Article

A genomic mutational constraint map using variation in 76,156 human genomes

https://doi.org/10.1038/s41586-023-06045-0

Received: 16 March 2022

Accepted: 3 April 2023

Published online: 6 December 2023

Check for updates

Siwei Chen[1,2,230]✉, Laurent C. Francioli[1,2,230], Julia K. Goodrich[1], Ryan L. Collins[1,3,4], Masahiro Kanai[1,2], Qingbo Wang[1,5], Jessica Alföldi[1,2], Nicholas A. Watts[1,2], Christopher Vittal[1,2], Laura D. Gauthier[6], Timothy Poterba[1,2,7], Michael W. Wilson[1,2], Yekaterina Tarasova[1], William Phu[1,8], Riley Grant[1], Mary T. Yohannes[1], Zan Koenig[2,7], Yossi Farjoun[9], Eric Banks[6], Stacey Donnelly[10], Stacey Gabriel[11], Namrata Gupta[1,11], Steven Ferriera[11], Charlotte Tolonen[6], Sam Novod[6], Louis Bergelson[6], David Roazen[6], Valentin Ruano-Rubio[6], Miguel Covarrubias[6], Christopher Llanwarne[6], Nikelle Petrillo[6], Gordon Wade[6], Thibault Jeandet[6], Ruchi Munshi[6], Kathleen Tibbetts[6], Genome Aggregation Database Consortium*, Anne O'Donnell-Luria[1,3,8], Matthew Solomonson[1,2], Cotton Seed[2,7], Alicia R. Martin[1,2,7], Michael E. Talkowski[1,3,7], Heidi L. Rehm[1,3], Mark J. Daly[1,2,15], Grace Tiao[1,2], Benjamin M. Neale[1,2,230], Daniel G. MacArthur[1,16,17,230] & Konrad J. Karczewski[1,2,7]✉

The depletion of disruptive variation caused by purifying natural selection (constraint) has been widely used to investigate protein-coding genes underlying human disorders[1–4], but attempts to assess constraint for non-protein-coding regions have proved more difficult. Here we aggregate, process and release a dataset of 76,156 human genomes from the Genome Aggregation Database (gnomAD)—the largest public open-access human genome allele frequency reference dataset—and use it to build a genomic constraint map for the whole genome (genomic non-coding constraint of haploinsufficient variation (Gnocchi)). We present a refined mutational model that incorporates local sequence context and regional genomic features to detect depletions of variation. As expected, the average constraint for protein-coding sequences is stronger than that for non-coding regions. Within the non-coding genome, constrained regions are enriched for known regulatory elements and variants that are implicated in complex human diseases and traits, facilitating the triangulation of biological annotation, disease association and natural selection to non-coding DNA analysis. More constrained regulatory elements tend to regulate more constrained protein-coding genes, which in turn suggests that non-coding constraint can aid the identification of constrained genes that are as yet unrecognized by current gene constraint metrics. We demonstrate that this genome-wide constraint map improves the identification and interpretation of functional human genetic variation.

The expansion in the scale of human whole-genome or exome sequencing data has enabled characterization of the patterns of variation in human genes. With these data it is possible to directly assess the strength of negative selection on loss-of-function (LOF) and missense variation by modelling 'constraint', the depletion of variation in a gene compared to an expectation conditioned on that gene's mutability. Using coding variant data from sequencing thousands to hundreds of thousands of human genomes and exomes[5], we and others previously developed constraint metrics that classify each protein-coding gene along a spectrum of LOF and missense intolerance[5–7], providing a valuable resource for studying the functional importance of human genes[1–4]. Although protein-coding regions have fundamental roles in biology, they comprise less than 2% of the human genome, and the vast non-coding genome has been much less characterized, even though the importance of non-coding variation in human complex diseases has been long recognized[8–12].

Several challenges arise when extending the gene constraint model to the non-coding space. First, the sample size of human whole-genome reference data has been relatively small compared with the exome, limiting the power of detecting depletions of variation at a fine scale.

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [4]Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [5]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [6]Data Science Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [9]Richards Lab, Lady Davis Institute, Montreal, Quebec, Canada. [10]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland. [16]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, New South Wales, Australia. [17]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. [230]These authors contributed equally: Siwei Chen, Laurent C. Francioli, Benjamin M. Neale, Daniel G. MacArthur. *A list of authors and their affiliations appears online. ✉e-mail: siwei@broadinstitute.org; konradk@broadinstitute.org

Second, in coding regions, the gene model enables accurate prediction of the effect of specific variants on amino acid translation; such nucleotide-specific models of the consequences of base pair changes are not available in non-coding regions. Third, there is a strong expectation from Mendelian genetics and existing constraint analyses that the coding regions—although they comprise a small fraction of the genome—are grossly overrepresented among rare and common disease mutations under selection. Fourth, the mutation rate in non-coding regions is highly heterogeneous and can be affected by local sequence context as commonly modelled in gene constraint metrics as well as by a variety of genomic features at larger scales[13,14].

Current methods attempting to evaluate non-coding constraint can be broadly divided into three categories: (1) context-dependent mutational models that assess the deviation of observed variation from an expectation based on the sequence composition of *k*-mers (for example, Orion[15], CDTS[16] and DR[17]); (2) machine learning classifiers that are trained to differentiate between disease-associated variants and benign variants (for example, GWAVA[18] and JARVIS[19]); and (3) phylogenetic conservation scores that use comparative genomics data to infer evolutionary constraint (for example, phastCons[20] and phyloP[21]). Although all of these methods aid in our understanding of the non-coding genome, they each have limitations and/or biases, such as overlooking the influence of regional genomic features beyond the scale of flanking nucleotides on mutation rate in method (1) above; a strong dependence on the availability of well-characterized functional mutations as training data in method (2) above; and compromised power to detect regions that have only recently been under selection in the human lineage and may have a functional effect on human-specific traits or diseases in method (3) above.

Here we present Gnocchi, a genome-wide map of human constraint, generated from a high-quality set of variant calls from 76,156 whole-genome sequences (gnomAD v3.1.2 https://gnomad.broadinstitute.org). We describe an improved model of human mutation rates that jointly analyses local sequence context and regional genomic features and quantifies the depletion of variation in tiled windows across the entire genome. Incorporating constraint evidence from functional elements linked to genes can enhance the identification of genes under strong constraint and aid in the functional interpretation of non-coding regions. Our study aims to depict a genome-wide view of how natural selection shapes patterns of human genetic variation and identify which functional genomic elements are likely to harbour variation with potential clinical implications.

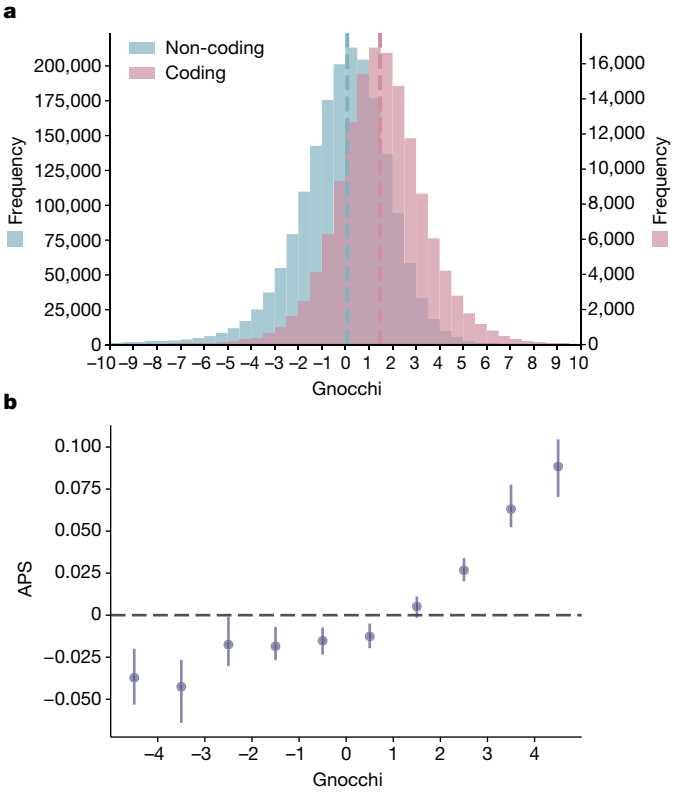


Fig. 1 | Distribution of Gnocchi scores across the genome. a, Histograms of Gnocchi scores for 1,984,900 1-kb windows across the human autosomes. Windows overlapping coding regions (n = 141,341 with ≥1 bp of coding sequence; red) overall exhibit a higher Gnocchi score (stronger negative selection) than windows that are exclusively non-coding (n = 1,843,559; blue). Dashed lines indicate median. **b**, The correlation between Gnocchi score and the APS score developed for structural variation constraint. A collection of 116,184 autosomal structural variations was assessed using Gnocchi by assigning each structural variation the highest Gnocchi score among all overlapping 1-kb windows, which shows a significant positive correlation with the structural variation constraint metric APS. Error bars indicate 100-fold bootstrapped 95% confidence intervals of the mean values.

Aggregating 76,156 whole genomes

We aggregated, reprocessed and performed joint variant calling on 153,030 whole genomes mapped to human genome reference build GRCh38, of which 76,156 samples were retained as high-quality sequences from unrelated individuals without known severe paediatric disease, and with appropriate consent and data use permissions for the sharing of aggregate variant data (Supplementary Figs. 1–5 and Supplementary Tables 1–3). Among these samples, 36,811 (48.3%) have non-European ancestry, including 20,744 individuals with African ancestry and 7,647 individuals with admixed American Indigineous ancestries. After stringent quality control, we identified a set of 644,267,978 high-confidence short nuclear variants (single nucleotide or insertion–deletion variants; gnomAD v3.1.2), of which 390,393,900 low-frequency (allele frequency (AF) ≤ 0.1%), high-quality single nucleotide variants were used to build the genome-wide constraint map. These correspond to approximately 1 variant every 4.9 bp (one low-frequency variant every 8 bp) of the genome, providing a high density of variation.

Gnocchi quantifies genomic constraint

To construct a genome-wide mutational constraint map, we divided the genome into continuous non-overlapping 1-kb windows, and quantified constraint for each window by comparing the expected and the observed variation in our gnomAD dataset. Here we implemented a refined mutational model, which incorporates trinucleotide sequence context, base-level methylation and regional genomic features to predict expected levels of variation under neutrality. In brief, we estimated the relative mutability for each single nucleotide substitution with one base of adjacent nucleotide context (for example, ACG>ATG), with adjustment for the effect of methylation on mutation rate at CpG sites, which become saturated for mutation at sample sizes above 10,000 genomes[22] (Extended Data Fig. 1a,b and Supplementary Fig. 6; Methods). Meanwhile, we adjusted the effects of regional genomic features for each trinucleotide mutation rate on the basis of the occurrence of de novo mutations (DNMs) (n = 413,304 previously detected in family-based whole-genome sequencing studies[23,24]; Extended Data Fig. 1c), and then applied it to establish the expected number of variants per 1 kb across the entire genome (Methods).

We quantified the deviation from expectation for each 1-kb window using a *Z*-score[7]—hereafter referred to as 'Gnocchi' (Extended Data Fig. 1d,e; Methods)—which was centred around zero for non-coding regions (median = 0.08), and was significantly higher (more constrained) for windows containing any protein-coding sequences (median = 1.47, Wilcoxon $P < 10^{-200}$; Fig. 1a). Gnocchi is positively correlated with the percentage of coding bases in a window and

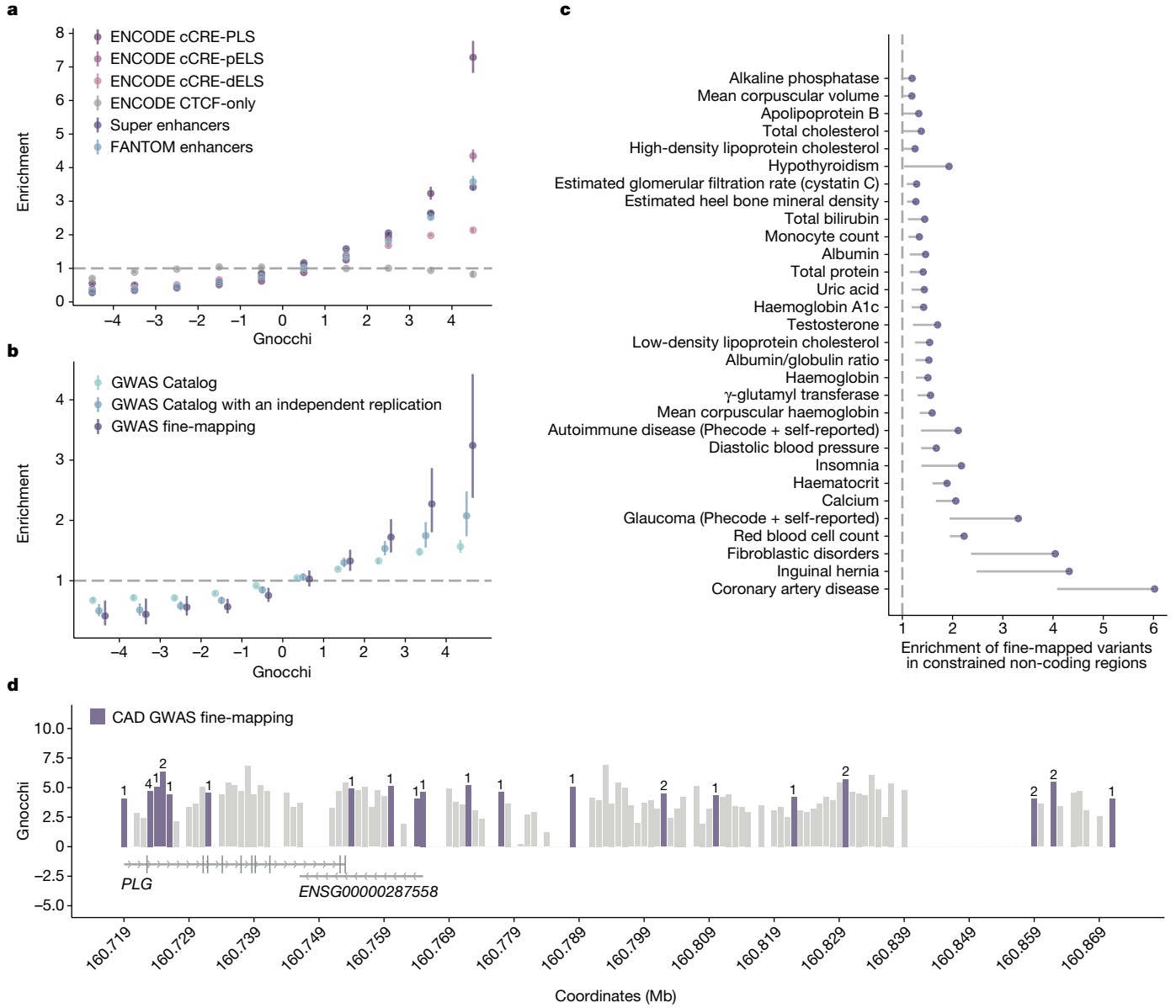


Fig. 2 | Correlation between Gnocchi and functional non-coding annotations. **a**,**b**, Distributions of candidate regulatory elements (**a**) and GWAS variants (**b**) along the spectrum of Gnocchi in non-coding regions. Enrichment was evaluated by comparing the proportion of non-coding 1-kb windows, binned by Gnocchi, that overlap with a given functional annotation to the genome-wide average. Error bars indicate 95% confidence intervals of the odds ratios. **a**, cCRE: $n = 34{,}803$ with a promoter-like signature (PLS), $n = 141{,}830$ with a pELS, $n = 667{,}599$ with a dELS, $n = 56{,}766$ CTCF-only. Super enhancers: $n = 331{,}601$; FANTOM enhancers: $n = 63{,}285$. **b**, GWAS Catalog: $n = 111{,}308$ variants with an association $P \leq 5.0 \times 10^{-8}$, $n = 9{,}229$ with an independent replication. GWAS fine-mapping: $n = 2{,}191$ variants fine-mapped with posterior inclusion probability of causality ≥ 0.9. See Methods for details on data collection. **c**, Enrichment of fine-mapped variants in constrained non-coding regions (Gnocchi ≥ 4). Credible set–trait pairs with a significant enrichment are shown, ordered by the lower bound of 95% confidence interval; only lower bounds are shown for presentation purposes. **d**, The distribution of variants fine-mapped for CAD in constrained regions (Gnocchi ≥ 4) of *PLG*. Each bar shows the Gnocchi score of a 1-kb window (gaps indicate windows removed by quality filters); windows containing fine-mapped variants are represented in purple, and the number of variants in each window is annotated on top of the corresponding bar. Ten variants are located within *PLG* introns, 4 are mapped to the antisense gene of *PLG* (ENSG00000287558), and 14 reside in the downstream intergenic regions.

presented a substantial shift towards higher constraint for exonic sequences from directly concatenating coding exons into 1-kb windows (median = 3.17; Extended Data Fig. 2a–c). About 3.12% and 0.05% of the non-coding windows exhibited constraint as strong as the 50th percentile and 90th percentile of exonic regions, respectively (Extended Data Fig. 2d). Comparing Gnocchi against the adjusted proportion of singletons (APS) score, a measure of constraint developed for structural variation[25], we found a significant correlation (linear regression beta = 0.01, $P = 4.3 \times 10^{-65}$; Fig. 1b, Methods), providing an internal validation of our approach.

