The landscape of regional missense mutational intolerance quantified from 125,748 exomes

Katherine R. Chao[1,2]*, Lily Wang[1,2,3]*, Ruchit Panchal[1,2], Calwing Liao[4,5,6], Haneen Abderrazzaq[7], Robert Ye[4,5], Patrick Schultz[1,4,5], John Compitello[1,4,5], Riley H. Grant[1], Jack A. Kosmicki[3,4,5], Ben Weisburd[1,2], William Phu[1,2], Michael W. Wilson[1,2], Kristen M. Laricchia[1,2], Julia K. Goodrich[1,2], Daniel Goldstein[1,4,5], Jacqueline I. Goldstein[1,4,5], Christopher Vittal[1,4,5], Timothy Poterba[1,4,5], Samantha Baxter[1], Nicholas A. Watts[1,4], Matthew Solomonson[1,4], gnomAD Consortium, Grace Tiao[1,2], Heidi L. Rehm[1,2], Benjamin M. Neale[1,4], Michael E. Talkowski[1,2,5], Daniel G. MacArthur[1,8,9], Anne O'Donnell-Luria[1,2,10], Konrad J. Karczewski[1,4,5], Predrag Radivojac[7], Mark J. Daly[1,4,11], Kaitlin E. Samocha[1,2,4]

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[3]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA, USA
[4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Department of Medicine, Harvard Medical School, Boston, MA, USA
[7]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[8]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, New South Wales, Australia
[9]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia
[10]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
[11]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland

*Indicates equal contribution

Correspondence should be addressed to K.E.S. (samocha@broadinstitute.org)

Abstract

Missense variants can have a range of functional impacts depending on factors such as the specific amino acid substitution and location within the gene. To interpret their deleteriousness, studies have sought to identify regions within genes that are specifically intolerant of missense variation[1–12]. Here, we leverage the patterns of rare missense variation in 125,748 individuals in the Genome Aggregation Database (gnomAD)[13] against a null mutational model to identify transcripts that display regional differences in missense constraint. Missense-depleted regions are enriched for ClinVar[14] pathogenic variants, *de novo* missense variants from individuals with neurodevelopmental disorders (NDDs)[15,16], and complex trait heritability. Following ClinGen calibration recommendations for the ACMG/AMP guidelines, we establish that regions with less than 20% of their expected missense variation achieve moderate support for pathogenicity. We create a missense deleteriousness metric (MPC) that incorporates regional constraint and outperforms other deleteriousness scores at stratifying case and control *de novo* missense variation, with a strong enrichment in NDDs. These results provide additional tools to aid in missense variant interpretation.

Main text

Over the last decade, exome and genome sequencing have enabled variant discovery across hundreds of thousands of individuals[13,17–21]. These large reference databases have provided the opportunity to study selective forces acting on the human genome and to identify genomic regions under selective constraint by, for example, identifying regions with fewer variants than expected based on mutational models[13,18,22–25]. Gene-level metrics of predicted loss-of-function (pLoF) variant depletion have proven to be valuable in variant classification and identification of novel disease genes[15,16,26–28]. The functional impact and selective pressures relevant to missense variation, by contrast, remain challenging to predict, as the effect of a missense variant is governed by the gene housing the variant, the position of the variant in the gene, and the specific amino acid substitution caused by the variant. To address this, prior work has sought to identify regions within coding genes that are specifically intolerant of missense variation as a way to improve interpretation[1–12]. Here, we expand upon previous work[1] and show a sub-genic measure of missense intolerance leveraging population-level variation facilitates variant classification and risk stratification for association studies with *de novo*, rare, and common variants.

We explored the patterns of rare missense variant presence or absence in 125,748 exomes in the Genome Aggregation Database (gnomAD) v2.1.1 on GRCh37 to quantify missense depletion at the sub-genic level. We searched 18,629 canonical protein-coding transcripts for variability in missense constraint, quantified as the number of rare (allele frequency [AF] < 0.1%) missense variants observed in gnomAD divided by the number expected under neutral evolution as estimated from previously described mutational models[13](observed/expected [OE]). For each transcript, we applied a recursive search based on likelihood ratio tests over all potential rare missense sites looking for breaks that divide the transcript coding sequence (CDS) into distinct missense constraint regions (MCRs; **Fig. 1a, b**). We discover 5,127 transcripts (28%) harbor regional variability in missense constraint (**Fig. 1c**), i.e., have two or more MCRs (minimum coding length 49bp, median 461bp; **Supplementary Fig. 1**). We thus refine the resolution of missense constraint for 42% of coding sites (coding space in the 5,127 transcripts vs. 18,629 total assessed). After recalibrating the missense OE distribution over all potential sites of missense variants using MCR-wide rather than transcript-wide missense OE measurements, we discover widespread signatures of negative and neutral selection that are obscured when quantifying over the unit of whole transcripts (**Fig. 1d**). We find a larger proportion of the exome lies within strongly constrained sequences (5.6% vs. 1.7% at OE < 0.4;

see **Supplementary Note** for OE threshold selection), and the mode of the distribution shifts toward an OE indicative of evolutionary neutrality at approximately 1 (40.6% vs. 36.5% at 0.9 < OE ≤ 1.1).

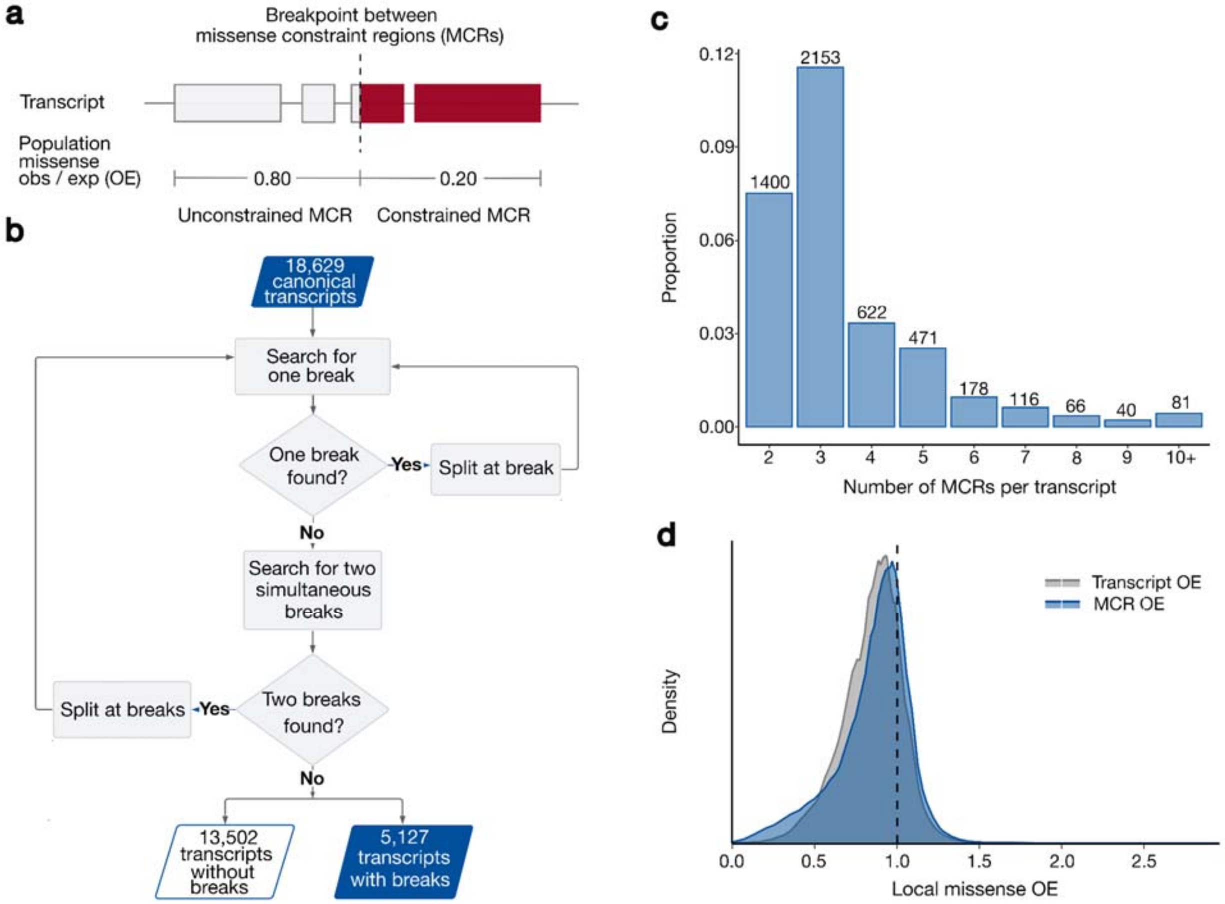

Fig. 1: 28% of protein-coding genes in the human genome are discovered to harbor regional variation in population-level missense depletion.
a, An example transcript that has two missense constraint regions (MCRs) with significantly different levels of population-wide missense depletion, defined as the number of missense variants observed in gnomAD at rare frequency (AF < 0.1%) divided by the number of rare missense variants expected under neutral evolution (observed/expected or OE). Lower OE values correspond to greater variant depletion in the population and suggest stronger constraint. **b**, Flow chart describing the process of searching for breakpoints that divide a transcript into multiple MCRs. Searching for breakpoints is recursive and leverages likelihood ratio tests at a significance threshold of p = 0.001. **c**, The number of MCRs within the 5,127 transcripts discovered to harbor regional differences in missense constraint. The other 13,502 transcripts are deemed to have a single MCR (that is, a constant level of constraint across their entirety) and are not shown. **d**, The distribution of local missense OE at all coding sites in canonical transcripts. Local missense OE is defined as the OE calculated over the whole transcript (for

"transcript OE") or over the MCR (for "MCR OE") where the site is located. Transcript OE and MCR OE are equivalent for transcripts with one MCR.

Furthermore, we find that constrained MCRs overlap established disease-associated mutational hotspots, including critical protein domains. One example is in the well-characterized *KCNQ1*, a voltage-gated potassium channel gene, in which pathogenic variants cause cardiac disorders such as long QT syndrome. We discover one moderately constrained MCR (missense OE = 0.60) overlapping the highly conserved C-terminus [29] and another (missense OE = 0.66) encompassing the voltage-sensing and pore domains (**Fig. 2a**). Both the C-terminus of *KCNQ1* and its voltage-sensing domain are established "hotspot" regions (specific missense-constrained regions with ACMG/AMP hotspot/functional domain moderate support [PM1] for pathogenicity)[29–31]. All but two ClinVar pathogenic/likely pathogenic (P/LP) missense variants in this gene fall within these two missense-constrained MCRs.

We also find that missense constraint within MCRs is able to identify regions associated with severe, early-onset disease. One example is in *BAP1*, which plays a key role in chromatin modeling by mediating histone deubiquitination. Disease-causing variants in this gene are linked to cancer or, as recently discovered, Kury-Isidor syndrome[32]. The first highly missense-constrained MCR (missense OE = 0.33) in *BAP1* encompasses the ubiquitin C-terminal hydrolase domain connected to Kury-Isidor syndrome[32] (**Fig. 2b**), and all 11 variants reported to be causal for Kury-Isidor fall within this MCR. The only ClinVar P/LP variants that do not fall within any missense-constrained MCRs in *BAP1* are associated with cancer phenotypes, which may be under weaker selection than neurodevelopmental disorders (NDDs).

