Olincy, V. K. Ota, C. Pantelis, G. N. Papadimitriou, M. Parellada, T. Paunio, R. Pellegrino, S. Periyasamy, D. O. Perkins, B. Pfuhlmann, O. Pietiläinen, J. Pimm, D. Porteous, J. Powell, D. Quattrone, D. Quested, A. D. Radant, A. Rampino, M. H. Rapaport, A. Rautanen, A. Reichenberg, C. Roe, J. L. Roffman, J. Roth, M. Rothermundt, B. P. F. Rutten, S. Saker-Delye, V. Salomaa, J. Sanjuan, M. L. Santoro, A. Savitz, U. Schall, R. J. Scott, L. J. Seidman, S. I. Sharp, J. Shi, L. J. Siever, E. Sigurdsson, K. Sim, N. Skarabis, P. Slominsky, H.-C. So, J. L. Sobell, E. Söderman, H. J. Stain, N. E. Steen, A. A. Steixner-Kumar, E. Stögmann, W. S. Stone, R. E. Straub, F. Streit, E. Strengman, T. S. Stroup, M. Subramaniam, C. A. Sugar, J. Suvisaari, D. M. Svrakic, N. R. Swerdlow, J. P. Szatkiewicz, T. M. T. Ta, A. Takahashi, C. Terao, F. Thibaut, D. Toncheva, P. A. Tooney, S. Torretta, S. Tosato, G. B. Tura, B. I. Turetsky, A. Üçok, A. Vaaler, T. van Amelsvoort, R. van Winkel, J. Veijola, J. Waddington, H. Walter, A. Waterreus, B. T. Webb, M. Weiser, N. M. Williams, S. H. Witt, B. K. Wormley, J. Q. Wu, Z. Xu, R. Yolken, C. C. Zai, W. Zhou, F. Zhu, F. Zimprich, E. C. Atbaşoğlu, M. Ayub, C. Benner, A. Bertolino, D. W. Black, N. J. Bray, G. Breen, N. G. Buccola, W. F. Byerley, W. J. Chen, C. R. Cloninger, B. Crespo-Facorro, G. Donohoe, R. Freedman, C. Galletly, M. J. Gandal, M. Gennarelli, D. M. Hougaard, H.-G. Hwu, A. V. Jablensky, S. A. McCarroll, J. L. Moran, O. Mors, P. B. Mortensen, B. Müller-Myhsok, A. L. Neil, M. Nordentoft, M. T. Pato, T. L. Petryshen, M. Pirinen, A. E. Pulver, T. G. Schulze, J. M. Silverman, J. W. Smoller, E. A. Stahl, D. W. Tsuang, E. Vilella, S.-H. Wang, S. Xu, R. Adolfsson, C. Arango, B. T. Baune, S. I. Belangero, A. D. Børglum, D. Braff, E. Bramon, J. Buxbaum, D. Campion, J. A. Cervilla, S. Cichon, D. A. Collier, A. Corvin, D. Curtis, M. D. Forti, E. Domenici, H. Ehrenreich, V. Escott-Price, T. Esko, A. H. Fanous, A. Gareeva, M. Gawlik, P. V. Gejman, M. Gill, S. J. Glatt, V. Golimbet, K. S. Hong, C. M. Hultman, S. E. Hyman, N. Iwata, E. G. Jönsson, R. S. Kahn, J. L. Kennedy, E. Khusnutdinova, G. Kirov, J. A. Knowles, M.-O. Krebs, C. Laurent-Levinson, J. Lee, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, D. Malhotra, A. McIntosh, A. McQuillin, P. R. Menezes, V. A. Morgan, D. W. Morris, B. J. Mowry, R. M. Murray, V. Nimgaonkar, M. M. Nöthen, R. A. Ophoff, S. A. Paciga, A. Palotie, C. N. Pato, S. Qin, M. Rietschel, B. P. Riley, M. Rivera, D. Rujescu, M. C. Saka, A. R. Sanders, S. G. Schwab, A. Serretti, P. C. Sham, Y. Shi, D. St Clair, H. Stefánsson, K. Stefansson, M. T. Tsuang, J. van Os, M. P. Vawter, D. R. Weinberger, T. Werge, D. B. Wildenauer, X. Yu, W. Yue, P. A. Holmans, A. J. Pocklington, P. Roussos, E. Vassos, M. Verhage, P. M. Visscher, J. Yang, O. A. Andreassen, K. S. Kendler, M. J. Owen, N. R. Wray, M. J. Daly, H. Huang, B. M. Neale, P. F. Sullivan, S. Ripke, J. T. R. Walters, M. C. O'Donovan, Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604, 502–508 (2022).