Gnocchi highlights non-coding function

To further validate the Gnocchi metric and investigate the functional relevance of non-coding regions under selection, we examined the correlation between Gnocchi and several annotations of functional non-coding sequences (Fig. 2a). First, we found that candidate *cis*-regulatory elements (cCREs) (derived from ENCODE[26] integrated DNase-based sequencing (DNase-seq) and chromatin immunoprecipitation with sequencing (ChIP–seq) data) are significantly enriched in the most constrained percentile of the genome (Gnocchi ≥ 4, odds

ratio (OR) = 2.77 compared to the genome-wide average, Fisher's exact $P < 10^{-200}$); cCREs with a promoter-like signature (cCRE-PLS) presented the strongest enrichment (OR = 7.28), followed by elements with a proximal enhancer-like signature (pELS) (OR = 4.35) or distal enhancer-like signature (dELS) (OR = 2.14), and as a negative control, elements bound by CTCF but not associated with a regulatory signature showed no enrichment (CTCF-only) (OR = 0.82). These patterns indicate that a large fraction of the constrained non-coding regions may serve a regulatory role, in line with previous findings[15,16,19]. Similarly, significant enrichment was found for an independent set of active, in vivo-transcribed enhancers (identified by FANTOM CAGE analyses[27]) (OR = 3.58) and super enhancers[28] (OR = 3.41), which are groups of enhancers in close genomic proximity regulating genes important for cell-type specification[29]. By aggregating the regulatory annotations, we estimated that approximately 10.4% and 6.3% of promoters and enhancers, respectively, are under selection as strong as the average constraint for coding exons (Extended Data Fig. 3a; Methods). A much higher proportion, 22.2%, was found for sequences encoding microRNAs (miRNAs), which are increasingly recognized as key mediators in various developmental and physiological processes[30]. By contrast, only 3.7% of long non-coding RNAs (lncRNAs) exhibited such strong constraint, similar to that of non-coding regions overall (3.1%; Extended Data Figs. 2d and 3b).

We next examined the distribution of putatively functional non-coding variants on the constraint spectrum. There was significant enrichment for non-coding variants implicated by genome-wide association studies (GWASs) in the constrained end of the genome: 837 out of 19,471 constrained windows (Gnocchi ≥ 4) overlapped with GWAS Catalog[31] annotations (OR = 1.57 compared with the genome-wide average of 51,430 out of 1,843,559, Fisher's exact $P = 2.5 \times 10^{-32}$; Fig. 2b, Methods). The enrichment became stronger when restricted to the subset of variants that had been replicated by an independent study (OR = 2.08, $P = 4.1 \times 10^{-13}$). Moreover, further strong signals were found for probably causal GWAS variants fine-mapped for 148 complex diseases and traits in large-scale biobanks[32] (OR = 3.24, $P = 3.0 \times 10^{-10}$; Methods). Across the 95% credible set–trait pairs, strong enrichment was predominantly seen in disease phenotypes, including coronary artery disease (CAD), inguinal hernia, fibroblastic disorders and glaucoma (ORs 3.31–6.02; Fig. 2c, Methods). In the 95% credible set of CAD, for instance, the highest Gnocchi score was found for rs1897107 and rs1897109 (both within the same genomic window chromosome (chr.) 6:160725000–160726000, Gnocchi = 6.32); high constraint (Gnocchi ≥ 4) was also found for 26 variants from the same credible set (a total of 28 out of 52), which together spanned a sequence of approximately 153 kb downstream of the gene *PLG* (Fig. 2d). *PLG* encodes the plasminogen protein that circulates in blood plasma and is converted to plasmin to dissolve the fibrin of blood clots. Although dysregulation of the PLG–plasmin system has been frequently associated with CAD[33–38], no specific variants in *PLG* have been implicated. Our results prioritized a set of non-coding variants in highly constrained regions of *PLG*, which adds quantitative evidence to the implication of *PLG* in CAD and may help direct or prioritize follow-up functional experiments.

Collectively, these results demonstrated a significant positive correlation between constraint and functional non-coding annotations, illustrating the utility of Gnocchi in characterizing non-coding regions. However, we suggest that Gnocchi provides additional information to existing annotations. For instance, prioritizing ENCODE cCREs by Gnocchi revealed increasingly stronger GWAS enrichment in the more constrained cCREs (Extended Data Fig. 4a), and constrained regions outside cCREs also captured significant signals, reflecting the value of Gnocchi independent of regulatory annotations. Moreover, besides prioritizing existing GWAS results, Gnocchi can be used as a prior for statistical fine-mapping. Using UK Biobank traits as examples, incorporating Gnocchi into the functionally informed fine-mapping model[39] predicted around 13,000 variant–trait pairs to have an increased posterior inclusion probability of causality (ΔPIP) ≥ 0.01, in which 164 likely causal associations were newly identified at PIP ≥ 0.8 (Extended Data Fig. 4b; Methods). Although only functional tests can ultimately validate the underlying causality, our constraint map presents a valuable resource for expanding or refining the catalogue of functional non-coding variants in the human genome.

Gnocchi versus other non-coding metrics

To benchmark the performance of Gnocchi in prioritizing non-coding variants, we extended the analyses of GWAS variants to compare it with other population genetics-based constraint metrics (Orion[15], CDTS[16], gwRVIS[19] and DR[17]). Specifically, we assessed the performance of different metrics in identifying putative functional non-coding variants (hereafter referred to as 'positive' variant set)—as previously mentioned: (1) GWAS Catalog[31] variants (n = 9,229 with an independent replication); (2) GWAS fine-mapping[32] variants (n = 2,191); (3) a subset of high-confidence causal variants from (2) (n = 140); and (4) likely pathogenic Mendelian variants (n = 1,026 from ClinVar[40] and the Human Gene Mutation Database (HGMD)[41])—against background variants in the population with a similar allele frequency ('negative' variant set; Methods). Overall, Gnocchi achieved the highest performance across all comparisons, as measured by the area under the curve (AUC) statistic (Fig. 3a,b and Extended Data Fig. 5). The performance was also more stable than others when varying the allele frequency threshold for the negative variant set (Extended Data Fig. 5). This may be owing to other metrics being informed by the site frequency spectrum, which made the classification performance sensitive to differences in allele frequency between the positive and negative variants. We also showed that our performance was robust to the artificial break of genomic windows (non-overlapping 1 kb) by reconstructing Gnocchi scores in a sliding window (1 kb stepped by 100 bp) approach as adopted by other metrics (Extended Data Fig. 6).

Extending the comparison to include phylogeny-based conservation scores (phyloP[21], phastCons[20] and GERP[42]) revealed relatively low performance compared to the population genetics-based constraint metrics (Fig. 3a,b). The conservation scores were weakly correlated with constraint (Spearman's rank correlation coefficient 0.017–0.19; Extended Data Fig. 7), suggesting that intraspecies (human lineage-specific) constrained regions complement, rather than reflect a subset of, regions that are conserved across species. Each individual metric also contributed to the classification when modelled as independent predictive variables (Fig. 3c,d; Methods), reinforcing the complementary nature of different approaches. Variants that were uniquely captured by Gnocchi, for instance, tended to be in regions with high recombination rates (3.45-fold versus the rest of the positive variant set) and high DNA methylation (2.74-fold; Methods), both associated with an increased mutation rate that had been adjusted in our refined mutational model. To further illustrate this improvement, we rebuilt our constraint model from solely the local sequence context—that is, without adjustment on mutation rate by regional genomic features—and confirmed that Gnocchi outperformed such metrics (Extended Data Fig. 6). Altogether, we demonstrate that Gnocchi is an effective metric for identifying functional variants in the non-coding genome; at the same time, we suggest that a combination of different metrics is likely to provide the most informative results for prioritizing functional variation.

Gnocchi prioritizes copy number variants

Besides single-nucleotide variants that have been extensively studied in GWASs, copy-number variants (CNVs) that cause dosage alterations (deletions causing loss or duplications causing gain) of DNA represent another

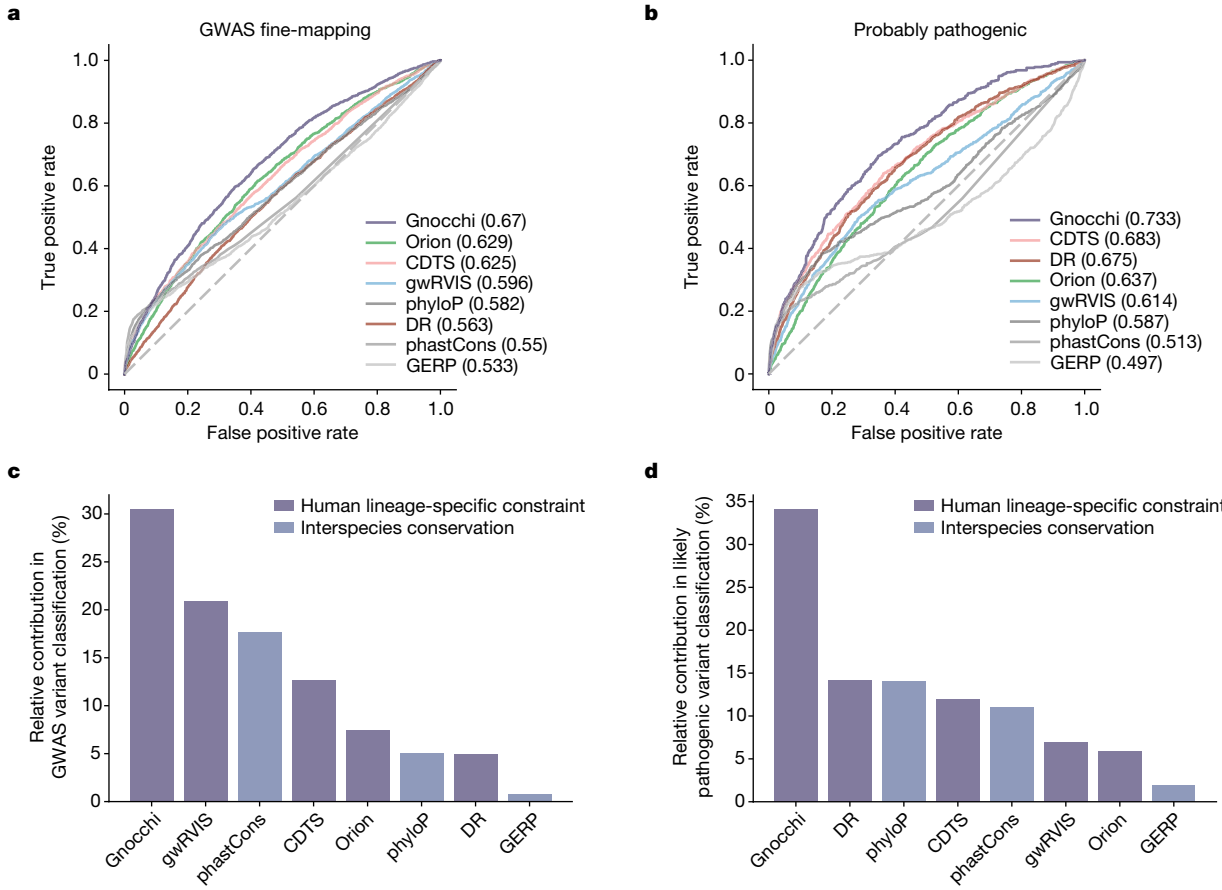


Fig. 3 | Performance of Gnocchi and other predictive metrics in prioritizing non-coding variants. a,**b**, Receiver operating characteristic (ROC) curves of Gnocchi and seven other metrics in classifying putative functional non-coding variants—2,191 GWAS fine-mapping variants (**a**) and 1,026 likely pathogenic variants (**b**)—against background variants in the population. The performance of each metric was measured and ranked by the AUC statistic. **c**,**d**, The relative contribution of different metrics in classifying GWAS variants (**c**) and likely pathogenic variants (**d**). The eight metrics were modelled as eight independent predictors for the classification, and the relative contribution of one predictor over another was evaluated by estimating their additional R^2 contributions across all subset models.

important class of variation for contributing variability in risk for human disease[43–48]. Yet, unlike single nucleotide variants, CNVs can be large and determining the 'minimal critical region'[49] with a pathogenic effect has been a major challenge. Although CNVs primarily affect non-coding sequences, the most commonly studied mechanism is still the dosage alteration of overlapping protein-coding genes[50]. Using our genome-wide constraint map, we explored the possibility that constrained non-coding regions are also sensitive to a dosage effect, which may underlie the pathogenicity of corresponding CNVs.

We surveyed a collection of around 100,000 CNVs from a genome-wide CNV morbidity map of developmental delay (DD)[51,52]. There was a substantial excess of CNVs that affected constrained non-coding regions (Gnocchi ≥ 4) among individuals with DD compared with healthy controls (42.6% versus 12.5%, OR = 5.21, Fisher's exact $P < 10^{-200}$; Fig. 4a, Methods). Moreover, out of the 19 loci that had been previously identified as pathogenic[51], all but one (94.7%) affected constrained non-coding regions; the high incidence was recapitulated in a curated set of around 4,000 putative pathogenic CNVs (85.5% in ClinVar[40]; Fig. 4a). Notably, the case–control enrichment remained significant, albeit attenuated, after adjusting for the size and gene content of each CNV and when being tested in the subset of CNVs that are exclusively non-coding (Fig. 4b; Methods). Non-coding constraint presented high association with DD CNVs conditioning on gene constraint (log(OR) = 1.06, logistic regression $P < 10^{-100}$), lending support to the possibility that dosage alteration of constrained non-coding regions may be an alternative explanation for the mechanism of CNVs underlying DDs.