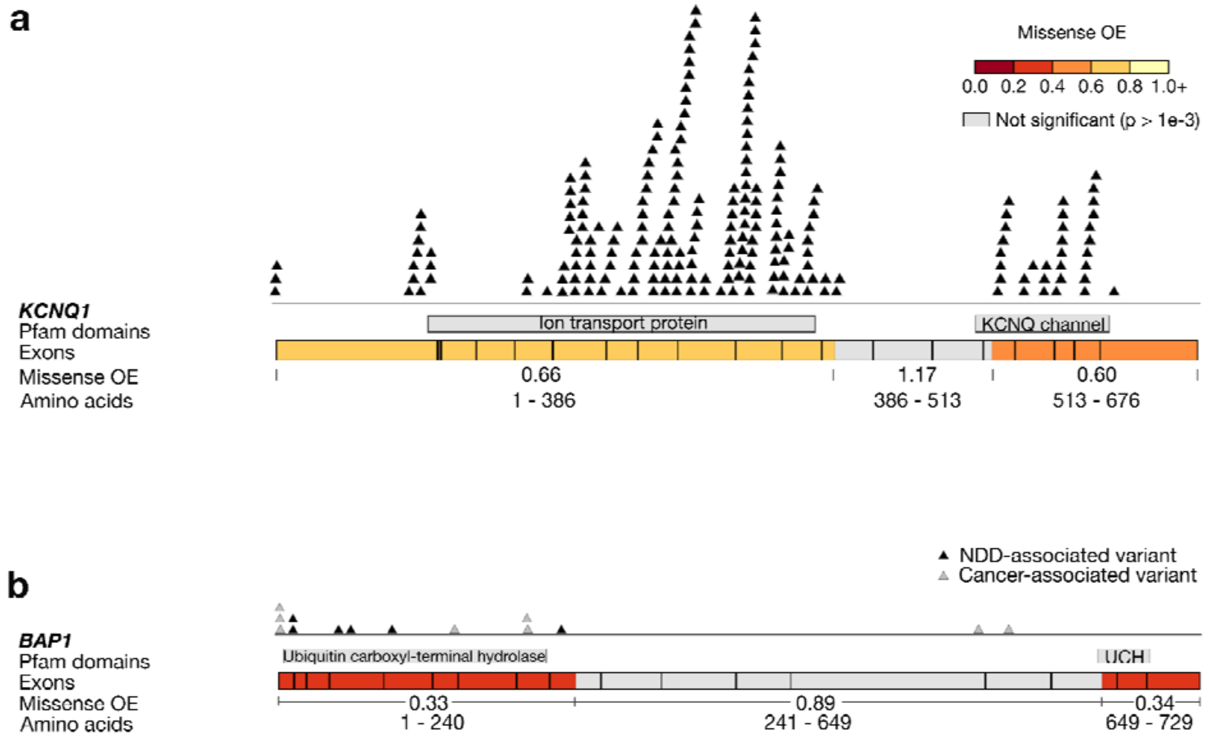

Fig. 2: Missense constraint regions (MCRs) and the distribution of ClinVar pathogenic/likely pathogenic (P/LP) missense variants in two genes associated with early-onset developmental disorders. Exons are delineated with black outlines and MCRs are delineated by color. MCRs are colored based on their missense observed/expected (OE) ratio, and MCRs with missense OEs not significantly different from 1 (p > 0.001) are shaded gray. **a,** *KCNQ1*. Only two of the 210 P/LP missense variants in *KCNQ1* do not fall within either constrained MCR. The first constrained MCR encompasses the voltage-sensing and pore domains of this gene, and the other constrained MCR overlaps the C-terminus. Both domains contain previously reported hotspot regions, with some regions reaching moderate level (PM1) support for pathogenicity[31]. Ion transport protein: domain that contains both the transmembrane voltage-sensing and pore domains. KCNQ channel: C-terminal cytoplasmic domain that overlaps four helices (A-D). **b,** *BAP1*. Variants in this gene can lead to cancer-predisposition syndromes, increased risk of certain cancers, or the neurodevelopmental disorder Kury-Isidor syndrome[32]. All of the ClinVar P/LP variants associated with Kury-Isidor fall within the first MCR with a highly depleted missense OE of 0.33. An additional five variants reported in Kury *et al*.[32] but not ClinVar fall within either highly constrained MCR in this gene. P/LP variants associated with Kury-Isidor are colored in black, and all other cancer-associated P/LP variants are colored in gray. UCH: Ubiquitin carboxyl-terminal hydrolase isozyme L5 domain. ClinVar data are from the October 15, 2023 release.

Next, we sought to determine if the signatures of selection revealed by MCRs recapitulated biological and disease relevance of coding sequences. Overall, most transcripts that are intolerant to pLoF variation (as measured by the loss-of-function observed/expected upper bound fraction [LOEUF] score[13]) also tend to be intolerant to missense variation. This trend is markedly more prominent when measuring missense constraint at the sub-genic level vs. the transcript-level (**Supplementary Note; Supplementary Fig. 2**). We also discovered that 64% (1697/2659) of genes that are both LOEUF- and MCR missense-constrained do not have disease associations in OMIM[33], suggesting the existence of many undocumented genes containing variants of significant consequence for disease (**Supplementary Fig. 3**). In a set of 730 strongly mutationally intolerant genes, defined here as exhibiting both population depletion of pLoF variants (first three LOEUF deciles) and association with a developmental phenotype (high-confidence membership in any non-cancer Gene2Phenotype [G2P][34] gene list with dominant inheritance), we observed strong transcript-wide missense depletion that was even stronger for genes with multiple MCRs (**Fig. 3a** and **Supplementary Fig. 4**; Wilcoxon $p < 10^{-50}$). Given that we have greater power to detect missense constraint variability over longer sequences (**Supplementary Fig. 5**), we controlled for transcript length but still found that intolerant transcripts are eight times more likely to harbor multiple MCRs ($p < 10^{-50}$). These strongly intolerant transcripts are highly enriched for severely depleted regions (three times more likely to have minimum MCR OE < 0.4 after regressing out transcript length, $p < 10^{-18}$), whereas the most constrained MCRs in not strongly intolerant transcripts are less depleted and more evenly distributed across the OE spectrum. Finally, we observe a group of genes with strong overall missense depletion in which we did not detect multiple MCRs (n = 459 with missense OE < 0.4; **Supplementary Table 1**), suggesting these genes are robustly intolerant to missense variants across their length. When comparing missense constraint to selection over longer timescales (measured by evolutionary conservation in placental mammals, phyloP[35]), we found that genes with more conserved coding sequences also tended to be more overall depleted of human missense variation (Spearman $\rho = 0.56$, $p < 10^{-50}$). However, a substantial number of strongly constrained MCRs appear widely unconserved across mammals, potentially pointing to human-specific negative selection pressures that are obscured at the whole-transcript level (**Supplementary Fig. 6**).

We next aggregated *de novo* missense variants from 31,058 individuals with a severe developmental disorder[15] (DD), 15,036 autistic individuals (AUT), and 5,492 siblings not diagnosed with a DD[16] (**Fig. 3b**). The distribution of *de novo* missense variants across the missense OE spectrum in unaffected siblings largely mirrored the exome-wide missense OE distribution. In contrast, *de novo* missense variants in autistic individuals are enriched in missense-constrained sequences, and this pattern is more striking in individuals with DDs. For example, relative to unaffected siblings, the rate of *de novo* missense variants in MCRs with OE < 0.2 is 2-fold higher in autistic individuals ($p < 10^{-23}$) and 6.6-fold higher in individuals with DDs ($p < 10^{-50}$) (**Supplementary Fig. 7;** see **Supplementary Note** for OE threshold selection). This is consistent with the expectations that a small subset of *de novo* missense variants in individuals with developmental phenotypes are causal for those traits and that variants causal for DD are generally more selectively deleterious than those for autism.

Beyond large-effect rare and *de novo* variation in traits under strong negative selection, we additionally investigated whether our MCR metric, which was calculated using rare variants, correlates with functional effects of common variants. Prior work found that pLoF-constrained genes and their flanking 100kb sequences are enriched for SNP heritability across hundreds of independent traits in the UK Biobank (UKBB) and other large genome-wide association studies (GWAS) [13]. We partitioned common (AF > 5%) variant heritability of the same 268 independent traits across MCRs to investigate relative enrichment. To establish a baseline, we computed the heritability enrichment over all coding sequences comprising MCRs (3-fold). The most constrained MCRs have the strongest heritability enrichment; the first quintile of MCR missense OE harbors a 41-fold enrichment (**Fig. 3c**). Coding SNPs in missense-unconstrained MCRs (e.g., in the two least constrained quintiles of MCR missense OE) harbor no detectable heritability enrichment relative to the average genome-wide SNP. These findings suggest that: 1) regions depleted of rare missense variation can help prioritize common coding variants important for complex traits (i.e., improve GWAS fine-mapping variant prioritization), and 2) there exists a subset of coding sequence with no appreciable heritability enrichment, which rare variant depletion can help identify.

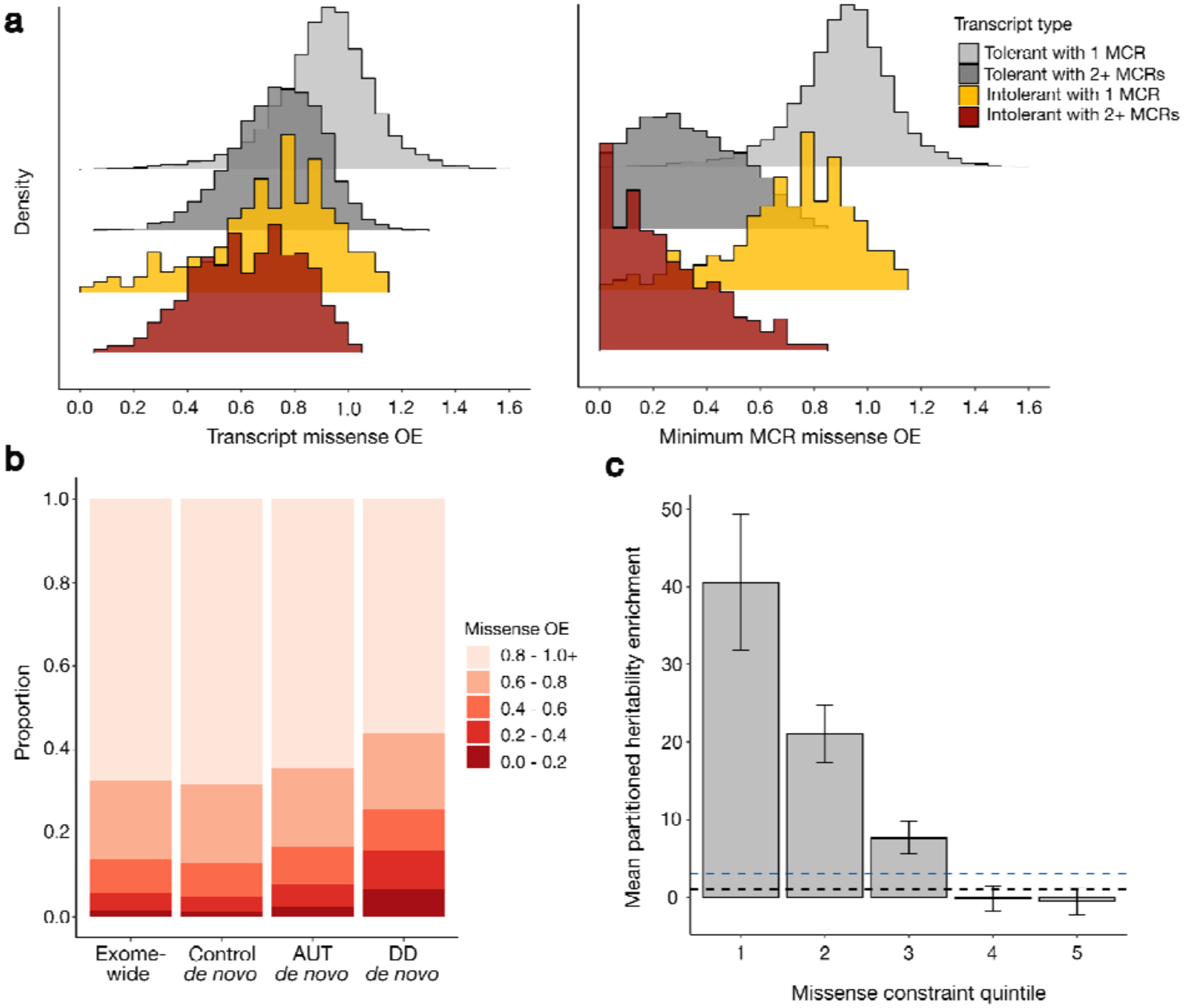

Fig. 3: Regional missense depletion reveals constraint obscured by gene-level measures. **a**, Left: The distribution of transcript-wide missense observed/expected (OE) across 18,629 transcripts stratified by the combination of two factors: whether the transcript is strongly mutationally intolerant (within first three LOEUF deciles and association with a developmental phenotype in Gene2Phenotype [G2P][34]) and whether we detect multiple missense constraint regions (MCRs). Number of transcripts in each category are: strongly intolerant with multiple MCRs (n=581; red), strongly intolerant with one MCR (n=149; yellow), not strongly intolerant with multiple MCRs (n=4,546; dark gray), not strongly intolerant with one MCR (n=13,353; light gray). X-axis is cut off at 1.6 for visibility. Right: Minimum MCR missense OE using the same groupings. Minimum MCR missense OE is the same as transcript missense OE for transcripts with a single MCR. **b**, MCR missense OE at all sites of possible exome-wide missense variants vs. sites of *de novo* missense variants in controls, autistic individuals (AUT), or individuals with DD. *De novo* variants from individuals with developmental phenotypes are enriched in more constrained sequences, with a more pronounced enrichment in DD than autism. **c**, Enrichment in per-variant heritability explained by common (AF > 5%) protein-coding SNPs stratified by MCR missense OE quintile, relative to the average SNP genome-wide. Enrichment is estimated by linkage disequilibrium score regression, accounting for number of SNPs in each quintile, and

is averaged across 268 independent traits in UKBB and other large genome-wide association studies. Black dashed line at 1 indicates no enrichment. Blue dashed line at 3 indicates average coding enrichment. Error bars represent 95% confidence intervals.

We examined the localization of high-quality ClinVar[36] missense variants classified as P/LP within genes with both unconstrained (missense OE > 0.9) and constrained (missense OE < 0.2) MCRs and found that P/LP variants occur much more frequently in missense constrained MCRs (odds ratio [OR] = 15.2; $p < 10^{-50}$). We also examined the localization of P/LP and benign/likely benign (B/LB) variants within MCRs in autosomal dominant disease-associated genes and found that P/LP variants tend to localize to regions that are more strongly missense-constrained than the overall transcript (Wilcoxon $p = 3.5 \times 10^{-10}$), while B/LB variants show the opposite effect and tend to occur in regions with OEs closer to 1 (Wilcoxon $p < 10^{-18}$; **Fig. 4a**). While more subtle, these same patterns are also significant in autosomal recessive disease-associated genes (**Supplementary Fig. 8**).

To enable use of our missense constraint metric in ACMG/AMP clinical variant classification, we applied previously established probabilistic frameworks[37] to determine the MCR missense OE thresholds that met different levels of clinical evidence strengths evaluated under the hotspot/functional domain (PM1) and benign *in silico* prediction (BP4) criteria codes[30]. MCR missense OE ≤ 0.37 met supporting (PM1_Supporting) and OE ≤ 0.21 met moderate (PM1) levels of evidence for pathogenicity (**Fig. 4b**), but no MCR missense OE threshold met any levels of evidence to support benignity. However, separate calibration specifically in transcripts with multiple MCRs found that MCR missense OE ≥ 1.56 met moderate and OE ≥ 0.97 met supporting evidence for BP4, indicating that in transcripts where we are powered to characterize regional constraint, MCRs with OEs close to one harbor an indication of benignity (**Supplementary Fig. 9**). Calibration of two additional regional constraint metrics, Constrained Coding Regions (CCRs[9]) and COntact Set MISsense tolerance (COSMIS[12]), which incorporates predicted 3D structure information, revealed that these metrics also reach moderate support for pathogenicity (PM1), and COSMIS only reaches supporting levels for benignity (**Supplementary Fig. 9; Supplementary Table 2**).

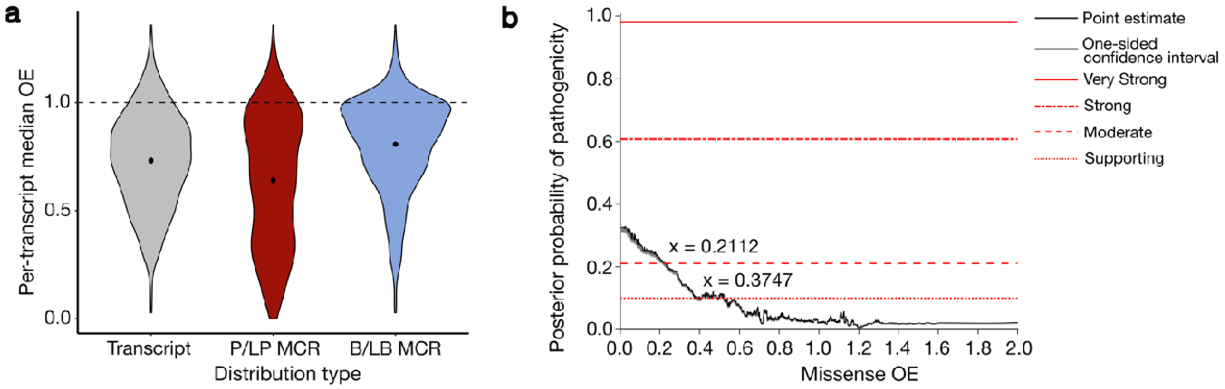

Fig. 4: ACMG/AMP calibration of missense constraint. **a,** The distribution within genes with autosomal dominant disease associations of transcript-wide missense observed/expected (OE; gray) and missense constraint region (MCR) OE for ClinVar pathogenic/likely pathogenic missense variants (P/LP; red) and benign/likely benign missense variants (B/LB; blue). We filtered to 1,007 transcripts with at least one P/LP and one B/LB missense variant. For the P/LP and B/LB distributions, we annotated each variant with the missense OE across the MCR they fell in and collapsed these values within each transcript by taking the respective medians. **b,** Local posterior probabilities of pathogenicity given MCR missense OE in all transcripts. Gray shading indicates the one-sided 95% confidence interval on the more stringent side. Horizontal lines indicate thresholds required to meet ACMG/AMP evidence levels. From bottom to top: supporting, moderate, strong, very strong. MCR missense OE reaches supporting (OE ≤ 0.37) and moderate (OE ≤ 0.21) level evidence for PM1 (hotspot/functional domain).

We transformed our regional missense constraint measure into a variant-level predictor of missense deleteriousness named MPC (Missense deleteriousness Prediction by Constraint) that additionally incorporates information about amino acid substitution type and local context. The logistic regression-based model integrates regional missense constraint-derived metrics together with BLOSUM[38], Grantham[39], and PolyPhen-2[40] and is trained on ClinVar pathogenic and gnomAD common (AF > 0.1%) variants in 2,987 genes defined as haploinsufficient in Collins *et al*.[41] and 366 genes with DD associations in G2P through non-LoF mechanisms. Higher scores predict greater deleteriousness (**Supplementary Fig. 10, 11**). We assessed the utility of MPC in prioritizing potentially disease-causing variation by evaluating its ability to stratify case and control rare and *de novo* missense variation. Consistent with the regional constraint results, the *de novo* missense variants from DD and AUT cases are enriched for high MPC scores compared to controls (**Supplementary Fig. 12**). We further stratified by presence in 373 genes previously associated with NDD[16] and three bins of MPC scores (< 1.6, 1.6-2.6, ≥ 2.6; see **Supplementary Note** for calibration of these bins), and found a very strong enrichment of *de novo* missense variants in the two most deleterious bins among both the DD (**Fig. 5a**) and

AUT cases (**Fig. 5b**) compared to unaffected individuals. However, while the enrichment in the 373 NDD-associated genes was significant for missense variants with MPC ≥ 2.6 (RR in DD cases = 22.7, $p < 10^{-50}$; RR in AUT cases = 6.9, $p < 10^{-21}$), as well as missense variants with MPC between 1.6-2.6 (RR in DD cases = 4.5, $p < 10^{-35}$; RR in AUT cases = 1.9, $p = 3.0 \times 10^{-5}$), it was only significant in NDD-unassociated genes for missense variants with MPC ≥ 2.6 (RR in DD cases = 3.1, $p < 10^{-28}$; RR in AUT cases = 1.5, $p = 5.9 \times 10^{-4}$). This suggests that while there is a sizable reservoir of potentially causal variants in genes yet to be associated with NDDs, they will be more difficult to find as they must reach stricter deleteriousness criteria. For autism, we additionally assessed inheritance rates of rare missense variants (AF < 0.1%) from parents to 13,384 probands and case-control rates for an additional 5,591 cases and 8,597 controls without *de novo* information. While we did not find substantial enrichment in inheritance rates in any missense category, we discovered substantial enrichment in the case-control analysis for variants in the 373 NDD-associated genes with MPC ≥ 2.6 (RR = 1.6, $p < 10^{-12}$), which we infer is from *de novo* variants that are not recognizable as such due to lack of parental information.

We extended our assessment of case-control *de novo* stratification for a comparison of our model against several other missense deleteriousness predictors: AlphaMissense[42], CCRs[9], M-CAP[43], REVEL[44], PrimateAI-3D[45], MVP[46], Polyphen-2[40], CADD[47,48], mammalian conservation phyloP[35], and SIFT[49]. For this assessment, we evaluated four additional early-onset development-related phenotypes: epileptic encephalopathy (EE), orofacial cleft (OFC), congenital heart disease (CHD), and congenital diaphragmatic hernia (CDH). To compare across predictors with different score distributions, we used a ranking-based performance assessment. For each predictor, we ranked the *de novo* missense variants from each case cohort against those in the 5,492 controls and computed the OR of case vs. control variants in the top percentiles of these rankings (**Fig. 5c**). At the top 10% of variants, MPC displays the highest OR for DD (OR = 5.2, Fisher's exact $p < 10^{-48}$), EE (OR = 3.1, $p = 2.2 \times 10^{-7}$), AUT (OR = 1.7, $p = 8.9 \times 10^{-9}$), and OFC (OR = 1.5, p = 0.025), although there is substantial confidence interval overlap with other predictors. This indicates that MPC effectively ranks high-impact *de novo* variants in the most deleterious prediction regimes. Of the other predictors, AlphaMissense also performs consistently well across all phenotypes. In particular, in CHD and CDH, which have the least *de novo* enrichment across predictors, we observe MPC lagging in performance, while AlphaMissense is one of the top performers. This may suggest that causal *de novo* variants in these phenotypes may occur at a narrow set of sites where 3D structure is important, which AlphaMissense can more deftly capture through integration of protein structure prediction. These observations are more or less consistent over a range of thresholds used to define the ranking top percentiles (**Supplementary Fig. 13**).