One known example of pathogenic non-coding dosage alteration is the duplication of the *IHH* regulatory domain in synpolydactyly and craniosynostosis[53–55]. The four implicated duplications covered approximately 102 kb of sequence upstream of *IHH*, with an overlapping region of approximately 10 kb (the 'critical region'[49]; Fig. 4c). The region contained no genes but exhibited high levels of constraint (median Gnocchi = 2.52, Wilcoxon $P = 1.3 \times 10^{-3}$ compared with the rest of the genome). The most constrained window (chr. 2:219111000–219112000, Gnocchi = 4.12) overlapped with the major enhancer of *IHH*, the duplication of which has been shown to result in dosage-dependent *IHH* misexpression and consequently syndactyly and malformation of the skull[55]. This result highlights a potential use of the Gnocchi metric to prioritize non-coding regions within large CNVs. As a further illustration, we examined a set of non-coding CNVs that had the highest Gnocchi score among the DD cases. The most constrained genomic window (chr. 11:133208000–133209000, Gnocchi = 8.87) was affected by 12 deletions spanning a non-coding sequence of around 400 kb (Fig. 4d). Although they varied in size, the deletions shared a common region of approximately 20 kb (the potential critical region), which encompassed the most constrained window and overall, showed a significantly higher constraint than the other affected regions (median Gnocchi = 1.63 versus 0.84, Wilcoxon $P = 1.6 \times 10^{-3}$; Fig. 4d). In addition, the approximately 400-kb sequence also contained two deletions from healthy controls,

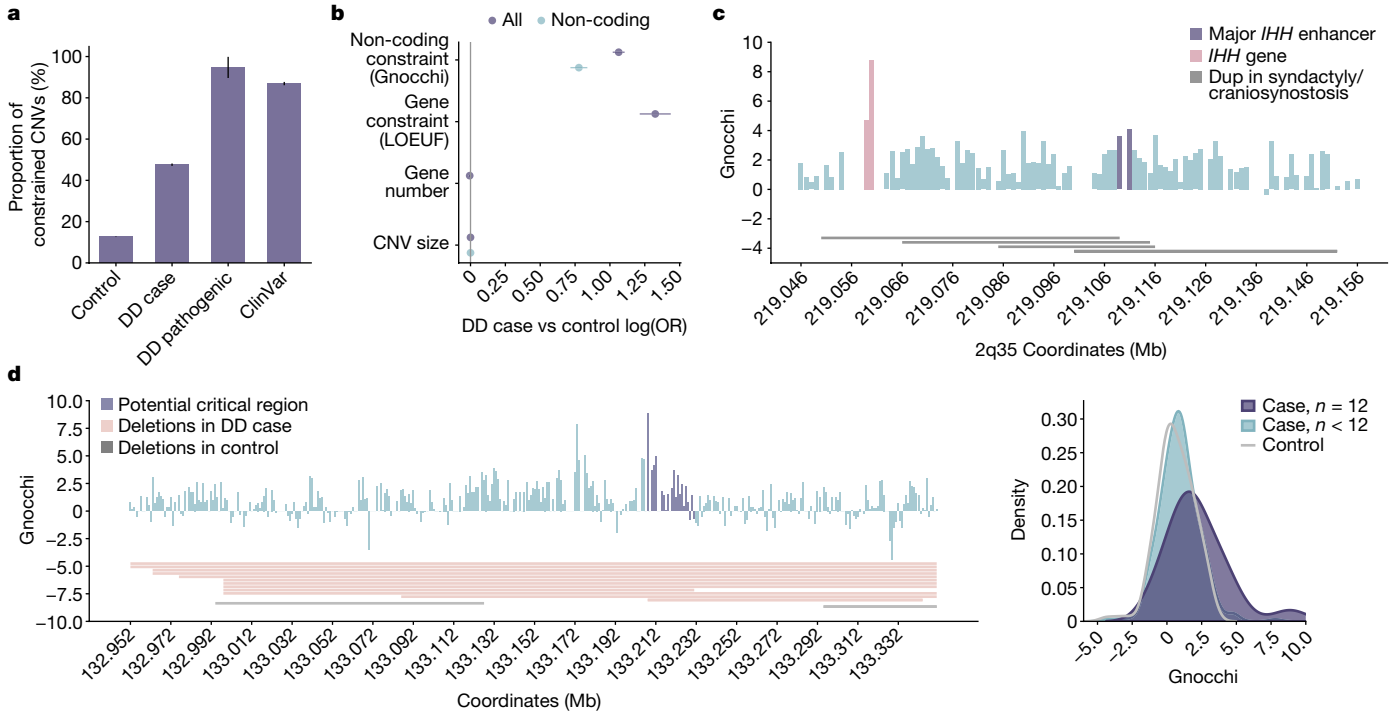


Fig. 4 | Contribution of non-coding constraint in evaluating CNVs. **a**, Proportions of constrained CNVs (Gnocchi ≥ 4) identified in individuals with DD versus healthy controls. Constrained CNVs are more common in DD than in controls (7,239 out of 17,004 (42.6%) versus 10,403 out of 83,526 (12.5%)) and are most frequent for CNVs previously implicated as pathogenic (18 out of 19 (94.7%) by DD and 3,433 out of 4,014 (85.5%) by ClinVar). Error bars indicate standard errors of the proportions. **b**, Contribution of non-coding constraint to predicting CNVs in DD cases versus controls. Non-coding constraint remains a significant predictor for the case or control status of CNVs after adjusting for gene constraint (LOEUF score), gene number and size of CNVs (cases, $n = 17{,}004$; controls, $n = 83{,}526$; purple), as well as being tested in the subset of non-coding CNVs (cases: $n = 8{,}702$; controls, $n = 66{,}795$; blue). Error bars indicate 95% confidence intervals of the log odds ratios. **c**, CNVs at the *IHH* locus associated with synpolydactyly and craniosynostosis. The 4 implicated duplications (grey horizontal bars) span an approximately 102-kb sequence upstream of *IHH*. Each vertical bar shows the Gnocchi score of a 1-kb window within the locus, with the highest score overlapping the *IHH* gene (red) and the highest non-coding score overlapping the major *IHH* enhancers (purple); gaps indicate windows removed by quality filters. **d**, Non-coding CNVs with the highest Gnocchi score identified in DD cases. Left, the highest-scored window is located within the potential critical region (purple vertical bars) shared by 12 DD deletions (red horizontal bars); grey indicates two deletions observed in controls. Right, kernel density estimate plot showing that the critical region overall has a significantly higher Gnocchi score than the other regions affected by DD or control deletions.

which notably overlapped with the two lowest Gnocchi scores within the region and were significantly less constrained than those from DD cases (median Gnocchi = 1.07 versus 0.62, Wilcoxon $P = 4.74 \times 10^{-4}$). These findings suggest that Gnocchi can be a useful indicator of critical regions affected by large CNVs, facilitating the interpretation of non-coding risk factors in CNV disease association studies.

Gnocchi informs gene function

Given the important role of non-coding regions in gene regulation, it is natural to expect that more constrained regulatory elements would regulate more constrained genes. To test this, we analysed the constraint for enhancers that had been linked to specific genes[56] (Methods). More constrained non-coding regions were more frequently linked to regulating a gene (Fig. 5a), and as expected, enhancers linked to constrained genes (predicted by LOF observed/expected upper bound fraction[5] (LOEUF) or curated disease genes from refs. 57–59; Methods) were significantly more constrained than those linked to presumably less constrained genes (median Gnocchi = 2.71 versus 1.99, Wilcoxon $P = 1.3 \times 10^{-26}$; Fig. 5b, Methods), thus supporting a correlated constraint between genes and their regulatory elements.

Conversely, the links between constrained enhancers and the 'unconstrained' genes predicted by LOEUF may reflect functional importance of the unconstrained genes that had previously been unrecognized. The lack of predicted gene constraint can be explained by the design of LOEUF as a measure of intolerance to rare LOF variation, where small genes with few expected LOF variants are probably underpowered. Indeed, stratifying genes by the number of expected LOF variants showed a significantly higher enhancer constraint for genes that were underpowered[5] (≤5 expected LOF variants) than for genes that were sufficiently powered but scored as unconstrained (median Gnocchi = 2.64 versus 2.27, Wilcoxon $P = 9.8 \times 10^{-4}$; Fig. 5b). This suggests that certain underpowered genes may be functionally important but were not recognized in gene constraint evaluation. For instance, *ASCL2*, a basic helix-loop-helix transcription factor, had only 0.57 expected LOF variants (versus 0 observed) across more than 125,000 exomes[5]; although depleted for LOF variation, the absolute difference was too small to obtain a precise estimate of LOF intolerance. Nevertheless, we found that *ASCL2* had a highly constrained enhancer (Gnocchi = 5.58), located about 16 kb upstream of the gene, where more than 40% of the expected variants were depleted (188.6 expected versus 112 observed, chr. 11:2286000–2287000). The same genomic window also contained an expression quantitative trait locus (chr. 11:2286192:G>T) that was predicted to be significantly associated with *ASCL2* expression[60], and increased *ASCL2* expression has been implicated in the development and progression of several human cancers[61–63]. This example highlights the value of non-coding constraint as a complementary metric to gene constraint for identifying functionally important genes.

A practical implementation of this finding is to integrate the constraint of regulatory elements into the modelling of gene constraint,

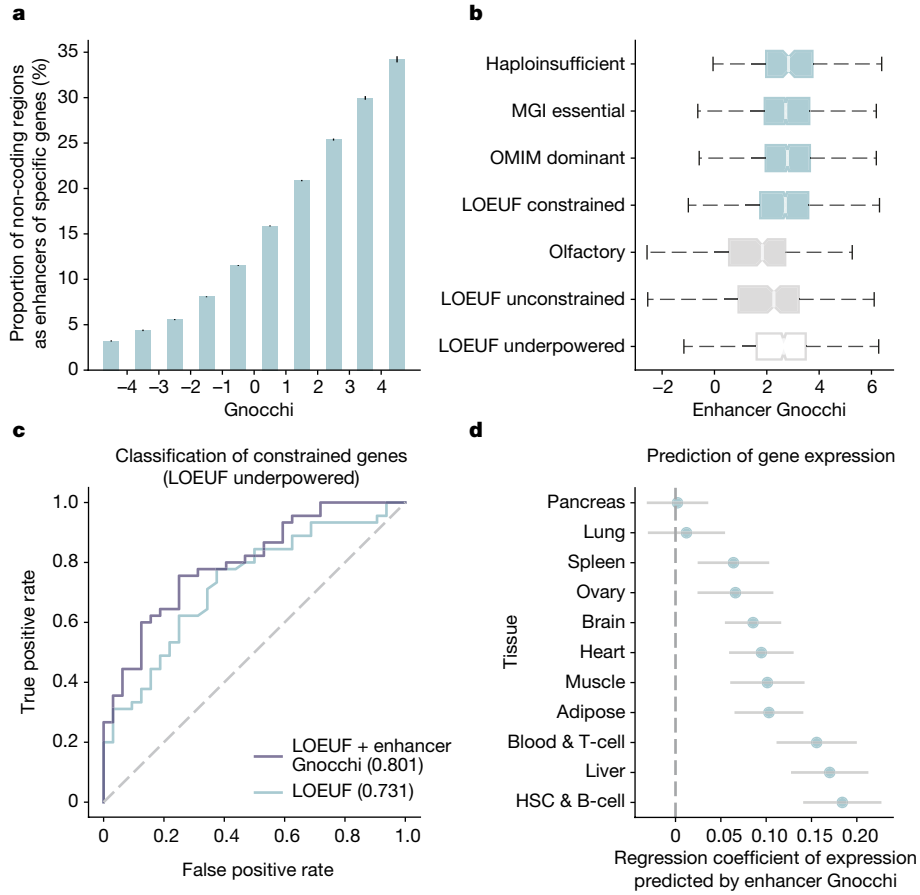


Fig. 5 | Correlation of constraint between non-coding regulatory elements and protein-coding genes. a, The proportion of non-coding 1-kb windows overlapping with enhancers that were predicted to regulate specific genes, as a function of their Gnocchi scores. More constrained non-coding regions are more frequently linked to a gene (left to right: n = 2,022 out of 62,894, 2,743 out of 62,653, 7,475 out of 134,279, 20,383 out of 252,354, 43,414 out of 376,829, 66,343 out of 417,743, 65,343 out of 313,110, 38,785 out of 152,787, 15,417 out of 51,439 and 6,663 out of 19,471). Error bars indicate standard errors of the proportions. **b**, Comparison of the Gnocchi scores of enhancers linked to constrained and unconstrained genes. Enhancers of established sets of constrained genes (blue boxes: n = 189 haploinsufficient genes, n = 2,454 MGI essential genes, n = 1,771 OMIM autosomal dominant disease genes, n = 1,920 LOEUF-predicted constrained genes) are more constrained than enhancers of presumably less constrained genes (grey boxes: n = 356 olfactory receptor genes, n = 189 LOEUF-predicted unconstrained genes). Enhancers of genes that are underpowered for gene constraint detection (LOEUF underpowered, n = 1,117) present a higher constraint than those powered yet unconstrained genes (LOUEF unconstrained). Box plots show the distribution of Gnocchi scores of enhancers linked to different gene sets, denoting the median, quartiles and range (excepting outliers). **c**, Improvement of incorporating enhancer constraint into LOEUF in prioritizing underpowered genes. ROC curves and AUCs show the performance of two logistic regression models using LOEUF and LOEUF plus enhancer Gnocchi score as independent predictive variables to classify constrained and unconstrained genes, tested on a set of 77 underpowered genes. **d**, Contribution of enhancer constraint to predicting gene expression in specific tissue types. The x-axis shows the linear regression coefficient of tissue-specific enhancer Gnocchi score predicting the expression level of target genes in matched tissue types (haematopoietic stem cell (HSC) and B cell, n = 11,970; brain, n = 11,555; heart, n = 10,759; pancreas, n = 10,572; blood and T cell, n = 10,403; muscle, n = 10,380; adipose, n = 9,316; liver, n = 8,838; spleen, n = 8,308; ovary, n = 7,926; lung, n = 7,499), conditioning on gene constraint (LOEUF score). Error bars indicate 95% confidence intervals of the coefficient estimates.

which essentially gains power from extending the functional unit of a gene to encompass its regulatory components. As a proof of principle, we tested whether adding the Gnocchi score of enhancer to LOEUF improves the prioritization of underpowered genes. The enhancer Gnocchi score was found to be a significant predictor of constrained genes (logistic regression $P = 7.4 \times 10^{-11}$ conditioning on LOEUF) and improved the performance of LOEUF in identifying constrained genes that were underpowered (AUC = 0.80 versus 0.73, bootstrap $P = 0.03$; Fig. 5c, Methods). Moreover, such approaches would enable incorporation of tissue- and cell-type-specific information into gene constraint modelling, given the diverse range of epigenomic data. We explored this by testing whether the constraint of tissue-specific enhancers is predictive of tissue-specific gene expression (as a proxy for tissue-specific gene function). The enhancer Gnocchi score, again conditioning on LOEUF, was a significant predictor of the expression level of target genes in matched tissue types (Fig. 5d; Methods). These results further support the application of the Gnocchi metric for improving the characterization of gene function. Although we acknowledge that the biological consequences of mutations in enhancers are not clearly understood and thus natural selection may differ in strength depending on mechanistic consequence, an extended model to incorporate non-coding variation information in a biologically informed way holds promise to facilitate our understanding of the molecular mechanisms underlying selection.

Discussion

We have previously developed constraint metrics that leverage population-scale exome and genome sequencing data to evaluate genic intolerance to coding variation for each protein-coding gene[5,22]. Here we adopted the same principle with an extended mutational model to

assess constraint across the entire genome, using our latest release of gnomAD (v3.1.2), a dataset of harmonized high-quality whole-genome sequences from 76,156 individuals of diverse ancestries. Improvements to constraint modelling include unified fitting of the mutation rate for all substitution and trinucleotide contexts and inclusion of regional genomic features to refine the expected variation in non-coding regions (Methods). We validated our metric—called Gnocchi—using a series of external functional annotations, with a focus on the non-coding genome, and demonstrated its value for prioritizing non-coding elements and identifying functionally important genes. We have made the Gnocchi scores publicly accessible via the gnomAD browser (https://gnomad.broadinstitute.org).

One key challenge in quantifying non-coding constraint is the estimation of the true base mutation rate, which can be affected by various genomic phenomena, potentially operating at different scales. To this end, we extended our previous mutational model, which computed the relative mutability of each substitution in a trinucleotide context, to include adjustments for regional genomic features that may index processes influencing mutagenesis. The adjustment was applied to each specific trinucleotide context and enabled a varying genomic scale for each specific feature (Methods). The added value of this adjustment was demonstrated by the improved performance of Gnocchi in identifying functional variants (Extended Data Fig. 6). Gnocchi also outperformed other genome-wide predictive scores, and each metric tended to provide complementary information. We note that all comparisons were restricted to non-coding regions for explicitly evaluating the metrics in prioritizing non-coding variants, and we further eliminated potential bias from nearby genes by recapitulating the results within regions more than 10 kb away from any protein-coding exons (Supplementary Fig. 7). Overall, Gnocchi presented consistent, high performance in identifying functional non-coding variants in the human genome.