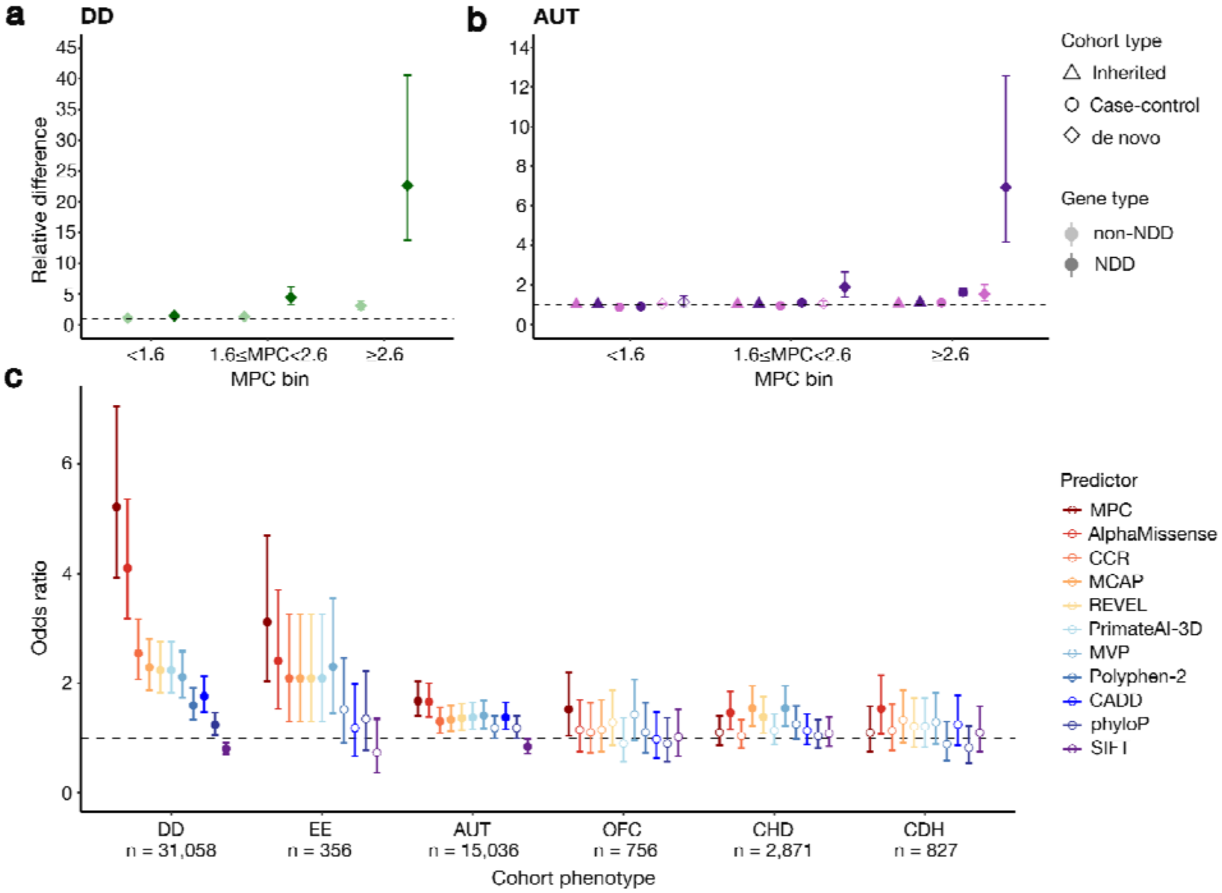

Fig. 5: MPC effectively stratifies case and control variation.
a, The difference relative to controls of missense variants stratified by MPC score and localization to genes associated with neurodevelopmental disorders (NDDs) for **a,** individuals with DD and **b**, autistic individuals (AUT). Relative difference is calculated as: for *de novo* variants, the average rate of variants in probands divided by that in sibling controls; for case-control, the average rate of variants in cases divided by that in controls from case-control data; for inherited, the average rate in probands of transmitted variants divided by that of untransmitted variants. Error bars represent 95% confidence intervals calculated from a binomial test. **c**, The odds ratio of case to control *de novo* missense variants in the top 10% vs. bottom 90% of respective rankings. *De novo* missense variants from each case cohort are ranked against those in the 5,492 controls for each predictor. DD: developmental disorders, EE: epileptic encephalopathy, AUT: autism, OFC: orofacial cleft, CHD: congenital heart disease, CDH: congenital diaphragmatic hernia. Error bars represent 95% confidence intervals. Only variants scored by all predictors are included. Points are solid colored if the difference from 1 is statistically significant (binomial or Fisher exact p < 0.05).

We have developed a method to identify sub-genic regions with differential intolerance to missense variation at base-level resolution. We demonstrate that coding regions depleted for missense variation in the general population are enriched for established disease-associated variation, *de novo* variants from individuals with NDDs, and heritability for 268 complex traits from the UK Biobank and other large GWAS. Additionally, we have calibrated these constraint scores to establish that regions with less than 20% of their expected variation can achieve moderate evidence for association to disease following ACMG/AMP guidelines. Finally, we incorporated regional missense intolerance information into the missense deleteriousness metric, MPC, and show that MPC effectively separates potentially risk-carrying variants identified in various developmental disorder cases from those seen in controls.

At current sample sizes, we are unable to characterize constraint at single amino acid resolution. Furthermore, because our approach relies on variant presence or absence in a large reference dataset, many of the constrained regions we find are linked to variants that cause severe, early-onset disease. However, the true nature of the variation we capture is more accurately linked to reproductive fitness and the strength of selection acting on heterozygotes[50]. Our methodology specifically searches for linear sub-genic regions in canonical transcripts that are depleted of missense variants compared to a null mutational model. This means that our model is unable to find depleted sequences that are clustered specifically in 3D space and is also currently ignorant of coding sequences not present in the Ensembl canonical transcript. However, we note that our linear metric achieves similar evidence for both pathogenicity and benignity as the structural constraint-based COSMIS model[12] (**Supplementary Fig. 9**).

In summary, we identify 28% of canonical transcripts with variable levels of missense constraint and demonstrate that coding regions specifically depleted of missense variation in the general population are enriched for disease-associated variation. Additionally, we show that this depletion of missense variation can be used as moderate evidence when classifying variants according to ACMG/AMP guidelines and that incorporation of regional missense constraint into an *in silico* predictor effectively prioritizes a subset of *de novo* missense variation in individuals with developmental phenotypes for association testing. We have publicly released these data for use in both research and clinical settings. We anticipate refined resolution of these metrics as datasets grow, both in size and in ancestral diversity, and with the incorporation of complementary structural or functional data.

Methods

Transcripts

This study analyzed only canonical, coding transcripts as defined by GENCODE v19/Ensembl v74. We excluded the same set of transcripts from this analysis that were excluded in the previous gnomAD v2.1.1 genic constraint estimates[13]. Briefly, we excluded transcripts that had outlier counts of variants expected under neutrality (zero expected pLoF, missense, or synonymous variants; too many observed pLoF, missense, or synonymous variants compared to expectation; or too few observed synonymous variants compared to expectation). In total, this study analyzed 18,629 transcripts.

gnomAD variants

All analyses in this paper were conducted using the 125,748 gnomAD v2.1.1 exomes[13] on GRCh37. Median coverage was calculated on a random subset of the gnomAD exomes as described previously[13]. We defined the set of sites with possible missense variants using a synthetic Hail Table (HT) containing all possible single nucleotide variants in the exome. We annotated this HT with the Variant Effect Predictor (VEP, version 85) against GENCODE version 19, and filtered to variants with the consequence "missense_variant" in the canonical, coding transcripts as defined in *Transcripts*. We then further filtered to variants that fit one of the following criteria: (1) allele count (AC) > 0 and AF < 0.001, variant QC PASS, and median coverage > 0 in gnomAD v2.1.1 exomes; or (2) AC = 0, i.e. variants not seen in gnomAD v2.1.1 exomes.

ClinVar variants

We annotated functional consequences for ClinVar[14] (v.20230305) variants using the VEP table described in *gnomAD variants.* Missense ClinVar variants with non-conflicting P, LP, B, LB classification and a review status of at least one star were selected for analysis.

Rare and *de novo* variants from developmental cohorts

Case *de novo* mutations for association analyses were obtained from studies of developmental disorders[15] (DD), autism[16] (AUT), congenital heart disease[51] (CHD), orofacial cleft[52] (OFC), congenital diaphragmatic hernia[53] (CDH), and epileptic encephalopathy[54] (EE). Control *de novo* mutations were obtained from neurotypical siblings of the autistic probands[16]. Variants from the autism study were lifted over from GRCh38 to GRCh37 using the “liftover” function in Hail. Variant functional consequences were re-annotated using the VEP table described in *gnomAD variants*. Variants transmitted and not transmitted from parents to autistic probands were procured from previously published ASC-SSC and SPARK cohorts, and case-control variants for autism were procured from previously published iPSYCH and Swedish cohorts[16]. Both the inherited/uninherited and case-control variant sets were filtered to AF < 0.1%.

Training, validation, and test datasets

To generate independent training and test sets, we selected 80% (14,894 transcripts) of the 18,629 canonical coding transcripts to comprise the training set and the remaining 20% (3,735 transcripts) to the test set. To ensure the training and test transcripts have similar distributions of features that may impact constraint estimates, we used stratified randomization to match the training and test transcripts on s_{het} coefficients (as a measure of selection) and number of potential missense sites (as a measure of power to detect transcript-wide constraint changes). The training set was used for MPC model training and MCR model selection, and the test set was used for MPC model evaluation. No similar hold-outs of data were performed for training of the mutational model used to compute expected variant counts (see *Modeling of mutation rates and expected neutral missense variation*).

Modeling of mutation rates and expected neutral missense variation

Expected missense variant counts were determined as described previously[13]. Briefly, we created a model using the 15,708 gnomAD v2.1.1 genomes that estimated the mutation rate for each single nucleotide substitution with one base of context (e.g., ACT > AGT) in non-coding regions of the genome. We then calibrated this mutation rate against the proportion observed of each context at synonymous sites to adjust for the larger size of the gnomAD v2 exomes, adjusting for low coverage regions (median coverage < 40x) and methylation levels at CpG sites using methylation data from the Roadmap Epigenomics Consortium[55]. We created three separate models (referred to as "plateau" models moving forwards): one for autosomal and pseudoautosomal sites, one for chromosome X sites, and one for chromosome Y sites. Each of these models contains mutation rate estimates for each substitution, context, and methylation level. We then applied the plateau models to the proportion observed of each substitution and its context, exome coverage, and methylation level. We counted all possible variants in our synthetic Hail Table (HT) that passed the following criteria: (1) Median coverage > 0; (2) no low-quality variant observed in gnomAD v2 exomes; (3) no variants above 0.1% AF observed in gnomAD v2 exomes. We then correlated this proportion observed value with the mutation rate calculated using the appropriate model above. For low coverage sites (median coverage below 40x), we calculated a scaling factor as described previously[13]: briefly, we computed the total number of observed synonymous variants in the gnomAD v2 exomes divided by the total number of possible synonymous variants in the synthetic HT multiplied by the mutation rate aggregated across all possible substitutions and their contexts and methylation levels. We used this scaling factor to create a model to adjust the proportion of expected variation for low coverage sites (coverage model).

Identifying breakpoints within transcripts of regional missense constraint

Observed missense variant counts were calculated using sites from the 125,748 gnomAD exomes that passed all the following criteria: (1) Allele count (AC) > 0; (2) allele frequency (AF) < 0.001; (3) variant QC PASS (passed gnomAD variant QC filters, including random forest filters); (4) median coverage > 0. We filtered the gnomAD v2 exomes Hail Table (HT) to the sites that matched the above criteria and then annotated the synthetic HT with whether that variant (chromosome/locus plus reference and alternate alleles) was observed in the gnomAD exomes. We then aggregated the total number of observed variant counts per locus by summing the number of observed variants for each possible substitution (reference and alternate allele) at each locus. Finally, we grouped the synthetic HT by transcript annotation to sum the total number of observed missense variants per transcript.

As previously described[13], we applied the two models (plateau and coverage) described in *Modeling of mutation rates and expected neutral missense variation* to calculate the total proportion of expected missense variation. Briefly, we summed the mutation rate (mu_agg) for each substitution, context, and methylation level across the exome. We then applied the appropriate plateau model (autosomal/pseudoautosomal, chromosome X, chromosome Y) and adjusted CpG vs. non-CpG sites separately. After applying the appropriate plateau model, we applied the coverage model to low coverage (median coverage < 40x) sites to create the final adjusted mutation rate (mu_adj). We then aggregated the raw mutation rate sum (mu_agg) and adjusted mutation rate (mu_adj) per transcript to get the total mutation rate sum and proportion of expected missense variation per transcript.

We implemented a minimum number of expected missense variants to prevent finding breakpoint positions that would create very small (i.e., a handful of base pairs in size) transcript subsections (see **Supplementary Note**).

We applied a likelihood ratio test to determine whether the missense observed/expected (OE) ratio was uniform along a transcript or whether a transcript had evidence of distinct sections of missense constraint. We used the observed and expected missense counts to search for positions that would divide a transcript into two or more regions with varying levels of missense depletion. For our analyses, we assume that the observed missense counts should follow a Poisson distribution around the expected missense counts. We defined our null model as transcripts not having any evidence of regional variability in missense depletion (where the expectation, the OE ratio, is consistent across the length of the transcript). Our alternative model was that transcripts exhibited evidence of distinct sections of missense depletion (OE ratio calculated per transcript subsection). Because the alternative model should always have a better fit than the null model, we require a chi square value above a given threshold (p = 0.001) to establish significance. We used the following formulas to determine the significance of a breakpoint that would split a transcript into two sections, A and B:

- $p_0 = Pois(obs_A, exp_A * OE) * Pois(obs_B, exp_B * OE)$
- $p_1 = Pois(obs_A, exp_A * OE_A) * Pois(obs_B, exp_B * OE_B)$
- $\chi^2 = 2(ln(p_1) - ln(p_0))$

where OE is the missense observed/expected ratio across the entire transcript, obs_A is the number of observed missense variants in transcript section A, exp_A is the number of expected missense variants in transcript section A, OE_A is the OE ratio across transcript section A, obs_B is the number of observed missense variants in transcript section B, exp_B is the number of expected missense variants in transcript section A, OE_B is the OE ratio across transcript section B, and Pois is the Poisson likelihood.

We used the following formulas to determine the significance of a breakpoint that would split a transcript into three sections, A, B, and C:

- $p_0 = Pois(obs_A, exp_A * OE) * Pois(obs_B, exp_B * OE) * Pois(obs_C, exp_C * OE)$
- $p_1 = Pois(obs_A, exp_A * OE_A) * Pois(obs_B, exp_B * OE_B) * Pois(obs_C, exp_C * OE_C)$
- $\chi^2 = 2(ln(p_1) - ln(p_0))$

where OE is the missense observed/expected ratio across the entire transcript, obs_A is the number of observed missense variants in transcript section A, exp_A is the number of expected missense variants in transcript section A, OE_A is the OE ratio across transcript section A, obs_B is the number of observed missense variants in transcript section B, exp_B is the number of expected missense variants in transcript section A, OE_B is the OE ratio across transcript section B, obs_C is the number of observed missense variants in transcript section C, exp_C is the number of expected missense variants in section C, and Pois is the Poisson likelihood.

For the purposes of our analyses, all transcript subsections with more observed variants than expected were capped at an OE of 1, as we were looking for areas of missense depletion and not missense enrichment. We also converted the expected counts for transcript subsections with zero expected variants from 0 to 10^{-9} to prevent nonfinite OE values.

To search for a single breakpoint that would divide a transcript into two subsections, we calculated chi square statistics (as discussed above) to conduct likelihood ratio tests simultaneously for every eligible position within a transcript. The positions we considered were positions with a possible missense variant substitution that had at least 16 expected missense counts in either direction (i.e., both transcript subsections created by dividing the transcript at this point would have at least 16 expected missense variants). We then aggregated chi square values across each transcript to find the maximum value per transcript, and we marked any positions as breakpoints if the chi square calculated at that position was equal to the maximum chi square value over all sites in the transcript and significant at $p = 0.001$.

Any transcripts that did not have a single significant breakpoint moved forwards into our two simultaneous breaks search flow. In this search flow, we again calculated chi square statistics to conduct likelihood ratio tests for every eligible position pair. For every position with a possible missense, we calculated the chi square statistic of that position paired with each possible position downstream as long as the two positions created transcript subsections with at least 16 expected missense variants (i.e., all three of the transcript subsections created would have at least 16 expected missense variants). Because of the large number of pairwise computations, this step is the most computationally intensive portion of our algorithm. After completing the single and two simultaneous break search workflows, we merged the results from both search types.

Our breakpoint search flow is recursive, and the steps are as follows: Search for a single significant breakpoint dividing a transcript into two subsections. If no single significant breakpoint was found in the transcript, search for two simultaneous breakpoints. Merge the results from the single and two simultaneous breakpoint searches. Repeat the steps above, treating each separate transcript subsection as if it were an independent transcript, until no more significant breakpoints are found.

Modeling deleteriousness of missense substitution classes with missense constraint

We incorporated two MCR OE-based metrics to measure the increased deleteriousness of amino acid substitution classes (e.g., Met to Tyr) in functionally important areas of proteins: the overall OE for each substitution and the second derivative of this OE value per OE bin of missense constraint (**Supplementary Fig. 14**). To calculate the first metric, the substitution overall OE, we divided the total number of rare, high quality variants (see *gnomAD variants*) causing that substitution by the total number of expected variants (see *Modeling of mutation rates and expected neutral missense variation*). To calculate the second metric, the substitution OE second derivative, we aggregated the OEs of each substitution by MCR OE bin in 10 bins from 0 to 1.0+ (i.e., for the 0-0.1 OE bin, we calculated all of the observed substitutions that occurred within regions with a OE between 0 and 0.1 and divided that number by the total number of expected substitutions occurring in those regions).

Modeling deleteriousness of individual missense variants

We designed a missense variant deleteriousness predictor to explicitly incorporate information on amino acid substitution class and position-specific variant effects. A logistic regression model was first trained to differentiate pathogenic from benign missense variants. The pathogenic training set consisted of high-quality ClinVar variants (see *ClinVar variants*) labeled as pathogenic or likely pathogenic in 2,987 likely-haploinsufficient genes, defined as having probability of haploinsufficiency (pHaplo) ≥ 0.86[41], or in 366 genes with DD associations in G2P through non-LoF mechanisms. The latter gene set was created by filtering on the G2P DD panel (accessed October 6, 2023) to select genes where: 1. confidence_category is either definitive or strong evidence, 2. allelic_requirement was monoallelic, and 3. mutation_consequence included altered gene product structure or increased gene product level. The benign training set consisted of high-quality common variants as described in *gnomAD variants*. Variants matching criteria for both the benign and pathogenic sets were removed from the training data. We evaluated models with all possible combinations of the following complementary features: amino acid substitution overall OE and OE second derivative (see *Modeling deleteriousness of missense substitution classes*); BLOSUM[38] and Grantham[39] scores of amino acid substitution class severity; the local missense constraint level of a variant (missense OE across the MCR if applicable, else across the transcript); and PolyPhen-2[40]. We selected BLOSUM, Grantham, and PolyPhen-2 because of the orthogonal information added on top of our OE-based (and therefore population allele frequency-dependent) metrics. For each model, only variants with all relevant annotations were used in training the regression model and the subsequent calculations to produce deleteriousness scores. The deleteriousness score prediction for any missense variant *i* is given as:

$$d_i = -log_{10}(m_i/M)$$
$$m_i = max(0.83, f_i)$$

where d_i is the deleteriousness score prediction, f_i is the number of common missense variants with a fitted value from the regression that is less than the fitted value for variant *i*, and *M* is the number of common missense variants (equivalent to the number of benign training variants for the regression). m_i is set to have a minimum value of 0.83 to avoid a mathematical error in the log when the fitted value for a given variant is less than those of all common variants. Larger values of d_i indicate stronger predicted-deleteriousness. The best model was chosen to be the model featurized with all six possible features. The training set for this model consisted of 64,023 benign variants and 12,955 pathogenic variants. This model was then applied to produce MPC scores for the 68,576,965 possible exome-wide missense variants with all features, and the distribution of these MPC scores is given in **Supplementary Figs. 10, 11,** and **12**.

Comparison of MPC to other predictors

We compared our model to the following missense deleteriousness predictors: AlphaMissense[42], Constrained Coding Regions (CCRs)[9], MVP[46], M-CAP[43], PrimateAI-3D[45], REVEL[44], CADD[47,48], PolyPhen-2[40], and SIFT[49]. We annotated the case and control *de novo* missense variants described in *Rare and de novo variants from developmental cohorts* and ranked the variants based on their annotated scores. To assess each predictor's ability to stratify case and control variation, we assessed the proportion of case to control variants among the variants with the top 10% for each score and compared this number to the overall proportion of case to control variation using a Fisher exact test.

Data availability

The missense constraint regions (MCRs) are displayed on the gnomAD v2 browser (https://gnomad.broadinstitute.org) and available for download on the gnomAD website (https://gnomad.broadinstitute.org/downloads#v2) and in the gnomAD v2 public datasets on Google, Amazon, and Microsoft clouds. MPC scores for all possible variants in canonical transcripts is available in the gnomAD v2 public datasets on Google (gs://gcp-public-data--gnomad/release/2.1.1/regional_missense_constraint/gnomad_v2.1.1_mpc.ht). gnomAD v2 exome, genome, and coverage data and the table of all possible single nucleotide polymorphisms used to calculate mutational models and search for MCRs are also available in the gnomAD public buckets and are easily accessed using code in the gnomAD Hail utilities GitHub repository (https://github.com/broadinstitute/gnomad_methods/blob/7c0c994883f321492a48962674d5cae b289df4c7/gnomad/resources/grch37/gnomad.py#L107 and https://github.com/broadinstitute/gnomad_methods/blob/7c0c994883f321492a48962674d5caeb289df4c7/gnomad/utils/vep.py#L161).