Despite the clear constraint signal identified for non-coding regions, many limitations exist. First, the lack of prior classification of the molecular consequences of non-coding variants—analogous to the classification of 'nonsynonymous' versus 'synonymous' informed by the genetic code in coding regions—limits the resolution of non-coding constraint assessment (for example, to measure constraint against LOF variation). Whereas there are rich resources defining regulatory elements in the non-coding genome, no method is available for determining the effect of each possible variant on gene regulation and the distribution of their effect sizes genome-wide. Further, the interpretation of non-coding constraint, especially in the context of gene regulation, can only be informative when considered in a particular context, such as a tissue or cell type, developmental stage or environment. Such information is not inherently built into our constraint metric nor in the mutational dataset; thus ad hoc integration of external annotations (for example, tissue-specific enhancers as analysed in this study) is often necessary for justifying specific biological implications. Also, since the detection of depletion of variation is immune to negative selection after reproductive age, genomic regions involved in late-onset phenotypes are likely to go underrecognized.

Finally, although this is among the largest datasets of human genomes examined to date for non-coding constraint, our method will substantially increase in power and resolution as sample sizes increase. Benchmarking on the depletion of variation seen in coding regions, we are currently well-powered to detect extreme non-coding constraint as strong as the 90th percentile of coding exons of similar size, and we estimate a sample size of around 340,000 genomes to detect constraint as strong as the 50th percentile of similarly sized coding exons (Extended Data Fig. 8a; Methods). Much larger sample sizes will be required to further increase the resolution, for instance from 1 kb to a 100 bp scale, we would need about 5.3 million samples (Extended Data Fig. 8b). With the current sample size, 1 kb presented optimal performance when compared with varying window sizes tested from 100 bp to 3 kb (Extended Data Fig. 8c). Meanwhile, we emphasize the importance of increasing genetic ancestral diversity in population-scale datasets such as gnomAD. A more diverse population would identify a larger number of rare variants, thereby increasing the power of detecting depletions of variation. We explicitly demonstrated this by reconstructing Gnocchi from the European population subset and comparing it to that from an equal-sized subset containing all populations—the latter achieved a higher predictive power (Extended Data Fig. 8d). Future efforts towards a larger, more diverse human reference dataset would enable improved studies of the influence of human demography on constraint metrics, facilitating a fuller understanding of the distribution and effect of human genetic variation.

Overall, our study demonstrates the value of the genome-wide constraint map in characterizing both non-coding regions and protein-coding genes, providing a critical step towards a comprehensive catalogue of functional genomic elements for humans.

Note added in proof: In an accompanying article[64], we describe an approach to statistically phase rare variant pairs and apply it to the 125,748 exome sequenced individuals in gnomAD v2, creating a resource to aid in variant interpretation especially in the context of recessive diseases.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06045-0.

1. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
2. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584. e523 (2020).
3. Singh, T. et al. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. *Nat. Genet.* **49**, 1167–1173 (2017).
4. Ganna, A. et al. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018).
5. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
6. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
7. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
8. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
9. Lanyi, J. K. Photochromism of halorhodopsin. *cis*/*trans* isomerization of the retinal around the 13–14 double bond. *J. Biol. Chem.* **261**, 14025–14030 (1986).
10. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered *cis*-regulatory elements in human disease. *Trends Genet.* **31**, 67–76 (2015).
11. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* **25**, R157–R165 (2016).
12. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).
13. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat. Rev. Genet.* **22**, 672–686 (2021).
14. Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030–1035 (2021).
15. Gussow, A. B. et al. Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS ONE* **12**, e0181604 (2017).
16. di Iulio, J. et al. The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
17. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
18. Ritchie, G. et al. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
19. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.* **12**, 1504 (2021).
20. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

21. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
22. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
23. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
24. An, J. Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
25. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
26. The ENCODE Project Consortium. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
27. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
28. Jiang, Y. et al. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.* **47**, D235–D243 (2019).
29. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Genet.* **47**, 8–12 (2015).
30. Bartel, D. P. Metazoan microRNAs. *Cell* **173**, 20–51 (2018).
31. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
32. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at *medRxiv* https://doi.org/10.1101/2021.09.03.21262975 (2021).
33. Jung, R. G. et al. Association between plasminogen activator inhibitor-1 and cardiovascular events: a systematic review and meta-analysis. *Thromb. J.* **16**, 12 (2018).
34. Song, C., Burgess, S., Eicher, J. D., O'Donnell, C. J. & Johnson, A. D. Causal effect of plasminogen activator inhibitor type 1 on coronary heart disease. *J. Am. Heart Assoc.* **6**, e004918 (2017).
35. Schaefer, A. S. et al. Genetic evidence for PLASMINOGEN as a shared genetic risk factor of coronary artery disease and periodontitis. *Circ. Cardiovasc. Genet.* **8**, 159–167 (2015).
36. Li, Y. Y. Plasminogen activator inhibitor-1 4G/5G gene polymorphism and coronary artery disease in the Chinese Han population: a meta-analysis. *PLoS ONE* **7**, e33511 (2012).
37. Drinane, M. C., Sherman, J. A., Hall, A. E., Simons, M. & Mulligan-Kehoe, M. J. Plasminogen and plasmin activity in patients with coronary artery disease. *J. Thromb. Haemost.* **4**, 1288–1295 (2006).
38. Lowe, G. D. et al. Tissue plasminogen activator antigen and coronary heart disease. Prospective study and meta-analysis. *Eur. Heart J.* **25**, 252–259 (2004).
39. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of *cis*-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
40. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
41. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
42. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
43. Greenway, S. C. et al. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.* **41**, 931–935 (2009).
44. Mefford, H. C. et al. Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am. J. Hum. Genet.* **81**, 1057–1069 (2007).
45. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
46. Stefansson, H. et al. Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
47. Walsh, T. et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
48. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
49. Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
50. Spielmann, M. & Mundlos, S. Structural variations, the regulatory landscape of the genome and their alteration in human disease. *Bioessays* **35**, 533–543 (2013).
51. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
52. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
53. Klopocki, E. et al. Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am. J. Hum. Genet.* **88**, 70–75 (2011).
54. Barroso, E. et al. Identification of the fourth duplication of upstream IHH regulatory elements, in a family with craniosynostosis Philadelphia type, helps to define the phenotypic characterization of these regulatory elements. *Am. J. Med. Genet. A* **167A**, 902–906 (2015).
55. Will, A. J. et al. Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat. Genet.* **49**, 1539–1545 (2017).
56. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
57. Rehm, H. L. et al. ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
58. Blake, J. A. et al. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39**, D842–D848 (2011).
59. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
60. Consortium, G. T. The Genotype–Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
61. Xu, H. et al. Elevated ASCL2 expression in breast cancer is associated with the poor prognosis of patients. *Am. J. Cancer Res.* **7**, 955–961 (2017).
62. Jubb, A. M. et al. Achaete-scute like 2 (ascl2) is a target of Wnt signalling and is upregulated in intestinal neoplasia. *Oncogene* **25**, 3445–3457 (2006).
63. Tian, Y. et al. MicroRNA-200 (miR-200) cluster regulation by achaete scute-like 2 (Ascl2): impact on the epithelial-mesenchymal transition in colon cancer cells. *J. Biol. Chem.* **289**, 36101–36115 (2014).
64. Guo, M. H. et al. Inferring compound heterozygosity from large-scale exome sequencing data. *Nat. Genet.* https://doi.org/10.1038/s41588-023-01608-3 (2023).

Genome Aggregation Database Consortium

Maria Abreu[12], Carlos A. Aguilar Salinas[13], Tariq Ahmad[14], Christine M. Albert[18,19], Jessica Alföldi[1,2], Diego Ardissino[20], Irina M. Armean[1,2], Elizabeth G. Atkinson[21,22], Gil Atzmon[23,24], Eric Banks[6], John Barnard[25], Samantha M. Baxter[1], Laurent Beaugerie[26], Emelia J. Benjamin[27,28,29], David Benjamin[6], Louis Bergelson[6], Michael Boehnke[30], Lori L. Bonnycastle[31], Erwin P. Bottinger[32], Donald W. Bowden[33,34,35], Matthew J. Bown[36,37], Harrison Brand[3,38], Steven Brant[39,40,41], Ted Brookings[6,42], Sam Bryant[2,22], Sarah E. Calvo[1,3], Hannia Campos[43,44], John C. Chambers[45,46,47], Juliana C. Chan[48], Katherine R. Chao[1,2], Sinéad Chapman[1,2,7], Daniel I. Chasman[18,49], Siwei Chen[1,2,230], Rex Chisholm[50], Judy Cho[32], Rajiv Chowdhury[51], Mina K. Chung[52], Wendy K. Chung[53,54,55], Kristian Cibulskis[6], Bruce Cohen[56,57], Ryan L. Collins[1,3,4], Kristen M. Connolly[58], Adolfo Correa[59], Miguel Covarrubias[6], Beryl B. Cummings[1,4], Dana Dabelea[60], Mark J. Daly[1,2,15], John Danesh[51], Dawood Darbar[61], Phil Darnowsky[1], Joshua Denny[62], Stacey Donnelly[10], Ravindranath Duggirala[63], Josée Dupuis[64,65], Patrick T. Ellinor[1,66], Roberto Elosua[67,68,69], James Emery[6], Eleina England[1,70], Jeanette Erdmann[71,72,73], Tõnu Esko[1,74], Emily Evangelista[1], Yossi Farjoun[9], Diane Fatkin[75,76,77], Steven Ferriera[11], Jose Florez[49,78,79], Laurent C. Francioli[1,2,230], Andre Franke[80,81], Jack Fu[1,3,38], Martti Färkkilä[82,83,84], Stacey Gabriel[11], Kiran Garimella[6], Laura D. Gauthier[6], Jeff Gentry[6], Gad Getz[49,85,86], David C. Glahn[87,88], Benjamin Glaser[89], Stephen J. Glatt[90], David Goldstein[91,92], Clicerio Gonzalez[93], Julia K. Goodrich[1], Riley Grant[1], Leif Groop[94,95], Sanna Gudmundsson[1,2,8], Namrata Gupta[1,11], Andrea Haessly[6], Christopher Haiman[96], Ira Hall[97], Craig L. Hanis[98], Matthew Harms[99,100], Mikko Hiltunen[101], Matti M. Holi[102], Christina M. Hultman[103,104], Chaim Jalas[105], Thibault Jeandet[6], Mikko Kallela[106], Masahiro Kanai[1,2], Diane Kaplan[6], Jaakko Kaprio[95], Konrad J. Karczewski[1,2,7], Sekar Kathiresan[3,49,107], Eimear E. Kenny[108], Bong-Jo Kim[109], Young Jin Kim[109], Daniel King[1], George Kirov[110], Zan Koenig[2,7], Jaspal Kooner[46,111,112], Seppo Koskinen[113], Harlan M. Krumholz[114,115], Subra Kugathasan[116], Soo Heon Kwak[117], Markku Laakso[118,119], Nicole Lake[120], Trevyn Langsford[6], Kristen M. Laricchia[1,2], Terho Lehtimäki[121,122,123], Monkol Lek[120], Emily Lipscomb[1], Christopher Llanwarne[6], Ruth J. F. Loos[32,124,125], Wenhan Lu[1], Steven A. Lubitz[1,66], Teresa Tusie Luna[126,127], Ronald C. W. Ma[48,128,129], Daniel G. MacArthur[1,16,17,230], Gregory M. Marcus[130], Jaume Marrugat[131,132], Alicia R. Martin[1,2,7], Kari M. Mattila[121,122,123], Steven McCarroll[7,133], Mark I. McCarthy[134,135,136], Jacob L. McCauley[137,138], Dermot McGovern[139], Ruth McPherson[140], James B. Meigs[1,49,141], Olle Melander[142], Andres Metspalu[74], Deborah Meyers[143], Eric V. Minikel[1], Braxton D. Mitchell[144], Vamsi K. Mootha[1,145], Ruchi Munshi[6], Aliya Naheed[146], Saman Nazarian[147,148], Benjamin M. Neale[1,2,230], Peter M. Nilsson[149], Sam Novod[6], Anne O'Donnell-Luria[1,3,8], Michael C. O'Donovan[150], Yukinori Okada[5,151,152], Dost Ongur[49,56], Lorena Orozco[153,154], Michael J. Owen[150], Colin Palmer[155], Nicholette D. Palmer[33], Aarno Palotie[2,7,95], Kyong Soo Park[117,156], Carlos Pato[157], Nikelle Petrillo[6], William Phu[1,8], Timothy Poterba[1,2,7], Ann E. Pulver[158], Dan Rader[147,159], Nazneen Rahman[160], Heidi L. Rehm[1,3], Alex Reiner[161,162], Anne M. Remes[163,164], Dan Rhodes[1], Stephen Rich[165,166], John D. Rioux[167,168], Samuli Ripatti[10,95,169], David Roazen[6], Dan M. Roden[170,171], Jerome I. Rotter[172], Valentin Ruano-Rubio[6], Nareh Sahakian[6], Danish Saleheen[173,174,175], Veikko Salomaa[176], Andrea Saltzman[1], Nilesh J. Samani[37,177], Kaitlin E. Samocha[1,3], Alba Sanchis-Juan[3], Jeremiah Scharf[1,3,7], Molly Schleicher[1], Heribert Schunkert[178,179], Sebastian Schönherr[180], Eleanor G. Seaby[1,181], Cotton Seed[2,7], Svati H. Shah[182,183], Megan Shand[6], Ted Sharpe[6], Moore B. Shoemaker[184], Tai Shyong[185,186], Edwin K. Silverman[187,188], Moriel Singer-Berk[1], Pamela Sklar[189,190,191], Jonathan T. Smith[6], J. Gustav Smith[192,193,194], Hilkka Soininen[195], Harry Sokol[196,197,198,199], Matthew Solomonson[1,2], Rachel G. Son[1], Jose Soto[6], Tim Spector[200], Christine Stevens[1,2,7], Nathan O. Stitziel[201,202,203], Patrick F. Sullivan[103,204], Jaana Suvisaari[176], E. Shyong Tai[205,206,207], Michael E. Talkowski[1,3,7], Yekaterina Tarasova[1], Kent D. Taylor[172], Yik Ying Teo[205,208,209], Grace Tiao[1,2], Kathleen Tibbetts[6], Charlotte Tolonen[6], Ming Tsuang[210,211], Tiinamaija Tuomi[95,212,213], Dan Turner[214], Teresa Tusie-Luna[215,216], Erkki Vartiainen[169], Marquis Vawter[217], Christopher Vittal[1,2], Gordon Wade[6], Lily Wang[1,218], Qingbo Wang[1,5], Arcturus Wang[1,2,7], James S. Ware[1,219,220,221], Hugh Watkins[222], Nicholas A. Watts[1,2], Rinse K. Weersma[223], Ben Weisburd[6], Maija Wessman[95,224], Nicola Whiffin[1,225,226], Michael W. Wilson[1,2], James G. Wilson[227], Ramnik J. Xavier[228,229] & Mary T. Yohannes[1]