ClinVar data were downloaded from ClinVar's FTP server (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/). *De novo* variants were extracted from the supplemental files of the cited studies.

AlphaMissense scores were downloaded from https://github.com/google-deepmind/alphamissense. CCRs were downloaded from https://github.com/quinlan-lab/ccrhtml. MCAP scores were downloaded from http://bejerano.stanford.edu/MCAP/. REVEL scores were downloaded from https://sites.google.com/site/revelgenomics/. PrimateAI-3D scores were downloaded from https://primad.basespace.illumina.com/download. MVP scores were downloaded from https://figshare.com/articles/dataset/Predicting_pathogenicity_of_missense_variants_by_deep_learning/13204118. PolyPhen-2 and SIFT scores were obtained from VEP[56]. CADD scores were downloaded from the CADD website (https://cadd.gs.washington.edu/download). phyloP scores were downloaded from the UCSC browser (https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg38&g=cons241way).

Code availability

Code to determine missense constraint regions (MCRs) and calculate MPC is available at https://github.com/broadinstitute/regional_missense_constraint. Code used to generate the mutational models is available at https://github.com/broadinstitute/gnomad_lof and https://github.com/broadinstitute/gnomad-constraint. The Hail library is available at https://hail.is/.

Acknowledgments

We thank all of the individuals who contributed their data to gnomAD for enabling the research presented here and in numerous other publications. We also thank members of the gnomAD Production and Browser Teams for their behind-the-scenes efforts in producing and sharing the gnomAD dataset, Dan King and the rest of the Hail team for enabling data processing at scale, and Emma Pierce-Hoffman for her work curating gene lists. This work was supported by the National Human Genome Research Institute (NHGRI; U24HG011450 to H.L.R. and M.J.D.; U01HG011755 to A.O.D.L., H.L.R, and M.E.T.; U01HG012022 to P.R.; 5T32HG002295-18 to L.W.), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD; F31HD111109 to L.W.), the National Science Foundation (NSF; GRFP#2022339661 to L.W.; GRFP#2023361442 to H.A.) the Simons Foundation (SFARI 1009802 to M.E.T. and K.E.S.), a National Health and Medical Research Council investigator grant (2009982) to D.G.M., and the Massachusetts General Hospital Executive Committee on Research, Interim Support Funding (K.E.S.).

Author information

These authors contributed equally: Katherine R. Chao and Lily Wang.

Contributions

K.R.C., L.W., K.E.S., B.M.N, and M.J.D conceived and designed experiments. K.R.C, L.W., and K.E.S. performed primary writing of the manuscript. K.R.C., L.W., R.P., C.L., H.A., and R.Y. performed the analyses and generated figures. P.S. and J.C. were instrumental to developing methods. R.H.G., N.A.W., and M.S. developed visualizations for the web browser. B.W., W.P., M.W.W., K.M.L., J.K.G, K.J.K, and G.T. completed code review for methods. D.G., J.I.G., C.V., and T.P. helped debug runtime compute. J.A.K. provided data and analysis suggestions. S.B. contributed analysis suggestions. H.L.R., B.M.N., M.E.T, D.G.M, A.O.D.L., K.J.K., P.R., M.J.D., and K.E.S. supervised the research. All authors listed under the Genome Aggregation Database Consortium contributed to the generation of the primary data incorporated into the gnomAD resource. All authors reviewed the manuscript.

Ethics declarations

Competing interests/Declaration of interests

J.A.K is a current employee of Regeneron Genetics Center. T.P. and G.T. are founders of E9 Genomics, Inc.. H.L.R. has received support from Illumina and Microsoft to support rare disease gene discovery and diagnosis. B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora. M.E.T. has received research support and/or reagents from Illumina, Pacific Biosciences, Microsoft, Oxford Nanopore, and Ionis Therapeutics. D.G.M. is a paid advisor to GSK, Insitro, and Overtone Therapeutics, and receives research funding from Microsoft. A.O.D.L. has consulted for Tome Biosciences and Ono Pharma USA Inc, and is member of the scientific advisory board for Congenica Inc and the Simons Foundation SPARK for Autism study, and received research support from Pacific Biosciences for rare disease diagnosis. K.J.K. is a consultant for Vor Biopharma, Tome Biosciences, and is on the Scientific Advisory Board of Nurture Genomics. M.J.D. is a founder of Maze Therapeutics and Neumora Therapeutics, Inc. (f/k/a RBNC Therapeutics). K.E.S. has received support from Microsoft for work related to rare disease diagnostics. All other authors declare no competing interests.

References

1. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017) doi:10.1101/148353.
2. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, 9 (2016).
3. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum. Mutat.* **38**, 1454–1463 (2017).
4. Sivley, R. M., Dou, X., Meiler, J., Bush, W. S. & Capra, J. A. Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am. J. Hum. Genet.* **102**, 415–426 (2018).
5. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med.* **12**, 28 (2020).
6. Zhang, X. *et al.* Genetic constraint at single amino acid resolution improves missense variant prioritisation and gene discovery. (2022) doi:10.1101/2022.02.16.22271023.
7. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* **27**, 1715–1729 (2017).
8. Perszyk, R. E., Kristensen, A. S., Lyuboslavsky, P. & Traynelis, S. F. Three-dimensional missense tolerance ratio analysis. *Genome Res.* **31**, 1447–1461 (2021).
9. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
10. Silk, M. *et al.* MTR3D: identifying regions within protein tertiary structures under purifying selection. *Nucleic Acids Res.* **49**, W438–W445 (2021).
11. Hicks, M., Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Functional characterization of 3D protein structures informed by human genetic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8960–8965 (2019).
12. Li, B., Roden, D. M. & Capra, J. A. The 3D mutational constraint on amino acid sites in the human proteome. *Nat. Commun.* **13**, 3273 (2022).
13. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
14. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, (2020).
15. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
16. Fu, J. M. *et al.* Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
17. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
18. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**,

285–291 (2016).

19. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
20. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
21. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* (2024) doi:10.1038/s41586-023-06957-x.
22. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet.* **11**, e1005492 (2015).
23. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
24. Weghorn, D. *et al.* Applicability of the Mutation-Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans. *Mol. Biol. Evol.* **36**, 1701–1710 (2019).
25. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *Elife* **12**, (2023).
26. Kosmicki, J. A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
27. Bamshad, M. J., Nickerson, D. A. & Chong, J. X. Mendelian Gene Discovery: Fast and Furious with No End in Sight. *Am. J. Hum. Genet.* **105**, 448–455 (2019).
28. Seaby, E. G., Rehm, H. L. & O'Donnell-Luria, A. Strategies to Uplift Novel Mendelian Gene Discovery for Improved Clinical Outcomes. *Front. Genet.* **12**, 674295 (2021).
29. Kapplinger, J. D. *et al.* Enhancing the Predictive Power of Mutations in the C-Terminus of the KCNQ1-Encoded Kv7.1 Voltage-Gated Potassium Channel. *J. Cardiovasc. Transl. Res.* **8**, 187–197 (2015).
30. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
31. Whiffin, N. *et al.* CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. *Genet. Med.* **20**, 1246–1254 (2018).
32. Küry, S. *et al.* Rare germline heterozygous missense variants in BRCA1-associated protein 1, BAP1, cause a syndromic neurodevelopmental disorder. *Am. J. Hum. Genet.* **109**, 361–372 (2022).
33. Hamosh, A., Amberger, J. S., Bocchini, C., Scott, A. F. & Rasmussen, S. A. Online Mendelian Inheritance in Man (OMIM®): Victor McKusick's magnum opus. *Am. J. Med. Genet. A* **185**, 3259–3265 (2021).
34. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
35. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. *Science* **380**, eabn3943 (2023).
36. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
37. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity

classification and ClinGen recommendations for PP3/BP4 criteria. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).

38. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
39. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
40. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
41. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055.e25 (2022).
42. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
43. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
44. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
45. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
46. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).
47. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
48. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
49. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
50. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
51. Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
52. Bishop, M. R. *et al.* Genome-wide Enrichment of De Novo Coding Mutations in Orofacial Cleft Trios. *Am. J. Hum. Genet.* **107**, 124–136 (2020).
53. Qiao, L. *et al.* Rare and de novo variants in 827 congenital diaphragmatic hernia probands implicate LONP1 as candidate risk gene. *Am. J. Hum. Genet.* **108**, 1964–1980 (2021).
54. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project & Epi4K Consortium. De novo mutations in synaptic transmission genes including DNM1 cause epileptic encephalopathies. *Am. J. Hum. Genet.* **95**, 360–370 (2014).
55. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
56. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