[12]University of Miami Miller School of Medicine, Gastroenterology, Miami, FL, 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. [18]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [19]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [20]Department of Cardiology, University Hospital, Parma, Italy. [21]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [22]Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [23]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [24]Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [25]Department of Quantitative Health Sciences, Lerner Research Institute Cleveland Clinic, Cleveland, OH, USA. [26]Gastroenterology Department, Saint Antoine Hospital, Sorbonne Université, APHP, Paris, France. [27]Framingham Heart Study, NHLBI and Boston University, Framingham, MA, USA. [28]Department of Medicine, Boston University Chobanian and Avedisian School of Medicine, Boston, MA, USA. [29]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [30]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [31]National Human Genome Research Institute, National Institutes of Health Bethesda, Bethesda, MD, USA. [32]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [33]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [34]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [35]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [36]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [37]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [38]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [39]Department of Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA. [40]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. [41]Meyerhoff Inflammatory Bowel Disease Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [42]Fulcrum Genomics, Boulder, CO, USA. [43]Harvard School of Public Health, Boston, MA, USA. [44]Central American Population Center, San Pedro, Costa Rica. [45]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [46]Department of Cardiology, Ealing Hospital, NHS Trust, Southall, UK. [47]Imperial College, Healthcare NHS Trust Imperial College London, London, UK. [48]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [49]Department of Medicine, Harvard Medical School, Boston, MA, USA. [50]Northwestern University, Evanston, IL, USA. [51]University of Cambridge, Cambridge, UK. [52]Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH, USA. [53]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, USA. [54]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA. [55]Department of Medicine, Columbia University Medical Center, New York, NY, USA. [56]McLean Hospital, Belmont, MA, USA. [57]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [58]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [59]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [60]Department of Epidemiology Colorado School of Public Health Aurora, Aurora, CO, USA. [61]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [62]Vanderbilt University Medical Center, Nashville, TN, USA. [63]Department of Life Sciences, College of Arts and Scienecs, Texas A&M University—San Antonio, San Antonio, TX, USA. [64]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [65]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada. [66]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [67]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. [68]Centro de Investigación Biomédica en Red Enfermedades Cardiovasculares (CIBERCV), Madrid, Spain. [69]Departament of Medicine, Faculty of Medicine, University of Vic—Central University of Catalonia, Vic, Spain. [70]Clalit Genomics Center, Ramat-Gan, Israel. [71]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [72]German Research Centre for Cardiovascular Research Hamburg/Lübeck/Kiel, Lübeck, Germany. [73]University Heart Center Lübeck, Lübeck, Germany. [74]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [75]Victor Chang Cardiac Research Institute, Darlinghurst, New South Wales, Australia. [76]Faculty of Medicine, UNSW Sydney, Kensington, New South Wales, Australia. [77]Cardiology Department, St Vincent's Hospital, Darlinghurst, New South Wales, Australia. [78]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [79]Programs in Metabolism and Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [80]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. [81]University Hospital Schleswig-Holstein, Kiel, Germany. [82]Clinic of Gastroenterology, Helsinki University and Helsinki University Hospital, Helsinki, Finland. [83]Helsinki University and Helsinki University Hospital, Helsinki, Finland. [84]Abdominal Center, Helsinki, Finland. [85]Bioinformatics Program, MGH Cancer Center and Department of Pathology, Massachusets General Hospital, 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 Hospital and 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, 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 and Development Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [93]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [94]Lund University Sweden, Lund, 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, MI, USA. [98]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [99]Department of Neurology, Columbia University, New York, NY, USA. [100]Institute of Genomic Medicine, Columbia University, New York, 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, Cheongju-si, Republic of Korea. [110]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK. [111]Imperial College, Healthcare NHS Trust, London, UK. [112]National Heart and Lung Institute Cardiovascular Sciences, 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, CT, USA. [115]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA. [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, Tampere University, Tampere, Finland. [122]Fimlab Laboratories, Tampere, Finland. [123]Finnish Cardiovascular Research Center, Tampere Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [124]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [125]The Novo Nordisk

Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [126]National Autonomous University of Mexico, Mexico City, Mexico. [127]Salvador Zubirán National Institute of Health Sciences and Nutrition, Mexico City, Mexico. [128]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [129]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [130]Division of Cardiology, University of California San Francisco, San Francisco, CA, USA. [131]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain. [132]CIBERCV, Madrid, Spain. [133]Department of Genetics, Harvard Medical School, Boston, MA, USA. [134]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [135]Welcome Centre for Human Genetics, University of Oxford, Oxford, UK. [136]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals, NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [137]John P. Hussman Institute for Human Genomics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA. [138]The Dr. John T. Macdonald Foundation Department of Human Genetics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA. [139]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [140]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [141]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [142]Department of Clinical Sciences University, Hospital Malmo Clinical Research Center, Lund University, Malmö, Sweden. [143]University of Arizona Health Science, Tuscon, AZ, USA. [144]University of Maryland School of Medicine, Baltimore, MD, USA. [145]Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [146]International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh. [147]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [148]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [149]Department of Clinical Sciences, Skåne University Hospital, Lund University, Malmö, Sweden. [150]Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff, UK. [151]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan. [152]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan. [153]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [154]Laboratory of Immunogenomics and Metabolic Diseases, INMEGEN, Mexico City, Mexico. [155]Medical Research Institute, Ninewells Hospital and Medical School University of Dundee, Dundee, UK. [156]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea. [157]Department of Psychiatry, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [158]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [159]Children's Hospital of Philadelphia, Philadelphia, PA, USA. [160]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [161]University of Washington, Seattle, WA, USA. [162]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [163]Medical Research Center, Oulu University Hospital, Oulu, Finland. [164]Research Unit of Clinical Neuroscience, Neurology University of Oulu, Oulu, Finland. [165]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [166]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA. [167]Research Center Montreal Heart Institute, Montreal, Quebec, Canada. [168]Department of Medicine, Faculty of Medicine, Université de Montréal, Quebec, Canada. [169]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [170]Departments of Medicine, Pharmacology and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [171]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [172]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. [173]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [174]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [175]Center for Non-Communicable Diseases, Karachi, Pakistan. [176]National Institute for Health and Welfare, Helsinki, Finland. [177]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK. [178]Department of Cardiology, Deutsches Herzzentrum München, Technical University of Munich, DZHK Munich Heart Alliance, Munich, Germany. [179]Technische Universität München, Munich, Germany. [180]Institute of Genetic Epidemiology, Department of Genetics, Medical University of Innsbruck, Innsbruck, Austria. [181]Faculty of Medicine, University of Southampton, Southampton, UK. [182]Duke Molecular Physiology Institute, Durham, NC, USA. [183]Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, NC, USA. [184]Division of Cardiovascular Medicine, Nashville VA Medical Center, Vanderbilt University School of Medicine, Nashville, TN, USA. [185]Division of Endocrinology, National University Hospital, Singapore, Singapore. [186]NUS Saw Swee Hock School of Public Health, Singapore, Singapore. [187]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [188]Harvard Medical School, Boston, MA, USA. [189]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [190]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [191]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [192]The Wallenberg Laboratory, Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University, Gothenburg, Sweden. [193]Department of Cardiology, Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden. [194]Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden. [195]Institute of Clinical Medicine Neurology, University of Eastern Finad, Kuopio, Finland. [196]Gastroenterology Department, Centre de Recherche Saint-Antoine, CRSA, AP-HP, Saint Antoine Hospital, Sorbonne Université, INSERM, Paris, France. [197]INRA, UMR1319 Micalis, Jouy en Josas, France. [198]Paris Center for Microbiome Medicine (PaCeMM) FHU, Paris, France. [199]AgroParisTech, Jouy en Josas, France. [200]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [201]Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. [202]Department of Genetics, Washington University School of Medicine, St Louis, MO, USA. [203]The McDonnell Genome Institute at Washington University, St Louis, MO, USA. [204]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [205]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [206]Department of Medicine,Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [207]Duke–NUS Graduate Medical School, Singapore, Singapore. [208]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [209]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [210]Center for Behavioral Genomics, Department of Psychiatry, University of California San Diego, San Diego, CA, USA. [211]Institute of Genomic Medicine, University of California San Diego, San Diego, CA, USA. [212]Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [213]Institute of Genetics, Folkhalsan Research Center, Helsinki, Finland. [214]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [215]Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico. [216]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [217]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA. [218]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA, USA. [219]National Heart and Lung Institute, Imperial College London, London, UK. [220]Royal Brompton and Harefield Hospitals, Guy's and St. Thomas' NHS Foundation Trust, London, UK. [221]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [222]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [223]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, Netherlands. [224]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [225]Big Data Institute, University of Oxford, Oxford, UK. [226]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [227]Division of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [228]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [229]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA.

Methods

Aggregation, variant calling and quality control of gnomAD genome data

We aggregated whole-genome sequence data from 153,030 individuals spanning projects from case–control consortia and population cohorts, in a similar manner to previous efforts[5]. Informed consent was obtained for the original studies that generated sequencing data and we keep a blank copy of those consents on file with our local Office of Research Subject Protection (ORSP). The Institutional Review Board (IRB) has approved our study protocol, and we confirm that we have complied with all relevant ethical regulations relating to human research subjects.

We harmonized the sequencing data using the GATK Best Practices pipeline and joint-called all samples using Hail (https://github.com/hail-is/hail/commit/84fa81b9ea3d), and developed and utilized an updated pipeline of sample, variant, and genotype quality control to create a high-quality callset of 76,156 individuals, computing frequency information for several strata of this dataset based on attributes such as ancestry and sex for each of 644,267,978 short nuclear variants (see Supplementary Information).

Estimation of trinucleotide context-specific mutation rates

We estimated the probability of a given nucleotide mutating to one of the three other possible bases in a trinucleotide context ($XY_1Z>XY_2Z$), by computing the proportion of all possible variants observed per context in the human genome. Since CpG transitions begin to saturate (proportion observed approaching 1) at a sample size of ~10K genomes, we downsampled the gnomAD dataset to 1,000 genomes for this calculation. The computed proportion observed values, which represent the relative mutability of each trinucleotide context, were further scaled so that the weighted genome-wide average is the human per-base, per-generation mutation rate (1.2×10^{-8}) to obtain the absolute mutation rates μ. To estimate the proportion of variants expected to be observed in the full gnomAD dataset of 76,156 genomes, we fitted the actual proportion observed in the dataset against μ, using an exponential regression that caps at 1 for refining the estimates of (near-) saturated variant types ($R^2 = 0.999$, Extended Data Fig. 1a,b and Supplementary Data 1).

A total of 390,393,900 high-quality, rare (AF ≤ 0.1%) variants observed in 76,156 gnomAD genomes, a dataset of 6,079,733,538 possible variants at 2,026,577,846 autosomal sites (30–32× coverage), were used in the calculation of trinucleotide context-specific mutation rates. The estimates are well-correlated with the mutation rates reported in previous independent studies and are highly stable across different AF thresholds in gnomAD (Supplementary Fig. 6).

Adjustment of the effect of DNA methylation on CpG mutation rates

Given the strong effect of DNA methylation on increasing the mutation rate at CpG sites, we stratified all CpG sites by their methylation levels and computed the proportion observed within each context and methylation level. As an improvement to our previous methylation annotation (by averaging different tissues[5]), we analysed methylation data from germ cells across 14 developmental stages, comprising 8 from preimplantation embryos (sperm, oocyte, pronucleus, two-cell, four-cell, eight-cell, morula and blastocyst stage embryos)[65] and six from primordial germ cells (7 week, 10 week, 11 week, 13 week, 17 week and 19 week)[66]. For each stage, we computed methylation level at each CpG site as the proportion of whole-genome bisulfite sequencing reads corresponding to the methylated allele. To derive a composite score from the 14 stages, we regressed the observation of a CpG variant in gnomAD (0 or 1) on the methylation computed at the corresponding site (a vector of 14), and we used the coefficients from the regression model as weights to compute a composite methylation score for each CpG site. This metric was further discretized into 16 levels (by a minimum step of 0.05: [0,0.05], (0.05,0.1], (0.1,0.15], (0.15,0.2], (0.2,0.25], (0.25,0.3], (0.3,0.5], (0.5,0.55], (0.55,0.6], (0.6,0.65], (0.65,0.7], (0.7,0.75], (0.75,0.8], (0.8,0.85], (0.85,0.9], (0.9,1.0]) to stratify CpG variants in the mutation rate analysis.

Adjustment of the effects of regional genomic features on mutation rates

To estimate the effects of regional genomic features on mutation rates under neutrality, we utilized DNMs as a proxy of spontaneous mutations, and fitted logistic regression models using the genomic features as predictive variables. A set of 413,304 unique DNMs were compiled from two large-scale family-based whole-genome sequencing studies[23,24] and an exclusive set of 4,104,879 genomic sites (~10× the DNMs) randomly drew from the genome was used as the 'non-mutated' background. For each DNM or background site, we computed 13 genomic features (see Collection of genomic features) at four scales by taking the mean value of 1-kb, 10-kb, 100-kb and 1-Mb windows centring at the site. This generated a feature matrix of 13×4 = 52 columns and 413,304 + 4,104,879 = 4,518,183 rows. The matrix was further divided based on the trinucleotide context of each DNM or background site (by row) to assess the effects of genomic features on context-specific mutation rates. In particular, for CpG contexts, features that were correlated with DNA methylation (GC content, CpG island, short interspersed nuclear element and nucleosome density), which had been used for adjusting CpG mutation rates, were excluded from the analysis.

For each trinucleotide context, we first performed univariable logistic regression to select features that are significantly associated with an increased or decreased probability of observing a DNM. Features with a significant association surpassing the Bonferroni correction for 13×4 = 52 tests were selected; if a feature was significant at multiple genomic scales, the smallest window size was selected for the highest resolution (Extended Data Fig. 1c). Next, we fitted multivariable logistic regression using the selected features to predict DNMs from the background. To control for multicollinearity, we transformed the input feature matrix using principal component analysis[67] (PCA) to generate decorrelated predictive variables (that is, the principal components). The regression coefficients were the primary output of interest, which represent the effects of genomic features on increasing (a positive coefficient) or decreasing (a negative coefficient) the mutation rate, and were used for adjusting the expected number of variants in a given region. The selected features, the principal components and the coefficients are summarized in Extended Data Fig. 1c and are available as pickle files for implementation (see Code availability).

Prediction of expected number of variants per 1 kb

Using the trinucleotide mutation rate estimates and the above adjustments, we computed the expected number of variants in a given 1-kb genomic window as follows:

$$\exp(w) = \sum_{i}^{32} r(w)_i \sum_{j=1}^{3} \sum_{m=1}^{k} n(w)_{i,j,m} \times p_{i,j,m}$$

where i denotes one of the 32 trinucleotide contexts; j denotes one of the three bases substituting the central nucleotide; and m denotes one of the k DNA methylation levels, where $k = 16$ for CpG sites (see 'Adjustment of the effect of DNA methylation on CpG mutation rates') and $k = 1$ for non-CpG sites (that is, no stratification). Essentially, the expected value of variants in a genomic window w is calculated by multiplying the number of possible variants (n) in w by the probability of a variant (p) and summing across all trinucleotide contexts (i), substitutions (j) and methylation levels (m); $p_{i,j,m}$ is the trinucleotide mutation rate estimated in this study (as described in 'Estimation of trinucleotide context-specific mutation rates').

Additionally, Exp is adjusted by a factor r, which represents the effect of regional genomic features of w on mutation rate. For each i, specific

features have been pre-selected and their effects on mutation rate have been estimated using logistic regression models (see 'Adjustment of the effects of regional genomic features on mutation rates'). Denote the feature values (computed centring w and decorrelated by PCA) and the regression coefficients by $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t\}$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, .., \boldsymbol{\beta}_t\}$, respectively, where t is the number of selected features for i, the adjustment factor r is defined as the ratio of logit given $\mathbf{x}(w)$ to that of the genome-wide average $\bar{\mathbf{x}}$: $r = \boldsymbol{\beta} \cdot \mathbf{x}(w) / \boldsymbol{\beta} \cdot \bar{\mathbf{x}}$; since the adjustment is specific to each trinucleotide context, r is further subscribed by i.

Construction of Gnocchi

We created a signed score, Gnocchi, to quantify the depletion of variation (constraint) at a 1-kb scale by comparing the observed variation to an expectation:

$$\chi^2 = (\text{Obs} - \text{Exp})^2 / \text{Exp}$$

$$\text{Gnocchi} = \begin{cases} \sqrt{\chi^2} \ if \ \text{Obs} < \text{Exp} \\ -\sqrt{\chi^2} \ if \ \text{Obs} \geq \text{Exp} \end{cases}$$

The observed variant count (Obs) is the number of unique rare (AF ≤ 0.1%) variants in a 1-kb window identified in the gnomAD dataset of 76,156 genomes, and the expected number of variants (Exp) is established as described above based on the sequence context and the regional genomic features of the 1-kb window.