Genome Aggregation Database Consortium

Maria Abreu[12], Carlos A. Aguilar Salinas[13], Tariq Ahmad[14], Christine M. Albert[15,16], Jessica Alföldi[1,3], Diego Ardissino[17], Irina M. Armean[1,3], Gil Atzmon[18,19], Eric Banks[20], John Barnard[21], Samantha M. Baxter[1], Laurent Beaugerie[22], Emelia J. Benjamin[23,24,25], David Benjamin[20], Louis Bergelson[20], Michael Boehnke[26], Lori L. Bonnycastle[27], Erwin P. Bottinger[28], Donald W. Bowden[29,30,31], Matthew J. Bown[32,33], Harrison Brand[2,34], Steven Brant[35,36,37], Ted Brookings[20,38], Hannia Campos[39,40], John C. Chambers[41,42,43], Juliana C. Chan[44], Katherine R. Chao[1,3], Sinéad Chapman[1,3,5], Daniel I. Chasman[6,15], Siwei Chen[1,3], Rex Chisholm[45], Judy Cho[28], Rajiv Chowdhury[46], Mina K. Chung[47], Wendy K. Chung[48,49,50], Kristian Cibulskis[20], Bruce Cohen[51,52], Ryan L. Collins[1,2,53], Kristen M. Connolly[54], Adolfo Correa[55], Miguel Covarrubias[20], Beryl B. Cummings[1,53], Dana Dabelea[56], Mark J. Daly[1,3,57], John Danesh[46], Dawood Darbar[58], Phil Darnowsky[1], Joshua C. Denny[59], Stacey Donnelly[60], Ravindranath Duggirala[61], Josée Dupuis[62,63], Patrick T. Ellinor[1,64], Roberto Elosua[65,66,67], James Emery[20], Eleina England[1,68], Jeanette Erdmann[69,70,71], Tõnu Esko[1,72], Emily Evangelista[1], Yossi Farjoun[73], Martti Färkkilä[74,75,76], Diane Fatkin[77,78,79], Steven Ferriera[80], Jose Florez[6,81,82], Laurent Francioli[1,3], Andre Franke[83,84], Jack Fu[1,2,34], Stacey Gabriel[80], Kiran Garimella[20], Laura D. Gauthier[20], Jeff Gentry[20], Gad Getz[6,85,86], David C. Glahn[87,88], Benjamin Glaser[89], Stephen J. Glatt[90], David Goldstein[91,92], Clicerio Gonzalez[93], Julia Goodrich[1,2], Riley H. Grant[1], Leif Groop[94,95], Sanna Gudmundsson[1,3,10], Namrata Gupta[1,80], Andrea Haessly[20], Christopher Haiman[96], Ira Hall[97], Craig L. Hanis[98], Matthew Harms[99,100], Qin He[1], Mikko Hiltunen[101], Matti M. Holi[102], Christina M. Hultman[103,104], Steve Jahl[1,3], Chaim Jalas[105], Thibault Jeandet[20], Mikko Kallela[106], Diane Kaplan[20], Jaakko Kaprio[95], Konrad J. Karczewski[1,3,5], Sekar Kathiresan[2,6,107], Eimear E. Kenny[108], Bong-Jo Kim[109], Young Jin Kim[109], Daniel King[1], George Kirov[110], Zan Koenig[3,5], Jaspal Kooner[42,111,112], Seppo Koskinen[113], Harlan M. Krumholz[114,115], Subra Kugathasan[116], Soo Heon Kwak[117], Markku Laakso[118,119], Nicole Lake[120], Trevyn Langsford[20], Kristen M. Laricchia[1,3], Terho Lehtimäki[121,122], Monkol Lek[120], Emily Lipscomb[1], Christopher Llanwarne[20], Ruth J.F. Loos[28,123,124], Wenhan Lu[1], Steven A. Lubitz[1,64], Teresa Tusie Luna[125,126], Ronald C.W. Ma[44,127,128], Daniel G. MacArthur[1,129,130], Gregory M. Marcus[131], Jaume Marrugat[132,133], Daniel M. Marten[1,10], Alicia R. Martin[1,3,5], Kari M. Mattila[134], Steven McCarroll[5,135], Mark I. McCarthy[136,137,138], Jacob L. McCauley[139,140], Dermot McGovern[141], Ruth McPherson[142], James B. Meigs[1,6,143], Olle Melander[144], Andres Metspalu[145], Deborah Meyers[146], Eric V. Minikel[1], Braxton D. Mitchell[147], Ruchi Munshi[20], Aliya Naheed[148], Saman Nazarian[149,150], Benjamin M. Neale[1,3], Peter M. Nilsson[151], Sam Novod[20], Anne H. O'Donnell-Luria[1,2,10], Michael C. O'Donovan[152], Yukinori Okada[153,154,155], Dost Ongur[6,51], Lorena Orozco[156,157], Michael J. Owen[152], Colin Palmer[158], Nicholette D. Palmer[29], Aarno Palotie[3,5,95], Kyong Soo Park[117,159], Carlos Pato[160], Nikelle Petrillo[20], William Phu[1,10], Timothy Poterba[1,3,5], Ann E. Pulver[161], Dan Rader[149,162], Nazneen Rahman[163], Heidi Rehm[1,2], Alex Reiner[164,165], Anne M. Remes[166,167], Dan Rhodes[1], Stephen Rich[168,169], John D. Rioux[170,171], Samuli Ripatti[60,95,172], David Roazen[20], Dan M. Roden[173,174], Jerome I. Rotter[175], Valentin Ruano-Rubio[20], Nareh Sahakian[20], Danish Saleheen[176,177,178], Veikko Salomaa[179], Andrea Saltzman[1], Nilesh J. Samani[33,180], Kaitlin E. Samocha[1,2], Jeremiah Scharf[1,2,5], Molly Schleicher[1], Sebastian Schönherr[181], Patrick Schultz[1,3,5], Heribert Schunkert[182,183], Eleanor G. Seaby[1,184], Cotton Seed[3,5], Svati H. Shah[185,186], Megan Shand[20], Ted Sharpe[20], Moore B. Shoemaker[187], Tai Shyong[188,189], Edwin K. Silverman[190,191], Moriel Singer-Berk[1], Pamela Sklar[192,193,194], J. Gustav Smith[195,196,197], Hilkka Soininen[198], Harry Sokol[199,200,201], Matthew Solomonson[1,3], Rachel G. Son[1], Jose Soto[20], Tim Spector[202], Christine Stevens[1,3,5], Nathan O. Stitziel[203,204,205], Patrick F. Sullivan[103,206], Jaana Suvisaari[179], E. Shyong Tai[207,208,209], Michael E. Talkowski[1,2,5], Yekaterina Tarasova[1], Kent D. Taylor[175], Yik Ying Teo[207,210,211], Grace Tiao[1,3], Kathleen Tibbetts[20], Charlotte Tolonen[20], Ming Tsuang[212,213], Tiinamaija Tuomi[95,214,215], Dan Turner[216], Teresa Tusie-Luna[217,218], Erkki Vartiainen[219], Marquis Vawter[220], Christopher Vittal[1,3], Gordon Wade[20], Arcturus Wang[1,3,5], Lily Wang[221], Qingbo

Wang[1,153], James S. Ware[1,222,223], Hugh Watkins[224], Nicholas A. Watts[1,3], Rinse K. Weersma[225], Ben Weisburd[20], Maija Wessman[95,226], Nicola Whiffin[1,227,228], Michael W. Wilson[1,3], James G. Wilson[229], Ramnik J. Xavier[230,231], Mary T. Yohannes[1]

[12]University of Miami Miller School of Medicine, Gastroenterology, Miami, USA
[13]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico
[14]Peninsula College of Medicine and Dentistry, Exeter, UK
[15]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA
[16]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[17]Department of Cardiology University Hospital, Parma, Italy
[18]Department of Biology Faculty of Natural Sciences, University of Haifa, Haifa, Israel
[19]Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA
[20]Data Science Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[21]Department of Quantitative Health Sciences, Lerner Research Institute Cleveland Clinic, Cleveland, OH, USA
[22]Sorbonne Université, APHP, Gastroenterology Department Saint Antoine Hospital, Paris, France
[23]NHLBI and Boston University's Framingham Heart Study, Framingham, MA, USA
[24]Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA
[25]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
[26]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
[27]National Human Genome Research Institute, National Institutes of Health Bethesda, MD, USA
[28]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[29]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA
[30]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
[31]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
[32]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK
[33]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
[34]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[35]Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
[36]Department of Genetics and the Human Genetics Institute of New Jersey, School of Arts and Sciences, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[37]Meyerhoff Inflammatory Bowel Disease Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[38]Fulcrum Genomics, Boulder, CO, USA
[39]Harvard School of Public Health, Boston, MA, USA
[40]Central American Population Center, San Pedro, Costa Rica
[41]Department of Epidemiology and Biostatistics, Imperial College London, London, UK
[42]Department of Cardiology, Ealing Hospital, NHS Trust, Southall, UK
[43]Imperial College, Healthcare NHS Trust Imperial College London, London, UK
[44]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China

[45]Northwestern University, Evanston, IL, USA
[46]University of Cambridge, Cambridge, England
[47]Departments of Cardiovascular, Medicine Cellular and Molecular Medicine Molecular Cardiology, Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA
[48]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, USA
[49]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA
[50]Department of Medicine, Columbia University Medical Center, New York, NY, USA
[51]McLean Hospital, Belmont, MA, USA
[52]Department of Psychiatry, Harvard Medical School, Boston, MA, USA
[53]Division of Medical Sciences, Harvard Medical School, Boston, MA, USA
[54]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[55]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA
[56]Department of Epidemiology Colorado School of Public Health Aurora, CO, USA
[57]Institute for Molecular Medicine Finland, (FIMM) Helsinki, Finland
[58]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA
[59]Vanderbilt University Medical Center, Nashville, TN, USA
[60]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[61]Department of Life Sciences, College of Arts and Scienecs, Texas A&M University-San Antonio, San Antonio, TX, USA
[62]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[63]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
[64]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
[65]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain
[66]CIBER CV, Spain
[67]Departament of Medicine, Faculty of Medicine, University of Vic-Central University of Catalonia, Vic Catalonia, Spain
[68]Clalit Genomics Center, Ramat-Gan, Israel
[69]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
[70]German Research Centre for Cardiovascular Research, Hamburg/Lübeck/Kiel, Lübeck, Germany
[71]University Heart Center Lübeck, Lübeck, Germany
[72]Estonian Genome Center, Institute of Genomics University of Tartu, Tartu, Estonia
[73]Richards Lab, Lady Davis Institute, Montreal, QC, Canada
[74]Helsinki University and Helsinki University Hospital Clinic of Gastroenterology, Helsinki, Finland
[75]Helsinki University and Helsinki University Hospital, Helsinki, Finland
[76]Abdominal Center, Helsinki, Finland
[77]Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia
[78]Faculty of Medicine and Health, UNSW Sydney, Kensington, NSW, Australia
[79]Cardiology Department, St Vincent’s Hospital, Darlinghurst, NSW, Australia
[80]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[81]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[82]Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[83]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany
[84]University Hospital Schleswig-Holstein, Kiel, Germany

[85]Bioinformatics Program MGH Cancer Center and Department of Pathology, Boston, MA, USA
[86]Cancer Genome Computational Analysis, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[87]Department of Psychiatry and Behavioral Sciences, Boston Children's Hospitaland Harvard Medical School, Boston, MA, USA
[88]Harvard Medical School Teaching Hospital, Boston, MA, USA
[89]Department of Endocrinology and Metabolism, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Israel
[90]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA
[91]Institute for Genomic Medicine, Columbia University Medical Center Hammer Health Sciences, New York, NY, USA
[92]Department of Genetics & Development Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA
[93]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Mexico
[94]Lund University Sweden, Sweden
[95]Institute for Molecular Medicine Finland, (FIMM) HiLIFE University of Helsinki, Helsinki, Finland
[96]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA
[97]Washington School of Medicine, St Louis, MO, USA
[98]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA
[99]Department of Neurology Columbia University, New York City, NY, USA
[100]Institute of Genomic Medicine, Columbia University, New York City, NY, USA
[101]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
[102]Department of Psychiatry, Helsinki University Central Hospital Lapinlahdentie, Helsinki, Finland
[103]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[104]Icahn School of Medicine at Mount Sinai, New York, NY, USA
[105]Bonei Olam, Center for Rare Jewish Genetic Diseases, Brooklyn, NY, USA
[106]Department of Neurology, Helsinki University, Central Hospital, Helsinki, Finland
[107]Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[108]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[109]Division of Genome Science, Department of Precision Medicine, National Institute of Health, Republic of Korea
[110]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, Wales
[111]Imperial College, Healthcare NHS Trust, London, UK
[112]National Heart and Lung Institute Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK
[113]Department of Health THL-National Institute for Health and Welfare, Helsinki, Finland
[114]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut
[115]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut
[116]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA
[117]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea

[118]The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland
[119]Kuopio University Hospital, Kuopio, Finland
[120]Department of Genetics, Yale School of Medicine, New Haven, CT, USA
[121]Department of Clinical Chemistry Fimlab Laboratories, Tampere University, Finland
[122]innish Cardiovascular Research Center-Tampere Faculty of Medicine and Health Technology, Tampere University, Finland
[123]The Mindich Child Health and Development, Institute Icahn School of Medicine at Mount Sinai, New York, NY, USA
[124]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[125]National Autonomous University of Mexico, Mexico City, Mexico
[126]Salvador Zubirán National Institute of Health Sciences and Nutrition, Mexico City, Mexico
[127]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
[128]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China
[129]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, Australia
[130]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia
[131]Division of Cardiology, University of California San Francisco, San Francisco, CA, USA
[132]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain
[133]CIBERCV, Madrid, Spain
[134]Department of Clinical Chemistry Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere Faculty of Medicine and Health Technology, Tampere University, Finland
[135]Department of Genetics, Harvard Medical School, Boston, MA, USA
[136]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital Old Road Headington, Oxford, OX, LJ, UK
[137]Welcome Centre for Human Genetics, University of Oxford, Oxford, OX, BN, UK
[138]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals, NHS Foundation Trust, John Radcliffe Hospital, Oxford, OX, DU, UK
[139]John P. Hussman Institute for Human Genomics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA
[140]The Dr. John T. Macdonald Foundation Department of Human Genetics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA
[141]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute Cedars-Sinai Medical Center, Los Angeles, CA, USA
[142]Atherogenomics Laboratory University of Ottawa, Heart Institute, Ottawa, Canada
[143]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA
[144]Department of Clinical Sciences University, Hospital Malmo Clinical Research Center, Lund University, Malmö, Sweden
[145]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
[146]University of Arizona Health Science, Tuscon, AZ, USA
[147]University of Maryland School of Medicine, Baltimore, MD, USA
[148]International Centre for Diarrhoeal Disease Research, Bangladesh
[149]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[150]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[151]Lund University, Dept. Clinical Sciences, Skåne University Hospital, Malmö, Sweden
[152]Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, Wales
[153]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan

[154]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
[155]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan
[156]Instituto Nacional de Medicina Genómica, (INMEGEN) Mexico City, Mexico
[157]Laboratory of Immunogenomics and Metabolic Diseases, INMEGEN,Mexico City, Mexico
[158]Medical Research Institute, Ninewells Hospital and Medical School University of Dundee, Dundee, UK
[159]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea
[160]Department of Psychiatry Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA
[161]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[162]Children's Hospital of Philadelphia, Philadelphia, PA, USA
[163]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK
[164]University of Washington, Seattle, WA, USA
[165]Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[166]Medical Research Center, Oulu University Hospital, Oulu Finland
[167]Research Unit of Clinical Neuroscience Neurology University of Oulu, Oulu, Finland
[168]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[169]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA
[170]Research Center Montreal Heart Institute, Montreal, Quebec, Canada
[171]Department of Medicine, Faculty of Medicine Université de Montréal, Québec, Canada
[172]Department of Public Health Faculty of Medicine, University of Helsinki, Helsinki, Finland
[173]Departments of Medicine, Pharmacology, Biomedical Informatics Vanderbilt, University Medical Center, Nashville, TN, USA
[174]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[175]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA
[176]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[177]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[178]Center for Non-Communicable Diseases, Karachi, Pakistan
[179]National Institute for Health and Welfare, Helsinki, Finland
[180]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
[181]Institute of Genetic Epidemiology, Department of Genetics, Medical University of Innsbruck, 6020 Innsbruck, Austria
[182]Department of Cardiology, Deutsches Herzzentrum München, Technical University of Munich, DZHK Munich Heart Alliance, Germany
[183]Technische Universität München, Germany
[184]Faculty of Medicine, University of Southampton, Southampton, SO16 6YD, UK
[185]Duke Molecular Physiology Institute, Durham, NC
[186]Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, NC, USA
[187]Division of Cardiovascular Medicine, Nashville VA Medical Center, Vanderbilt University School of Medicine, Nashville, TN, USA
[188]Division of Endocrinology, National University Hospital, Singapore
[189]NUS Saw Swee Hock School of Public Health, Singapore

[190]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA
[191]Harvard Medical School, Boston, MA, USA
[192]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[193]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[194]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[195]The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University
[196]Department of Cardiology, Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden
[197]Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden
[198]Institute of Clinical Medicine Neurology, University of Eastern Finad, Kuopio, Finland
[199]Sorbonne Université, INSERM, Centre de Recherche Saint-Antoine, CRSA, AP-HP, Saint Antoine Hospital, Gastroenterology department, F-75012 Paris, France
[200]INRA, UMR1319 Micalis, Jouy en Josas, France
[201]Paris Center for Microbiome Medicine, (PaCeMM) FHU, Paris, France
[202]Department of Twin Research and Genetic Epidemiology King's College London, London, UK
[203]Department of Medicine, Washington University School of Medicine, Saint Louis, MO, USA
[204]Department of Genetics, Washington University School of Medicine, Saint Louis, MO, USA
[205]The McDonnell Genome Institute at Washington University, Saint Louis, MO, USA
[206]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
[207]Saw Swee Hock School of Public Health National University of Singapore, National University Health System, Singapore
[208]Department of Medicine, Yong Loo Lin School of Medicine National University of Singapore, Singapore
[209]Duke-NUS Graduate Medical School, Singapore
[210]Life Sciences Institute, National University of Singapore, Singapore
[211]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[212]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA
[213]Institute of Genomic Medicine, University of California San Diego, San Diego, CA, USA
[214]Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland
[215]Institute of Genetics, Folkhalsan Research Center, Helsinki, Finland
[216]Juliet Keidan Institute of Pediatric Gastroenterology Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel
[217]Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico
[218]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico
[219]Department of Public Health Faculty of Medicine University of Helsinki, Helsinki, Finland
[220]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA
[221]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA; Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA, USA
[222]National Heart and Lung Institute, Imperial College London, London
[223]UK/MRC London Institute of Medical Sciences, Imperial College London, London, UK
[224]Radcliffe Department of Medicine, University of Oxford, Oxford, UK
[225]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, Netherlands
[226]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland

[227]Big Data Institute, University of Oxford, UK
[228]Wellcome Centre for Human Genetics, University of Oxford, UK
[229]Division of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA USA
[230]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[231]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA

Authors received funding as follows:
Emelia J. Benjamin: R01HL092577; American Heart Association AF AHA_18SFRN34110082
Matthew J. Bown: British Heart Foundation awards CS/14/2/30841 and RG/18/10/33842
Steven Brant: National Institutes of Health DK062431
Ravindranath Duggirala: U01 DK085524 National Institute for Diabetes and Digestive and Kidney Diseases (NIDDK)
Josée Dupuis: National Institute for Diabetes and Digestive and Kidney Diseases (NIDDK) R DK
Roberto Elosua: Agència de Gestió d'Ajuts Universitaris i de Recerca: 2021 SGR 00144
Jeanette Erdmann: VIAgenomics, Leducq network PlaqOmics, Deutsche Forschungsgemeinschaft Cluster of Excellence "Precision Medicine in Chronic Inflammation" (EXC2167);
Martti Färkkilä: State funding for university level health research
Laura D. Gauthier: Intel, Illumina
Benjamin Glaser: 5U01 DK085584
Stephen J. Glatt: U.S. NIMH Grant R MH
Leif Groop: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312063 and 336822), Sigrid Jusélius Foundation; IMI 2 (grant No 115974 and 15881)
Christopher Haiman: U01CA164973
Mikko Hiltunen: Academy of Finland (grant 338182), Sigrid Jusélius Foundation, the Strategic Neuroscience Funding of the University of Eastern Finland
Chaim Jalas: Bonei Olam
Mikko Kallela: Grants from State funding for university level health research and from Department of Neurology, Helsinki University, Central Hospital; Grant from Maire Taponen Foundation
Jaakko Kaprio: Academy of Finland (grants 312073 and 336823)
Ruth J.F. Loos: Novo Nordisk Foundation (NNF18CC0034900, NNF20OC0059313); NIH (R01DK110113; R01DK124097)
Ronald C.W. Ma: Research Grants Council of the Hong Kong Special Administrative Region (CU R4012-18), Research Grants; Council Theme-based Research Scheme (T12-402/13N), University Grants Committee Research Grants Matching Scheme
Jaume Marrugat: Agència de Gestió d'Ajuts Universitaris i de Recerca: 2021 SGR 00144
Jacob L. McCauley: National Institute of Diabetes and Digestive and Kidney Disease Grant R01DK104844
Michael C. O'Donovan: Medical Research Council UK: Centre Grant No. MR/L010305/1, Program Grant No. G0800509
Yukinori Okada: JSPS KAKENHI (19H01021, 20K21834), AMED (JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217, JP21ek0410075), JST Moonshot R&D (JPMJMS2021)
Michael J. Owen: Medical Research Council UK: Centre Grant No. MR/L010305/1, Program Grant No. G0800509
Aarno Palotie: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant numbers 312074 and 336824) and Sigrid Jusélius Foundation

John D. Rioux: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK; DK062432), from the Canadian Institutes of Health (CIHR GPG 102170), from Genome Canada/Génome Québec (GPH-129341), and a Canada Research Chair (#230625)
Samuli Ripatti: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant number) Sigrid Jusélius Foundation
Jerome I. Rotter: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center
Veikko Salomaa: Juho Vainio Foundation and Finnish Foundation for Cardiovascular Research
Jeremiah Scharf: NIH Grants U01 NS40024, K02 NS085048, NS102371
Eleanor G. Seaby: Kerkut Charitable Trust, Foulkes Fellowship, University of Southampton Presidential Scholarship
Edwin K. Silverman: NIH Grants U01 HL089856 and U01 HL089897
J. Gustav Smith: The Swedish Heart-Lung Foundation (2022-0344, 2022-0345), the Swedish Research Council (2021-02273), the European Research Council (ERC-STG-2015-679242), Gothenburg University, Skåne University Hospital, governmental funding of clinical research within the Swedish National Health Service, a generous donation from the Knut and Alice Wallenberg foundation to the Wallenberg Center for Molecular Medicine in Lund, and funding from the Swedish Research Council (Linnaeus grant Dnr 349-2006-237, Strategic Research Area Exodiab Dnr 2009-1039) and Swedish Foundation for Strategic Research (Dnr IRC15-0067) to the Lund University Diabetes Center
Harry Sokol: AgroParisTech, Jouy en Josas, France
Nathan O. Stitziel: National Human Genome Reserach Institute Grant UM1HG008853
Kent D. Taylor: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and

general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center
Tiinamaija Tuomi: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312072 and 336826), Folkhalsan Research Foundation, Helsinki University Hospital, Ollqvist Foundation, Liv och Halsa foundation; NovoNordisk Foundation
Teresa Tusie-Luna: CONACyT Project 312688
James S. Ware: Royal Brompton & Harefield Hospitals, Guy's and St. Thomas' NHS Foundation Trust, London, UK
Rinse K. Weersma: The Lifelines Biobank initiative has been made possible by subsidy from the Dutch Ministry of Health Welfare and Sport the Dutch Ministry of Economic Affairs the University Medical Centre Groningen (UMCG the Netherlands) the University of Groningen and the Northern Provinces of the Netherlands

Conflicts of interest are as follow:
Mikko Kallela: No related COI
Eimear E. Kenny: EEK has received personal fees from Regeneron Pharmaceuticals, 23&Me, Allelica, and Illumina; has received research funding from Allelica; and serves on the advisory boards for Encompass Biosciences, Foresite Labs, and Galateo Bio
Ronald C.W. Ma: No related COI
Benjamin M. Neale: B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora.
Veikko Salomaa: VS has had research collaboration with Bayer Ltd (not related to the present study)
Edwin K. Silverman: Research grants from GSK and Bayer
James S. Ware: JSW has received consultancy fees or grant support from MyoKardia (now Bristol-Myers Squibb), Pfizer, and Foresite Labs