Gnocchi scores were created for 2,689,987 non-overlapping 1-kb windows across the human genome, comprising 2,561,056 on autosomes and 128,931 on chromosome X. Due to the lack of DNM data on chromosome X, the genomic feature adjustment factor r was assessed using autosomal regions and extrapolated to chromosome X. We performed downstream analyses separately for autosomes and chromosome X and presented the former as primary, with the latter provided in Supplementary Fig. 8. For the analyses, we filtered the dataset to windows where (1) the sites contained at least 1,000 possible variants; (2) at least 80% of the observed variants passed all variant call filters (INFO/FILTER equals to "PASS"); and (3) the mean coverage in the gnomAD genomes was between 25–35× (or 20–25× for chromosome X). This resulted in 1,984,900 autosomal windows (77.5% of initial) for the primary analyses, of which 141,341 overlapped with coding regions and 1,843,559 were exclusively non-coding. The computed Gnocchi scores are available in Supplementary Data 2. We also computed the sores in a sliding window approach (1 kb stepped by 100 bp) and provided them in Supplementary Data 3.

Collection of genomic features

The 13 regional genomic features used for adjusting trinucleotide mutation rate are (1) GC content[68], (2) low-complexity region[69], (3) short and (4) long interspersed nuclear element[68], distance from the (5) telomere and the (6) centromere[68], (7) male and (8) female recombination rate[23], (9) DNA methylation, (10) CpG island[68], (11) nucleosome density[70], (12) maternal and (13) paternal DNM cluster[71]. Data were downloaded from the referenced resources, lifted over to GRCh38 coordinates when needed using CrossMap[72], and files in .bed or .BigWig format were processed using bedtools[73] and bigWigAverageOverBed[74] to obtain feature values within specific genomic windows.

Correlation between Gnocchi and APS

As an internal validation, we compared our Gnocchi score against the structural variation constraint score APS[25]. For each structural variation from the original study[25], we assessed its constraint by assigning the highest Gnocchi score among all overlapping 1-kb windows. The correlation between Gnocchi and APS was evaluated across 116,184 high-quality autosomal structural variations scored by both metrics, using a linear regression test. In Fig. 1b, the correlation was presented by the mean value of APS across ascending constraint Gnocchi score bins, with 95% confidence intervals computed from 100-fold bootstrapping.

Correlation between Gnocchi and putative functional non-coding annotations

We validated the Gnocchi metric using a number of external functional annotations, including 926,535 ENCODE cCREs[26] (34,803 PLS, 141,830 pELS, 667,599 dELS and 56,766 CTCF-only elements; 25,537 elements with undetermined regulatory categories were not included in analysis), 63,285 FANTOM5[27] enhancers, 331,601 super enhancers (SEdb[28]), 111,308 GWAS Catalog[31] variants (with an association $P \leq 5.0 \times 10^{-8}$; 9,229 with an independent replication), 2,191 GWAS variants fine-mapped across population biobanks with a posterior inclusion probability of causality≥0.9[75], and 100,530 CNVs from a CNV morbidity map of DD[51,52].

To assess the correlation between Gnocchi and the collected functional elements, we intersected each annotation with the scored 1-kb windows binned by Gnocchi score (<−4, [−4,−3), [−3,−2), [−2,−1), [−1,−0), [0,1), [1,2), [2,3), [3,4), ≥4), and counted the frequency of overlapping windows within each bin. The enrichment of a given annotation (except CNVs) at a constraint level was evaluated by comparing the corresponding frequency to the genome-wide average using a Fisher's exact test. In the analysis of CNVs, we assessed their enrichment in constrained regions by assigning each CNV the highest Gnocchi score among its overlapping windows and comparing the proportions of constrained CNVs (Gnocchi ≥ 4) from cases of DD and healthy controls (Supplementary Data 4). The enrichment was further examined using a logistic regression model to adjust for the size and gene content (gene constraint[5] and gene number) of each CNV. We note that we performed all above analyses restricting to exclusively non-coding windows to evaluate the use of Gnocchi in characterizing the non-coding genome.

Estimation of constraint for aggregated regulatory annotations

We estimated how constrained the sequences encoding regulatory elements overall compared to coding exons by aggregating the regulatory annotations at a 1kb scale. These included 7,246 promoter, 154,003 enhancer, 117 microRNA (miRNA) and 414,084 long non-coding RNA (lncRNA) 1-kb elements, created from concatenating ENCODE cCREs-PLS, cCREs-dELS, GENCODE[76] miRNA and FANTOM5 lncRNA[77] annotations, respectively, into 1-kb windows. Similarly, 27,875 exonic 1-kb elements were created from aggregating all protein-coding exons. Gnocchi scores were computed for the created 1-kb elements and the percentiles of each regulatory annotation were compared against the exonic region. Benchmarking on the 50th percentile (median) of exonic regions, we estimated the proportion of the regulatory elements that are under selection as strong as the coding exons.

Incorporation of Gnocchi into GWAS fine-mapping

To demonstrate the use of Gnocchi in statistical fine-mapping, we performed approximate functionally informed fine-mapping[39] incorporating Gnocchi score and our previous fine-mapping results for 119 UK Biobank traits[75]. The Gnocchi scores were normalized and used as functional prior probabilities to update the posterior inclusion probabilities (PIPs; denoted as $\text{PIP}_{\text{Gnocchi}}$) based on the previous UK Biobank fine-mapping (using a uniform prior, PIP_{unif}) and SuSiE[78]. To exclude signals that potentially correspond to coding variants, we restricted our analysis to 60,121 non-coding variants in 6,592 SuSiE 95% credible set–trait pairs that do not contain variants within 1 kb of exonic regions. A total of 13,069 variant–trait pairs were predicted to have an increased PIP ($\Delta\text{PIP} \geq 0.01$) of causality. The variants, associated traits and PIP scores (PIP_{unif} and $\text{PIP}_{\text{Gnocchi}}$) are provided in Supplementary Data 5.

Comparison of Gnocchi and other predictive metrics

We compared the Gnocchi metric with other seven genome-wide predictive scores: Orion[15], CDTS[16], gwRVIS[19], DR[17], phyloP[21], phastCons[20] and GERP[42]. Each score was downloaded from the original study, lifted

over to GRCh38 coordinates (for Orion) and multiplied by −1 (for CDTS, gwRVIS and DR) when needed so that a higher value represents a higher constraint/conservation for all metrics. Pairwise correlation between the scores was assessed by comparing the mean value of each score on 1kb windows, using a Spearman's rank correlation test.

We evaluated the predictive performance of each metric in distinguishing functional non-coding variants (positive variant set) from background variants (negative variant set). Four positive variant sets were compiled from public databases: (1) 9,229 variants from GWAS Catalog[31] (with an independent replication), (2) 2,191 variants from a recent fine-mapping study[75] (with a posterior inclusion probability of causality ≥ 0.9), (3) 140 high-confidence variants from (2), and (4) 1,026 variants from ClinVar[40] (annotated as 'pathogenic' or 'likely pathogenic') and HGMD (annotated as 'disease-causing mutation' curated in ref. 16). All variants were filtered to non-coding regions; in particular, pathogenic variants were more strictly filtered to intergenic/intron variants given its strong predominance of variants close to protein-coding exons (>90% were splice site/region variants). A further stringent non-coding subset was generated by excluding variants within 10 kb of any exons, which resulted in (1) 4,379, (2) 967, (3) 59 and (4) 45 variants. For each positive variant set, a negative variant set was created by randomly drawing variants from the Trans-Omics for Precision Medicine (TOPMed) whole-genome sequencing dataset (Freeze 8)[79] to ~10× the size of corresponding positive variant set, of which the most severe molecular consequence is intergenic or intron and the AF approximates the positive variant set; AF > 5% and allele count = 1 were applied respectively for matching positive variant set (1)–(3) and (4), based on their AF distributions in TOPMed (Fig. 3b). The selected variants were scored by each of the eight metrics, using bedtools[73] (for .bed files) and bigWigAverageOverBed[70] (for .BigWig files), and the performance of each metric in classifying positive and negative variants was assessed by the AUC statistic, as presented by the ROC curve.

To investigate whether different metrics capture complementary information in the classification, we fitted logistic regression models using all eight metrics as independent variables. The relative contribution of each metric was evaluated by the dominance analysis[80,81], which estimates the dominance of one predictor over another by comparing their additional R^2 contributions across all subset models. We further explored whether specific features were particularly captured by (and may have contributed to the performance of) our metric. We merged all positive variant sets and focused on a set of variants (n = 204) that were uniquely prioritized by our metric, defined as being captured in the 99th percentile of Gnocchi score but not in that of any other scores. Specific features associated with these variants were evaluated by comparing values of the 13 genomic features of these variants to the rest of the positive variant set. The fold change was used to indicate the extent to which a feature is distinguished in variants captured by Gnocchi from others.

Correlation of constraint between non-coding regulatory elements and protein-coding genes

To examine whether constraint of non-coding regulatory elements informs the constraint of their target genes, we compared Gnocchi scores of enhancers linked to constrained genes and unconstrained genes. The former included well-established gene sets of 189 ClinGen[57] haploinsufficient genes, 2,454 MGI[58] essential genes mapped to human orthologues, 1,771 OMIM[59] autosomal dominant genes and 1,920 LOEUF[5] first-decile genes; and the latter included a curated list of 356 olfactory receptor genes and 189 LOEUF last-decile genes with at least 10 expected LOF variants (which are sufficiently powered to be classified into the most constrained decile[5]). The LOEUF underpowered list included 1,117 genes with ≤5 expected LOF variants. Enhancers linked to each gene were obtained from the Roadmap Epigenomics Enhancer-Gene Linking database, which used correlated patterns of activity between histone modifications and gene expression to predict enhancer-gene links[82,83]. For each gene, we aggregated and merged enhancers predicted from all 127 reference epigenomes and assigned the most constrained enhancer to each gene for the analysis of enhancer-gene constraint correlation (Supplementary Data 6).

In the analysis of correlation between tissue-specific enhancer constraint and tissue-specific gene expression, we processed the enhancer-gene links with the same principle as described above but within specific tissue types (as defined in the Roadmap Epigenomics metadata[56]). For each gene and tissue type, we searched for tissue-specific gene expression in the Genotype-Tissue Expression (GTEx[60]) database (RNASeQCv1.1.9) and computed a normalized median expression for each gene ($\log_2$(TPM+1)). Enhancer constraint and gene expression values were calculated for 11 matched tissue types, and the correlation within each tissue type was evaluated by regressing gene expression on enhancer constraint, including gene constraint (LOEUF score) as a covariate.

Incorporation of non-coding constraint of regulatory elements into gene constraint modelling

To demonstrate the practical value of non-coding constraint in improving gene constraint modelling, we compared two models using—(1) LOEUF and (2) LOEUF plus enhancer Gnocchi score (as described in 'Correlation of constraint between non-coding regulatory elements and protein-coding genes')—in predicting constrained genes, with a particular focus on genes that were underpowered in LOEUF. A set of 3,220 unique constrained genes were curated from ClinGen[57], MGI[58] and OMIM[59] (see 'Correlation of constraint between non-coding regulatory elements and protein-coding genes'), and a set of 356 olfactory receptor genes was used as the unconstrained genes. We trained logistic regression models on 50% of the genes and tested the performance on 77 underpowered genes in the remaining 50%. The predictive performance of the two models were measured by AUC, and the significance of the difference in AUCs was assessed using a bootstrap test[84].

Power of constraint detection

We estimated the power of our metric in detecting non-coding constraint as the percentage of the non-coding genome to obtain a high Gnocchi score (Gnocchi ≥ 4) under a certain strength of negative selection, which was quantified by the level of depletion of variation (that is, 1 − observed/expected). For a given depletion of variation, the minimum number of expected variants to achieve a Gnocchi ≥ 4 was determined, and the number of samples required to achieve the expected number of variants was estimated using a linear model of log(number of expected variants) ~ log(number of samples) from downsampling the gnomAD dataset. The power was estimated at two scales, 1 kb (used in this study) and 100 bp, and benchmarked by the depletion of variation observed in coding exons of similar size.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The aggregated allele frequency dataset is available in a browser at https://gnomad.broadinstitute.org, with bulk downloads for VCF files and Hail tables, as well as all constraint statistics described in this manuscript. Additionally, we provide a subset of the dataset that includes individual-level data for the HGDP[85] and 1000 Genomes projects[86]—the generation and use of this dataset is described in a companion manuscript[75]. There are no restrictions on the aggregate data released. External datasets used in this study are available in the following public resources: ENCODE cCREs, https://screen-v2.wenglab.org/; super enhancers, http://www.licpathway.net/sedb/download.php;

FANTOM5 enhancers, https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/enhancer/; miRNA, https://genome.ucsc.edu/cgi-bin/hgTables (All GENCODE V32 track); FANTOM5 lncRNA, https://fantom.gsc.riken.jp/cat/v1/#/genes; GWAS Catalog, https://genome.ucsc.edu/cgi-bin/hgTables (GWAS Catalog track); GWAS fine-mapping, https://www.finucanelab.org/data; CNV morbidity map of DD, https://genome.ucsc.edu/cgi-bin/hgTables (Development Delay track); ClinVar, https://genome.ucsc.edu/cgi-bin/hgTables (ClinVar Variants track); TOPMed, https://bravo.sph.umich.edu/freeze8/hg38/downloads; ClinGen, https://genome.ucsc.edu/cgi-bin/hgTables (ClinGen track); MGI, https://www.informatics.jax.org/; OMIM, https://www.omim.org/; Roadmap Epigenomics Enhancer-Gene Linking, https://ernstlab.biolchem.ucla.edu/roadmaplinking/; GTEx https://gtexportal.org/home/datasets.

Code availability

All code to perform quality control of the resource is publicly available at https://github.com/broadinstitute/gnomad_qc, and many of the functions are documented in a Python package (gnomad) at https://broadinstitute.github.io/gnomad_methods/index.html. The code to compute the constraint statistics is available at https://github.com/atgu/gnomad_nc_constraint.

65. Zhu, P. et al. Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat. Genet.* **50**, 12–19 (2018).
66. Tang, W. W. et al. A unique gene regulatory network resets the human germline epigenome for development. *Cell* **161**, 1453–1467 (2015).
67. Ross, D. A., Lim, J., Lin, R.-S. & Yang, M.-H. Incremental learning for robust visual tracking. *Int. J. Comput. Vision* **77**, 125–141 (2008).
68. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
69. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
70. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
71. Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
72. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
75. Koenig, Z. et al. A harmonized public resource of deeply sequenced diverse human genomes. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.23.525248 (2023).
76. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
77. Hon, C. C. et al. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
78. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
79. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
80. Budescu, D. V. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psych. Bull.* **114**, 542 (1993).
81. Azen, R. & Budescu, D. V. The dominance analysis approach for comparing predictors in multiple regression. *Psych. Methods* **8**, 129 (2003).
82. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
83. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
84. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 1–8 (2011).
85. Bergstrom, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
86. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Acknowledgements The authors thank the individuals whose data is in gnomAD for their contributions to research. Development of the Genome Aggregation Database was supported by NIDDK U54DK105566 and the NHGRI of the National Institutes of Health under award number U24HG011450. Additional funding for Genome Aggregation Database Consortium members is listed in the Supplementary Information. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions S.C., L.C.F., J.K.G., Q.W., A.O.-L., H.L.R., M.J.D., B.M.N., D.G.M. and K.J.K. contributed to the writing of the manuscript and generation of figures. S.C., R.L.C., M.K. and K.J.K. contributed to the analysis of data. L.C.F., Q.W., C.V., L.D.G., T.P., C.S., M.E.T., B.M.N. and K.J.K. developed tools and methods. L.C.F., J.K.G., J.A., M.W.W., Y.T., W.P., M.T.Y., Z.K., Y.F., E.B., S.D., S.G., N.G., S.F., C.T., S.N., L.B., D.R., V.R.-R., M.C., C.L., N.P., G.W., T.J., R.M., K.T., A.R.M., G.T. and K.J.K. contributed to the production and quality control of the gnomAD dataset. N.A.W., R.G., M.S. and K.J.K. contributed to the gnomAD browser. 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.

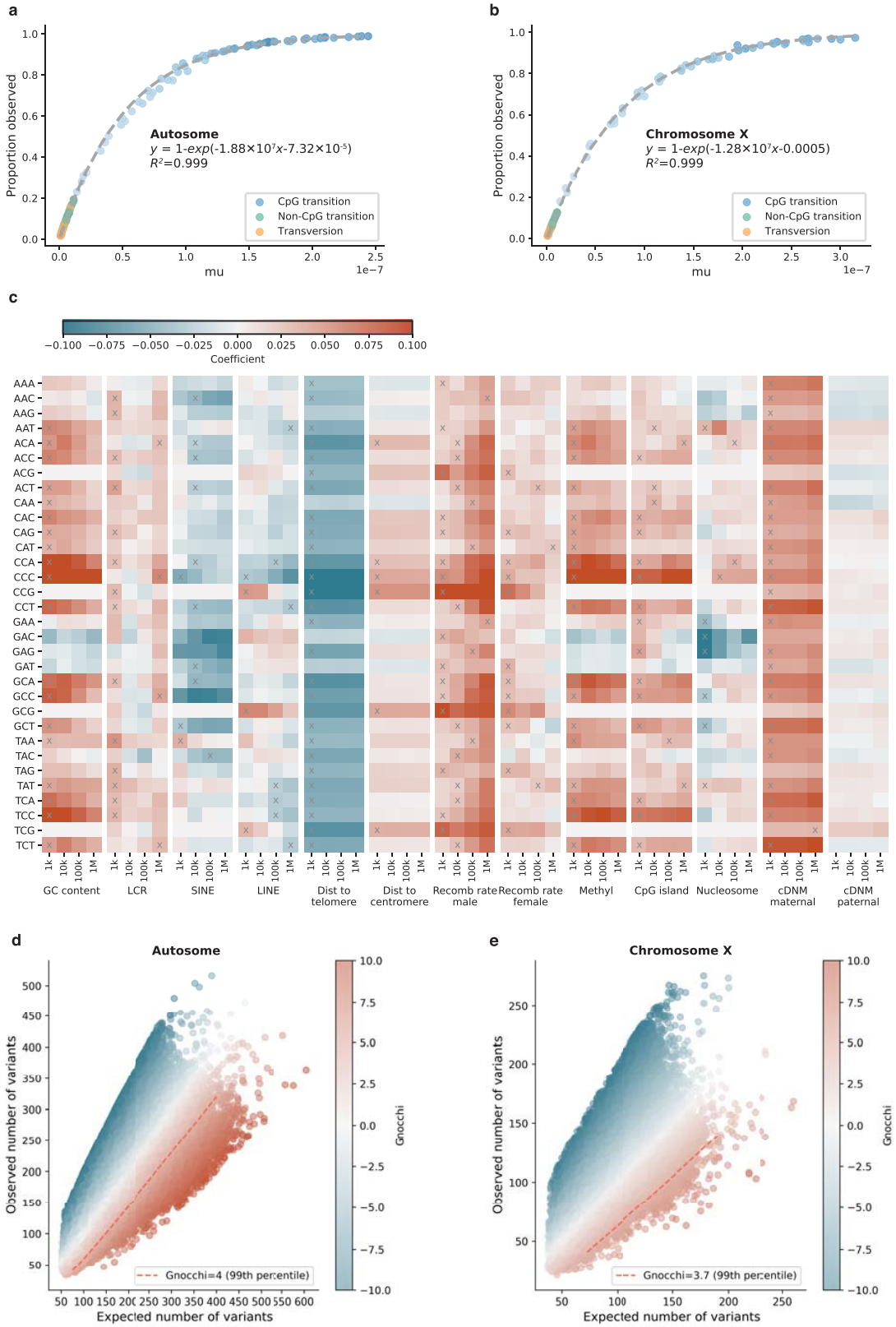
Competing interests K.J.K. is a consultant for Vor Biopharma, Tome Biosciences, and is on the Scientific Advisory Board of Nurture Genomics. D.G.M. is a paid advisor to GSK, Insitro, Variant Bio and Overtone Therapeutics, and has previously received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer and Sanofi-Genzyme.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06045-0.

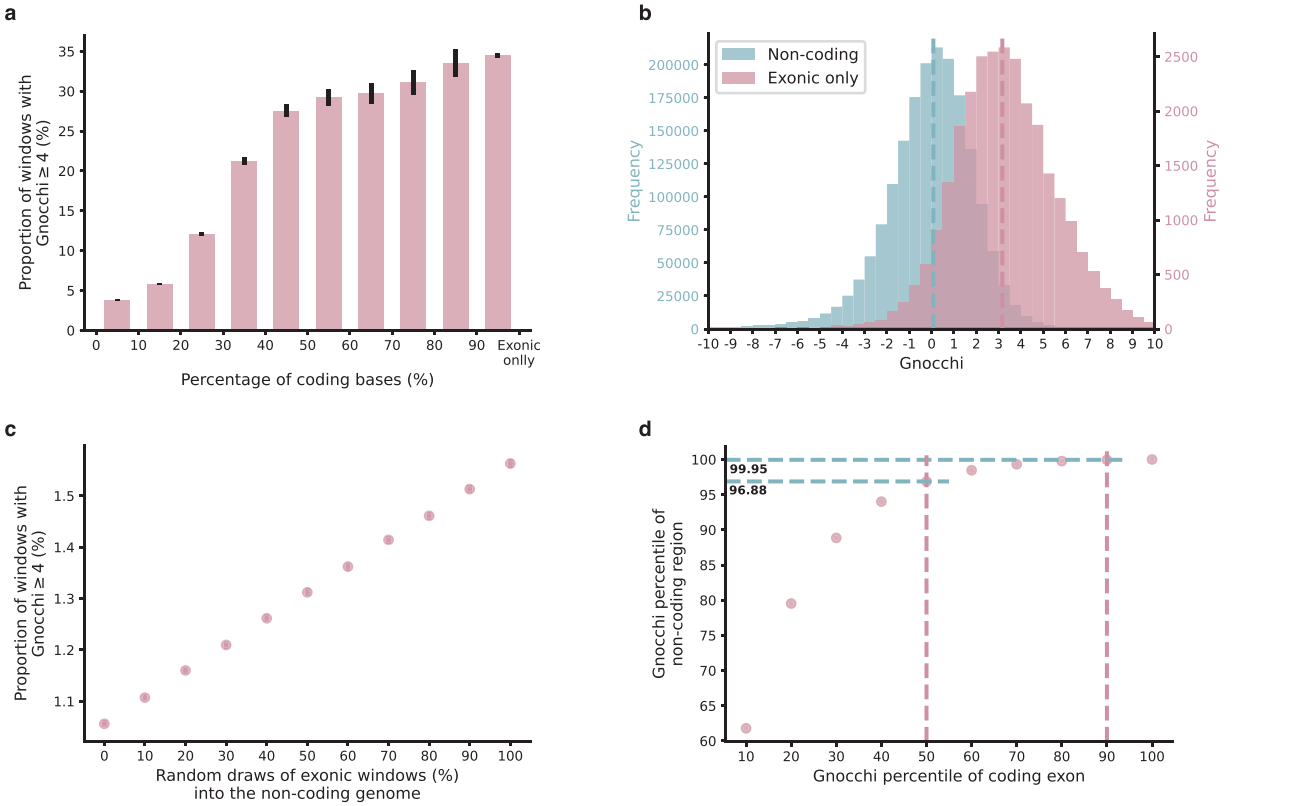
Correspondence and requests for materials should be addressed to Siwei Chen or Konrad J. Karczewski.

Peer review information *Nature* thanks Slavé Petrovski, Ryan Dhindsa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

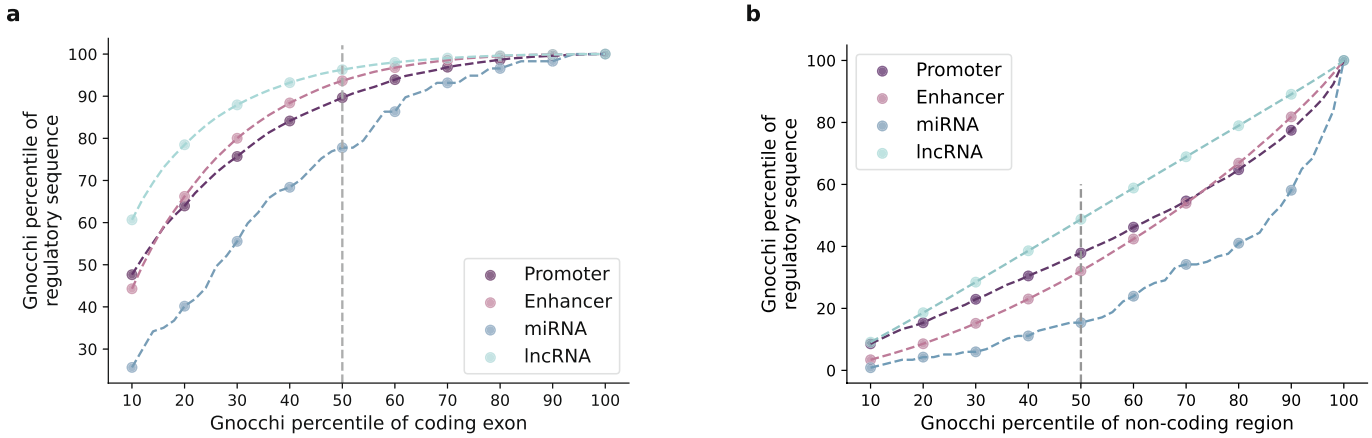


Extended Data Fig. 1 | See next page for caption.

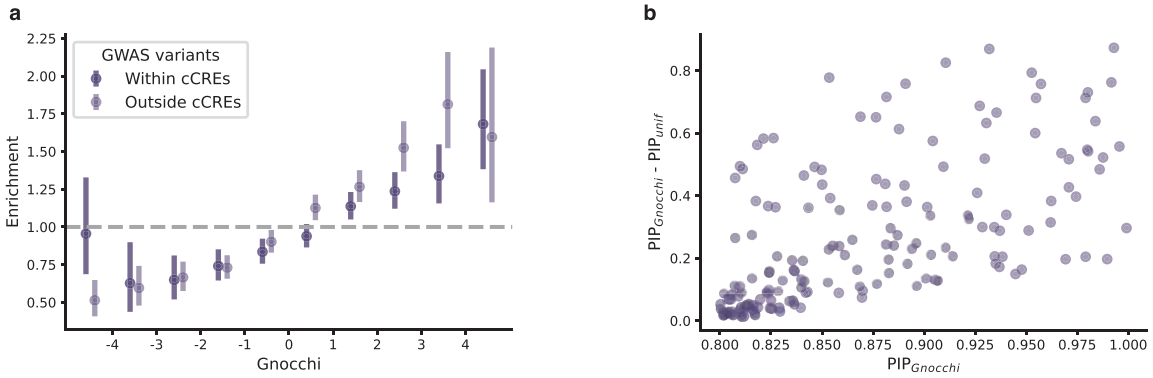
Extended Data Fig. 1 | Construction of mutational model and Gnocchi score. a,**b**, Estimation of trinucleotide context-specific mutation rates. The proportion of possible variants observed for each substitution and context in 76,156 gnomAD genomes (y-axis) is exponentially correlated with the absolute mutation rate estimated from 1,000 downsampled genomes (x-axis). Fit lines were modeled separately for human autosomes (**a**) and chromosome X (**b**). **c**, Estimation of the effects of regional genomic features on mutation rates. The effects of 13 genomic features at four scales (window sizes 1kb-1Mb; x-axis) on the mutation rate of 32 trinucleotide contexts (y-axis) are shown, colored by the coefficient from regressing *de novo* mutations (DNMs) on each specific feature and window size. Red/Blue color indicates a positive/negative effect of increasing the feature value on mutation rates; grey crosses indicate significant features at the smallest possible window size after Bonferroni correction for 13×4 = 52 tests. Abbreviations: LCR=low-complexity region, SINE/LINE=short/long interspersed nuclear element, Dist=Distance, Recomb=Recombination, Methyl=Methylation. **d**,**e**, The distribution of Gnocchi score as a function of expected and observed variation. Each point represents the Gnocchi score of a 1kb window on the genome (N = 1,984,900 on autosomes (**d**) and N = 57,729 on chromosome X (**e**)), which quantifies the deviation of observed variation from expectation. A positive Gnocchi score (red) indicates depletion of variation (observed<expected) and the higher the score the stronger the depletion; the red dashed line indicates the 99th percentile of Gnocchi scores across the autosomes (**d**) or chromosome X (**e**).



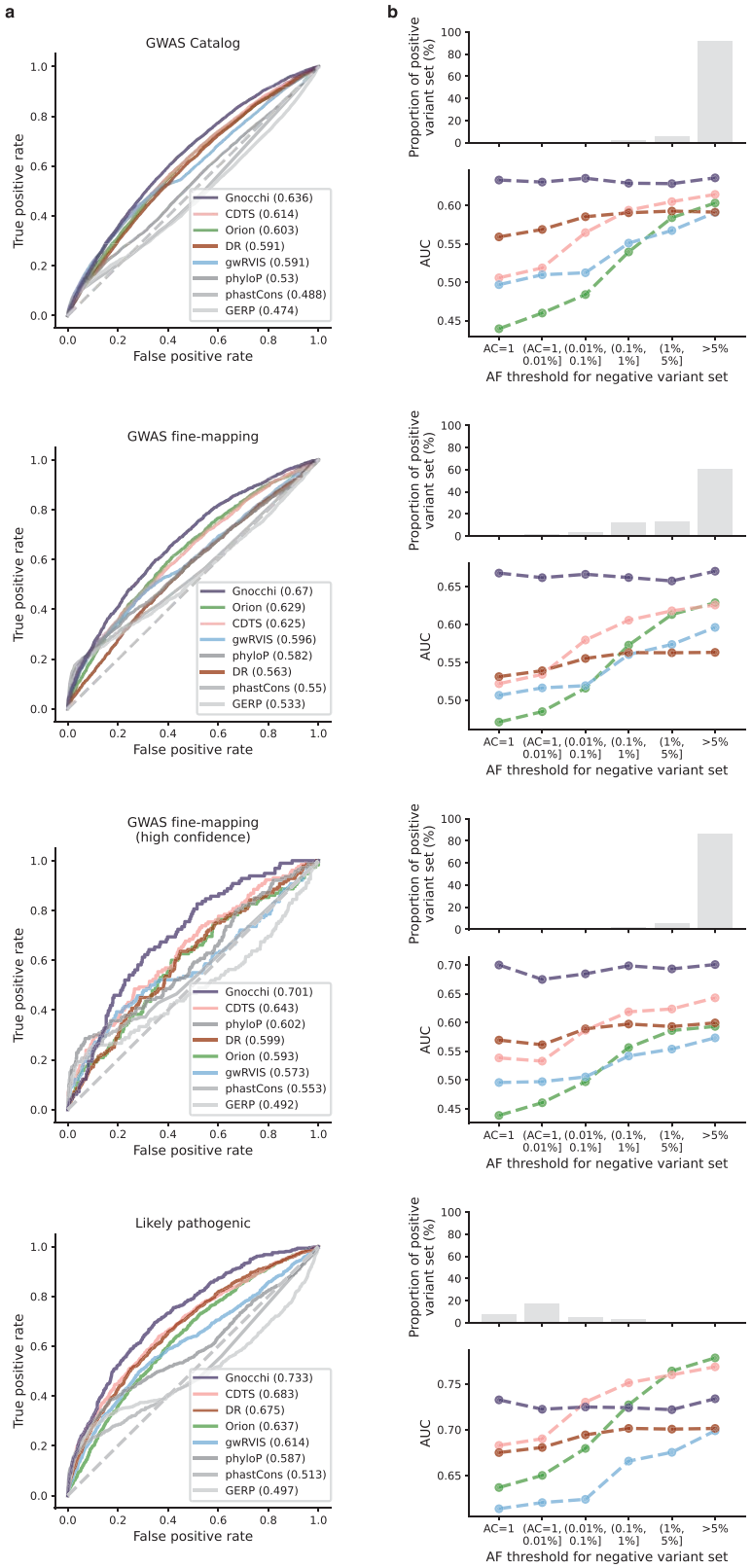
Extended Data Fig. 2 | Comparison of Gnocchi score between coding and non-coding regions. a, The proportion of highly constrained windows (Gnocchi ≥ 4) as a function of the percentage of coding sequences in a window (left to right: N = 1,906/49,525, 3,244/55,676, 2,240/18,461, 1,506/7,094, 969/3,519, 569/1,946, 364/1,223, 283/910, 243/724, 10,392/30,138). The intervals (x-axis) are left exclusive and right inclusive. "Exonic only" refers to the 1kb windows created from directly concatenating coding exons into 1kb sequences. Error bars indicate standard errors of the proportions. **b**, The exonic-only regions (N = 27,875; purple) present a significantly higher Gnocchi score than regions that are exclusively non-coding (N = 1,843,559; blue). Dashed lines indicate the medians. **c**, The proportion of highly constrained windows (Gnocchi≥4) as a function of the proportion of exonic windows being added to the dataset of non-coding windows. **d**, Gnocchi score percentiles of non-coding versus exonic windows. About 0.05% (100-99.95%) and 3.12% (100-96.88%) of the non-coding windows exhibit similar constraint to the 90th and 50th of exonic regions, respectively.



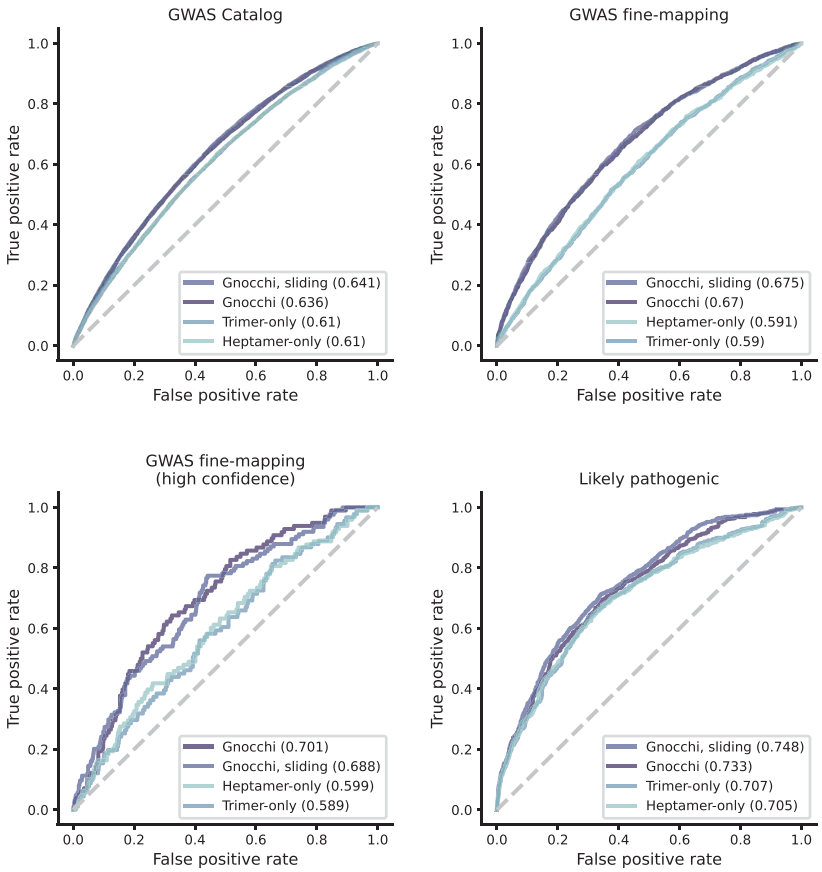
Extended Data Fig. 3 | Estimation of constraint for aggregated regulatory annotations. a,**b**, Gnocchi scores of aggregated promoter (dark purple), enhancer (light purple), microRNA (miRNA; dark blue), and long non-coding RNA (lncRNA; light blue) annotations are compared against those of exonic (**a**) and non-coding (**b**) regions at a 1kb scale. The Gnocchi score percentiles of each annotation (y-axis) are benchmarked by the score deciles of exonic or non-coding regions (10–100 percentiles; x-axis); the grey dashed vertical line indicates the median (50th percentile).



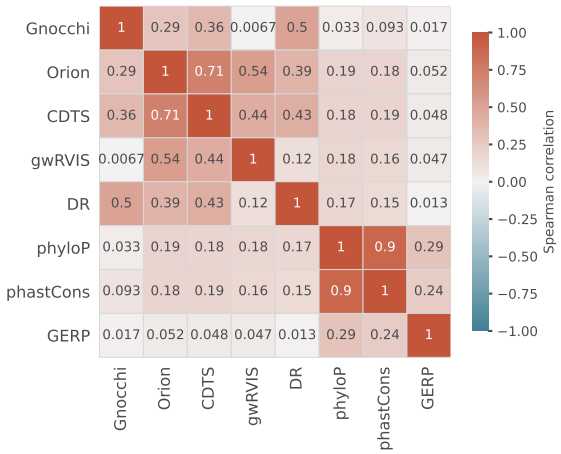
Extended Data Fig. 4 | Applications of Gnocchi for characterizing non-coding regions in addition to existing functional annotations. a, Use of Gnocchi for prioritizing non-coding regions with or without a regulatory annotation (N = 464,504 and 1,379,055, respectively). Constrained non-coding regions are enriched for GWAS variants, independent of the candidate cis-regulatory element (cCRE) annotation from ENCODE. Error bars indicate 95% confidence intervals of the odds ratios. **b**, Use of Gnocchi in statistical fine-mapping. The increase in posterior inclusion probability (PIP) when incorporating Gnocchi score as a functional prior into previous fine-mapping results (that used a uniform prior; denoted as $PIP_{Gnocchi}$ and PIP_{unif}, respectively) is shown for 164 new likely causal associations with a $PIP_{Gnocchi} \geq 0.8$ as a function of $PIP_{Gnocchi}$.



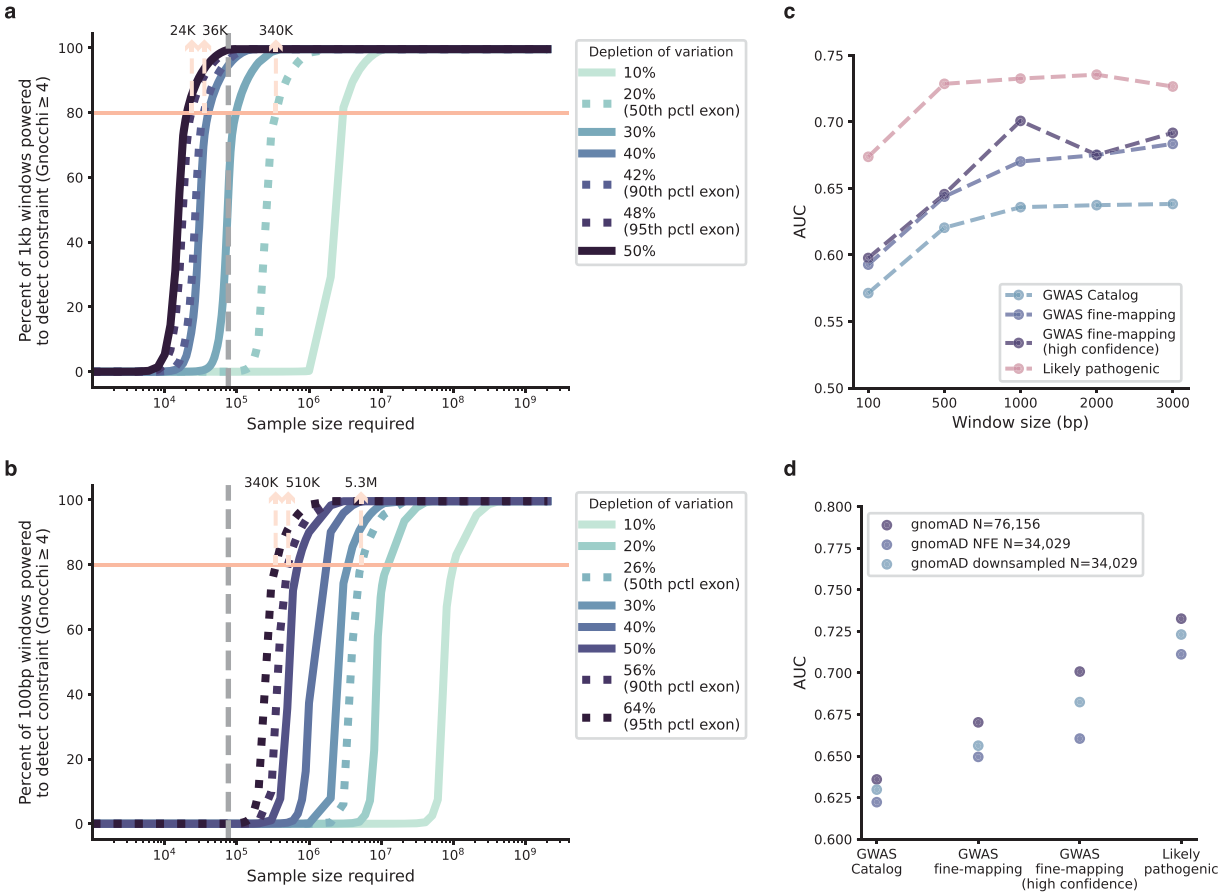
Extended Data Fig. 5 | Comparison of Gnocchi and other predictive metrics in prioritizing non-coding variants. a, Receiver operating characteristic (ROC) curves of Gnocchi and other seven metrics in classifying putative functional non-coding variants ("positive" variant set) – left to right: 9,229 GWAS Catalog variants, 2,191 GWAS fine-mapping variants, a subset of 140 high-confidence fine-mapped variants, and 1,026 likely pathogenic variants – against "negative" variant set randomly drew from the population with a similar allele frequency (AF). AF>5% and allele count (AC) = 1 were applied respectively for matching the three GWAS variant sets and the likely pathogenic variant set, based on their AF distributions in TOPMed (shown in **b**). **b**, AUCs of the classification with a varying AF threshold for the negative variant set. As most GWAS variants are common and most likely pathogenic variants are very rare (not seen in the population), AF>5% and AC = 1 were applied respectively in the primary analyses shown in **a**.



Extended Data Fig. 6 | Comparison of constraint scores built from different mutational models and genomic windows. Gnocchi (presented in this study) outperforms the scores rebuilt from mutational models that only consider local sequence context – trinucleotide (trimer-only) or heptanucleotide (heptamer-only) – without adjustment on mutation rate by regional genomic features, and the performance is robust to the artificial break of genomic windows when computed at a 1kb sliding by 100bp scale.



Extended Data Fig. 7 | Pairwise correlations between different constraint/conservation metrics. The Spearman's rank correlation between each pair of the eight metrics was computed based on the mean value of each score on 1kb windows across the genome.



Extended Data Fig. 8 | Power of constraint detection. a,**b**, The sample size required for well-powered non-coding constraint detection. The percentage of non-coding regions powered to detect constraint (Gnocchi ≥ 4) at a 1kb (**a**) and 100bp (**b**) scale under varying levels of selection (depletion of variation) is shown as a function of log-scaled sample size. Lighter color indicates milder deletion of variation (weaker selection), which requires a larger sample size to detect constraint; the grey dashed vertical line indicates the current sample size of 76,156 genomes. Dotted curves (left to right) benchmark the 95th, 90th, and 50th percentile of depletion of variation observed in coding exons of similar size. The number of samples required to obtain an 80% detection power is labeled at corresponding benchmarks. **c**, AUCs of Gnocchi scores computed on different window sizes in identifying putative functional non-coding variants. 1kb (used in this study) presents the optimal window size with high performance while maintaining reasonable resolution. **d**, AUCs of Gnocchi scores computed from different subsets of gnomAD in identifying putative functional non-coding variants. While with an equal sample size, the downsampled dataset with diverse ancestries presents higher performance than the Non-Finnish European (NFE)-only dataset.

nature portfolio

Corresponding author(s): Konrad Karczewski
Siwei Chen

Last updated by author(s): Mar 23, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for the collection of data, as this was an opportunistic study.
Data analysis	All code to perform quality control of the resource is publicly available at https://github.com/broadinstitute/gnomad_qc, and many of the functions are documented in a Python package (gnomad v.0.4.0) at https://broadinstitute.github.io/gnomad_methods/index.html. The code to compute the constraint statistics is available at https://github.com/atgu/gnomad_nc_constraint. bedtools v2.29.0: https://bedtools.readthedocs.io/en/latest/. CrossMap v0.6.1: https://crossmap.readthedocs.io/en/latest/. bigWigAverageOverBed v2: http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We release the aggregated allele frequency dataset at https://gnomad.broadinstitute.org, in a browser and bulk downloads for VCFs and Hail Tables, as well as all constraint statistics described in this manuscript. Additionally, we provide a subset of the dataset that includes individual level data for the HGDP65 and the 1000 Genomes projects66: the generation and use of this dataset is described in a companion manuscript67. There are no restrictions on the aggregate data released. External datasets used in this study are available in the following public resources: ENCODE cCREs https://screen-v2.wenglab.org/, super enhancers http://www.licpathway.net/sedb/download.php, FANTOM5 enhancers https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/extra/enhancer/, miRNA https://genome.ucsc.edu/cgi-bin/hgTables (All GENCODE V32 track), FANTOM5 lncRNA https://fantom.gsc.riken.jp/cat/v1/#/genes, GWAS Catalog https://genome.ucsc.edu/cgi-bin/hgTables (GWAS Catalog track), GWAS fine-mapping https://www.finucanelab.org/data, CNV morbidity map of developmental delay https://genome.ucsc.edu/cgi-bin/hgTables (Development Delay track), ClinVar https://genome.ucsc.edu/cgi-bin/hgTables (ClinVar Variants track), TOPMed https://bravo.sph.umich.edu/freeze8/hg38/downloads, ClinGen https://genome.ucsc.edu/cgi-bin/hgTables (ClinGen track), MGI https://www.informatics.jax.org/, OMIM https://www.omim.org/, Roadmap Epigenomics Enhancer-Gene Linking https://ernstlab.biolchem.ucla.edu/roadmaplinking/, GTEx https://gtexportal.org/home/datasets.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The term 'sex' was used to indicate biological attribute. The method of sex inference is described Supplementary Information, and the sex assignment is provided in Supplementary Table 1; there are in total 46,361 females and 45,129 males inferred the initial dataset, and 8,947 females and 37,209 males were included in the downstream analysis after sample QC as well as with permissions for public release of aggregate data. No sex-based analyses were performed in this study.
Population characteristics	As an opportunistic collection of data, the participants in this study were not selected based on age, sex or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. The populations are provided in Supplementary Table 3. These data were obtained primarily from case-control studies of adult-onset common diseases, including cardiovascular disease, type 2 diabetes, and psychiatric disorders.
Recruitment	As this was an opportunistic secondary use study, we did not recruit any participants.
Ethics oversight	This study was overseen by the Broad Institute's Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study was opportunistic, and involved secondary use of all available genome and exome data. No sample size was predetermined.
Data exclusions	Sample QC and variant QC for gnomAD are described extensively in the Supplementary Information. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded, as previously established and described [Lek et al., 2016].
Replication	We did not attempt to reproduce any findings in a separate dataset, as no other open-access data set of comparable size exists.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.

Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging