A genome-wide mutational constraint map quantified from variation in 76,156 human genomes

Siwei Chen[1,2,†], Laurent C. Francioli[1,2,†], Julia K. Goodrich[1,2], Ryan L. Collins[1,3,4], 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[1,2], Mary T. Yohannes[1], Zan Koenig[1], Yossi Farjoun[6], Eric Banks[6], Stacey Donnelly[7], Stacey Gabriel[1,7], Namrata Gupta[1,7], Steven Ferriera[7], 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], gnomAD Project Consortium[*], Anne O'Donnell-Luria[1,2,8], Matthew Solomonson[1,2], Cotton Seed[2,9], Alicia R. Martin[1,2], Michael E. Talkowski[1,3], Heidi L. Rehm[1,3], Mark J. Daly[1,2,10], Grace Tiao[1,2], Benjamin M. Neale[1,2,9,‡], Daniel G. MacArthur[1,2,11,12,‡], Konrad J. Karczewski[1,2,9]

[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]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[8]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA
[9]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[10]Institute for Molecular Medicine Finland, (FIMM) Helsinki, Finland
[11]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, Australia
[12]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia
[*]Lists of authors and their affiliations appear at the end of the paper

[‡]These authors contributed equally: Benjamin M. Neale, Daniel G. MacArthur.
[†]These authors contributed equally: Siwei Chen, Laurent C. Francioli.
Correspondence should be addressed to K.J.K (konradk@broadinstitute.org) and S.C (siwei@broadinstitute.org)

Abstract

The depletion of disruptive variation caused by purifying natural selection (constraint) has been widely used to investigate protein-coding genes underlying human disorders, but attempts to assess constraint for non-protein-coding regions have proven 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 reference dataset, and use this dataset to build a mutational constraint map for the whole genome. We present a refined mutational model that incorporates local sequence context and regional genomic features to detect depletions of variation across the genome. As expected, protein-coding sequences overall are under stronger constraint than non-coding regions. Within the non-coding genome, constrained regions are enriched for regulatory elements and variants 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 genes, while non-coding constraint captures additional functional

information underrecognized by gene constraint metrics. We demonstrate that this genome-wide constraint map provides an effective approach for characterizing the non-coding genome and improving the identification and interpretation of functional human genetic variation.

Introduction

The expansion in the scale of human whole-genome or exome reference data has characterized 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 modeling "constraint," the reduction of variants in a gene compared to an expectation conditioned on that gene's mutability. Using coding variant data from sequencing of more than 125K humans[1], we have previously developed a constraint metric that classifies each protein-coding gene along a spectrum of LoF intolerance[1], which provides a valuable resource for studying the functional significance of human genes[2-5]. Although of outsized biological importance, protein-coding regions 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[6-10].

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 to the exome, limiting the power of detecting depletions of variation at a fine scale. Second, our detailed understanding of coding region exon structure and effect of specific variants on amino acid translation enables a precision not available in non-coding analysis. Third, strong expectation from Mendelian genetics and existing constraint analyses that the coding regions, while a small fraction of the genome, harbor a massively disproportionate amount of rare and common disease mutations under selection. Fourth, the mutation rate in non-coding regions is highly heterogeneous, which can be affected not only by local sequence context as commonly modeled in gene constraint metrics but also by a variety of genomic features at larger scales[11,12].

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 (e.g., Orion[13], CDTS[14]); 2) machine-learning classifiers that are trained to differentiate between disease-associated variants and benign variants (e.g., CADD[15], GWAVA[16], gwRVIS[17]); 3) evolutionary methods that employ phylogenetic conservation scores to predict the consequence of non-coding variants on fitness (e.g., LINSIGHT[18]). While all these methods aid in our understanding of the non-coding genome, each suffer from limitations/biases, respectively as 1) overlooking the influence of regional genomic features beyond the scale of flanking nucleotides on mutation rate; 2) strong dependence on the availability of well-characterized functional mutations as training data; 3) compromised ability to detect regions that have only been under selection recently in the human lineage, which may have a functional impact on human phenotypic traits or diseases.

Here we present 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 analyzes local sequence context and regional genomic features, and quantify depletion of variation in a tiled window across the entire genome. By building a fuller picture of genic constraint than solely focusing on coding variation, we demonstrate that it facilitates the functional interpretation of non-coding regions and improves the characterization of gene function in the context of the regulatory network. . Our study aims to depict a genome-wide view

of how natural selection shapes patterns of human genetic variation and generate a more comprehensive catalog of functional genomic elements with potential clinical significance.

Results

Aggregation and quality control of genome sequence data

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 pediatric disease, and with appropriate consent and data use permissions for the sharing of aggregate variant data. Among these samples include 36,811 (48.3%) of non-European ancestry, including 20,744 individuals with African ancestries and 7,647 individuals with admixed Amerindigineous ancestries. After stringent quality control (see Supplemental Information), we were left with a set of 644,267,978 high-confidence short nuclear variants (gnomAD v3.1.2), of which 390,393,900 rare (allele frequency [AF]≤1%) single nucleotide variants were used for building the genome-wide constraint map. These correspond to approximately one variant every 4.9 bp (one rare variant every 8 bp) of the genome, providing a high density of variation.

Quantifying mutational constraint across the genome

To construct a genome-wide mutational constraint map, we divided the genome into continuous non-overlapping 1kb 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 (e.g., ACG -> ATG), with adjustment for the effect of methylation on mutation rate at CpG sites, which become saturated for mutation at sample sizes above ~10K genomes[19] (Extended Fig. 1a,b; Methods). We then jointly analyzed these local-context-specific estimates with regional genomic features (e.g., replication timing and recombination rate) to establish an expected number of variants. We trained our model on 413,304 *de novo* mutations previously detected in family-based whole-genome sequencing studies[20,21] (Extended Fig. 1c) and applied it to assess constraint across the entire genome (Methods).

We quantified the deviation from expectation for each 1kb window using a Z score[22] (Methods; Extended Fig. 1d,e), which was centered slightly below zero for non-coding regions (median=-0.50), and was significantly higher (more constrained) for windows containing any protein-coding sequences (median=0.17, Wilcoxon $P<10^{-200}$; **Fig. 1a**). The constraint Z score positively correlated with the percentage of coding bases in a window and presented a substantial shift towards higher constraint for exonic sequences calculated from directly concatenating coding exons into 1kb windows (median=1.48; Extended Fig. 2a-c). About 13.5% and 0.5% of the non-coding windows exhibited constraint as strong as the 50th and 90th percentile of exonic regions (Extended Fig. 2d). Comparing our Z score against the adjusted proportion of singletons (APS) score, a measure of constraint developed for structural variation (SV)[23], we found a strong correlation (linear regression $P=1.54\cdot10^{-40}$, **Fig. 1b**; Methods), providing an internal validation of our approach.

Investigating genomic properties of non-coding regions under constraint

To further validate our constraint metric and investigate the functional relevance of non-coding regions under selection, we examined the correlation between our Z score and several annotations of putatively functional non-coding sequences (**Fig. 2a**). First, we found that candidate cis-regulatory elements (cCREs, derived from ENCODE[24] integrated DNase- and ChIP-seq data) are significantly enriched in the most constrained percentile of the genome (Z≥4, OR=1.28 compared to the genome-wide average, Fisher's exact $P=2.72\cdot10^{-44}$), indicating that a large fraction of the constrained non-coding regions may serve a regulatory role. Similarly, a significant enrichment was found for an independent set of active, *in vivo*-

transcribed enhancers (identified by FANTOM CAGE analyses[25]; OR=1.36, P=1.03·10^{-12}). Moreover, we found that super enhancers[26] – groups of enhancers in close genomic proximity – exhibited a stronger enrichment (OR=1.44, P=4.98·10^{-110}), which is in line with their prominent roles in regulating genes important for cell type specification[27]. We also detected a modest but significant enrichment for long non-coding RNAs (lncRNAs, implicated by the FANTOM CAT catalog[28]; OR=1.21, P=2.71·10^{-25}), lending support to their potential functionality in the pervasive transcription in the human genome[29].

A pronounced level of constraint was observed for the 9p21 locus (OR=7.10, Fisher's exact P=2.90·10^{-4}; **Fig. 2b**), a widely recognized and replicated risk factor for coronary artery disease (CAD) and type 2 diabetes (T2D)[30]. Although devoid of genes, the locus tended to be as constrained as regions encompassing coding sequences (median Z=0.38 versus 0.17, Wilcoxon P=0.19; P=2.11·10^{-4} when compared to the rest of the genome). The signal of constraint was correlated with the presence of genetic markers identified by CAD/T2D genome-wide association studies[31] (GWAS; median Z=2.04 versus 0.003, Wilcoxon P=0.017) and coincided with the high density of regulatory elements[24] within the locus (median Z=0.74 versus -0.70, Wilcoxon P=0.011; **Fig. 2c**). A well-replicated CAD marker rs10738607[32-34], for example, was found to reside in a region under strong selection (Z=4.48, chr9:22088000-22089000) ~80kb upstream of the cell cycle suppressor gene *CDKN2B*, where three consecutive enhancer signatures occupied >80% of its sequence[24]. The risk allele of rs10738607 was predicted to disrupt a binding site for STAT5A[30], a known transcriptional activator of *CDKN2B* that negatively regulates cell proliferation[35,36]; reduced expression of *CDKN2B* and enhanced proliferation have been previously-recognized phenotypes linked to the 9p21 CAD risk locus[37-41]. Therefore, our constraint analysis adds weight to the hypothesis that loss of *STAT5A-CDKN2B* antiproliferative effect may present one mechanism that explains the genotype-phenotype association for 9p21 risk alleles.

Prioritizing non-coding variants through the constraint map

The genome-wide constraint map allows us to systematically evaluate each genetic variant in the genome, particularly expanding our ability to study variants in the vast and under-characterized non-coding regions. Examining the distribution of non-coding variants identified by GWAS on the constraint spectrum, we found a significant enrichment for non-coding GWAS hits in the constrained end of the genome (621/14,808 constrained windows [Z≥4] overlapped with GWAS Catalog[31] annotations, OR=1.47 compared to the genome-wide average of 48,045/1,665,599, Fisher's exact P=3.76·10^{-19}, **Fig. 3a**; Methods). The enrichment appeared to become increasingly stronger for hits that had a replication experiment (internally replicated within the same study: OR=1.54, P=1.39·10^{-9}; externally replicated by a different study: OR=1.76, P=1.51·10^{-6}). Furthermore, substantial selection signals were found for likely causal GWAS variants fine-mapped from 94 complex diseases and traits in UK Biobank (UKB)[42,43] (**Fig. 3b**; Methods). Leading enrichments were found for pathological conditions including CAD, inguinal hernia, insomnia, fibroblastic disorders, and hypothyroidism (OR>2.5), as well as quantitative molecular traits such as hemoglobin, uric acid, and bilirubin levels. These results revealed a high positive correlation between the level of constraint and the occurrence of candidate functional variants in the non-coding genome. Thus, our constraint metric could effectively serve to prioritize non-coding variants discovered in large-scale human disease or trait association studies.

For example, in the constraint analysis of CAD GWAS, we found seven variants from a 95% credible set located in highly constrained non-coding regions related to the gene *PLG* – four were mapped to its antisense transcript ENST00000659713 (rs144679016, rs116480834, rs10945688, rs148517610), one resided in its 9th intron (rs143556831), and two colocalized with its enhancer marks ~55-57kb downstream of the gene (rs12529023, rs139095347). *PLG* encodes the plasminogen protein that circulates in blood

plasma and is converted to the active protease, plasmin, which dissolves the fibrin of blood clots. Dysregulation of the PLG-plasmin system has been frequently implicated in the pathogenesis of CAD[44-49]. Therefore, together with previous evidence on *PLG*, our findings suggest a plausible role of the prioritized non-coding variants in increasing CAD risk, likely acting through changes in the constrained regulatory elements of *PLG*.

Exploring non-coding dosage sensitivity in the constrained genome

Copy number variants (CNVs) or dosage alterations (deletions/loss or duplications/gain) of DNA are well-established risk factors for human developmental disorders[50-55]. CNVs can exert pathogenic effects in more than one mechanism[56,57], such as direct gene dosage effect, positional effect, and transvection effect. Still, the most common is alteration of dosage-sensitive genes within the loci, whereas the functional relevance of the large proportion of non-coding sequences being affected are much less understood. With 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 ~90K CNVs from a genome-wide CNV morbidity map of developmental delay and congenital birth defects[58,59]. We observed a substantial excess of CNVs that affect constrained non-coding regions (Z≥4) among individuals with developmental disorders (DD cases) in comparison to healthy controls (42.2% versus 10.9%, OR=6.0, Fisher's exact $P<10^{-200}$, **Fig. 4a**; Methods). More strikingly, of the 18 loci that had been previously classified as pathogenic[58], 16 (88.9%) coincided with constrained non-coding regions. The high incidence was recapitulated in a curated set of ~4K putatively pathogenic CNVs (ClinVar[60]), of which 82.8% exhibited high non-coding constraint (**Fig. 4a**). Importantly, 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 subtype of CNV deletions and duplications (**Fig. 4b**; Methods). Non-coding constraint manifested a comparable level of association as gene constraint when modeled simultaneously to predict DD CNVs (log[OR]=1.13 and 1.48, Logit $P<2\cdot10^{-16}$ for both), suggesting that dosage imbalance of constrained non-coding regions could confer risk for developmental disorders in addition to gene dosage sensitivity.

One classic example of non-coding dosage imbalance is a set of duplications involving the regulatory domain of *IHH* associated with synpolydactyly and craniosynostosis[61-63]. The four implicated duplications covered a stretch of ~102kb sequence upstream of *IHH*, with a smallest ~10kb critical overlapping region (**Fig. 4c**). The critical region contained no exons but showed high non-coding constraint (median Z=1.22, Wilcoxon P=0.011 compared to the rest of the genome). Interestingly, the highest constraint signal colocalized 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[63]. Collectively, we suggest that non-coding constraint can be a useful indicator of dosage-sensitive regulatory CNVs and improves understanding of potential non-coding mechanisms underlying CNV disease associations.

Leveraging non-coding constraint to improve gene function characterization

Given the significant roles of constrained 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 surveyed enhancer-gene links from the Roadmap Epigenomics Enhancer-Gene Linking database[64] (Methods), where we found overall, more constrained non-coding regions were more frequently linked to a gene (**Fig. 5a**), and more constrained enhancers tend to be associated with more constrained genes (e.g., haploinsufficient genes; median Z=2.21 versus 1.43, Wilcoxon $P=4.24\cdot10^{-18}$, **Fig. 5b**; Methods).

On the other hand, a particular interesting set of associations are the links between constrained enhancers and the "unconstrained" genes classified by gene constraint metrics, because these links may reflect functional significance of the "unconstrained" genes that had been previously underrecognized. More than 40% of the least constrained genes (last decile scored by LOEUF [loss-of-function observed/expected upper bound fraction][1]) were linked with a constrained enhancer (last decile with constraint Z≥2.23); the lack of predicted gene constraint can be explained by the intrinsic design of LOEUF as a measure of intolerance to rare LoF variation, where small genes with few expected LoF variants are likely underpowered. Indeed, when stratifying genes by the number of expected LoF variants, we found a significantly higher enhancer constraint for genes that were underpowered (with ≤5 expected LoF variants by LOEUF) compared to those that were sufficiently powered and scored as unconstrained (median Z=2.06 versus 1.68, Wilcoxon $P=1.68\cdot10^{-3}$, **Fig. 5b**). This pattern suggests that a portion of these underpowered genes may play an important functional role but were not recognized in gene constraint evaluation. For instance, *ASCL2*, a basic helix-loop-helix (bHLH) transcription factor, had only 0.57 expected LoFs (versus 0 observed) across >125K exomes[1]; although being depleted for LoF variation, the absolute difference was too small to obtain a precise estimate of LoF intolerance. Yet, we found *ASCL2* had a highly constrained enhancer (Z=6.26), located ~16kb upstream of the gene, where >40% of the expected variants were depleted (200.6 expected versus 112 observed, chr11:2286000-2287000). The same genomic window also contained an eQTL chr11:2286192:G>T that was predicted to be significantly associated with *ASCL2* expression[65]; elevated *ASCL2* expression has been implicated in the development and progression of several human cancers[66-68]. This example highlights the value of non-coding constraint – as a complementary metric to gene constraint – for identifying functionally important genes.

Gene-set enrichment analysis of the underpowered genes with constrained enhancers identified strong enrichment for genes encoding histones, secreted proteins involved in cell signaling and communication (e.g., cytokines, hormones, growth factors; **Fig. 5c**). These genes are generally small and are thus likely to be underrecognized by LoF-intolerance-based metrics. Applying an alternative, phylogeny-based conservation score[69] to evaluate these underpowered genes, we found a significantly higher evolutionary constraint on these genes compared to those with less constrained enhancers (Wilcoxon $P=4.28\cdot10^{-12}$; **Fig. 5d**), providing additional evidence for their functional significance. Moreover, by dissecting enhancers in a tissue-specific manner, we examined how tissue-specific enhancer constraint correlates with tissue-specific gene regulation. We found that even conditioning on gene constraint, enhancer constraint remained a significant predictor for the expression level of target genes in matched tissue types (**Fig. 5e**; Methods). These results further support the application of our constraint metric to improve the characterization of human gene function.

A practical implementation of this notion would be refining gene constraint models to incorporate contributions from constrained regulatory elements. This essentially borrows power from extending the functional unit of a gene in the context of the gene regulatory network. Such an approach would allow constraint modeling for specific tissue/cell types and conditions given the diverse range of epigenomics data. Taking enhancer-gene links from brain tissue as a proof-of-principle, a simple logistic regression test conditioning on gene constraint indicated significant contribution from enhancer constraint for predicting functionally relevant genes (e.g., targets of Fragile X Mental Retardation Protein (FMRP)[70], Logit $P=2.84\cdot10^{-8}$; Methods). While we acknowledge that the biological consequences of mutations in enhancers are not clearly understood and thus natural selection may differ in its interest depending on mechanistic consequence, an extended model to incorporate non-coding variation information in a biologically-informed way hold the promise to provide a finer characterization of gene function and facilitate a better 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 LoF variation for each protein-coding gene[1,19]. 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 modeling include unified fitting of mutation rate for all substitution and trinucleotide context and inclusion of regional genomic features to refine the expected variation in non-coding regions (Methods). We validated our metric using a series of external functional annotations, with a focus on the non-coding genome, and demonstrated the value of our metric for prioritizing non-coding elements and identifying functionally important genes. We have made the constraint scores publicly accessible via the gnomAD browser (https://gnomad.broadinstitute.org).

One key challenge in quantifying non-coding constraint is to distinguish between selection and neutral variation. To this end, we intended not to include features that are directly reflective of regulatory functions in our model, such as histone modifications as commonly used in classifiers predicting functional non-coding variants. Instead, we employed such features as part of our validation, where we showed that our constraint Z score is positively correlated with the regulatory elements derived from epigenomic signatures. Meanwhile, to demonstrate the ability of our metric in prioritizing non-coding variants independent of established regulatory elements, we re-examined the enrichment analyses with previously annotated regulatory sequences excluded – all results persisted (Extended Fig. 3a,b). Likewise, we demonstrated that our constraint metric captures additional information to phylogeny-based conservation scores, which are likely to be blind to functional non-coding regions that have high evolutionary turnover (e.g., enhancers[71-76]; Extended Fig. 4). We also note that we restricted all our validation analyses to non-coding regions to explicitly evaluate the metric for characterizing the non-coding genome, and we further eliminated potential bias from nearby genes by recapitulating the results within regions >10kb away from any protein-coding genes (Extended Fig. 5). Further validating our metric on experimental data from multiplex assays of variant effect (MAVE)[77], we showed that our constraint Z score correlates well with the experimental measurements, along with comparison to other genome-wide predictive scores (Extended Fig. 6; Methods). Altogether, our constraint metric presented reliable and consistent performance in identifying important non-coding regions in the human genome.

Our analyses revealed considerable correlation between constraint of the non-coding regulatory elements and the functional importance of their target genes. The implication is twofold. First, the constraint metric can be applied to further prioritize existing regulatory annotations. For instance, categorizing ENCODE cCREs by constraint Z score identified an increasingly stronger GWAS signals in the more constrained cCREs (Extended Fig. 3c). This also suggests the integration of multiple lines of evidence indicative of function. Second, the prioritization of regulatory elements can be applied to improve identification of important genes. The example of *ASCL2* demonstrates how leveraging the non-coding functional variation of a gene can build out a clearer picture of its importance from a regulatory point of view. The analyses of tissue-specific enhancer constraint further demonstrate how non-coding constraint could be applied to refine gene constraint modeling.

Despite the clear constraint signal identified for non-coding regions, many limitations exist. First, the mutational model explains ~75% of the variation (R^2=0.746) in *de novo* mutation rate, indicating

underrecognized sources or properties of spontaneous human mutation. This situation can be partially alleviated by an increased sample size of *de novo* mutation data, which would allow inclusion of additional genomic features at smaller scales as well as more accurate estimation of feature contributions. Even with a sufficiently large dataset, however, comprehensive modeling of the heterogeneity in the mutation rate requires advanced knowledge about the underlying mutational mechanisms and ideally a more interpretable, biologically-informed statistical framework. Further, the practical 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/cell type, developmental stage, or environment. Such information is not directly built in current constraint scores nor in the mutational dataset, thus downstream analysis of functional genomics data (e.g., incorporating tissue-specific enhancer signatures as in this study) is often necessary for justifying specific biological implications. It should also be noted that, since the detection of depletion of variation is immune to negative selection after reproductive age, genomic regions involved in late-onset phenotypes are therefore likely to go underrecognized.

Finally, while this is the largest dataset of human genomes examined to date for non-coding constraint, our method will substantially increase in power and resolution as sample sizes increase.Using constraint seen in coding regions as benchmarks (Extended Fig. 7), we are currently well-powered to detect non-coding regions with constraint as strong as the 90th percentile of coding exons of similar size, and we estimate a sample size of ~700K to detect constraint that is about the average of coding regions. Moreover, as functional non-coding elements are generally small, to further increase our resolution, for example at a 100bp scale, we would need ~4.8M individuals to detect individual constrained non-coding elements.

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

Methods

Genome data aggregation, variant-calling, and quality control

We aggregated whole genome sequence data from 153,030 individuals spanning projects from case-control consortia and population cohorts, in a similar fashion to previous efforts[1]. We harmonized these data using the GATK Best Practices pipeline and joint-called all samples using Hail[78], and developed and utilized an updated pipeline of sample, variant, and genotype QC 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).

Mutation rate model

To estimate the baseline mutation rate for each substitution and context, we tallied each trinucleotide context in the human genome. As previously shown[19], the methylated CpG variants begin to saturate at a sample size of ~10K genomes, and therefore we downsampled the gnomAD dataset to 1,000 genomes for use in calculating the mutation rate. Sites were further excluded if there were variants observed in gnomAD but flagged as low quality, or found in greater than 5 copies in the downsampled set. Using this dataset, we computed the proportion of possible variants observed for each trinucleotide change (XY_1Z -> XY_2Z), with CpG sites stratified by their methylation levels (see DNA methylation analysis). These proportions represent the relative probability of a given nucleotide mutating into one of the three other possible bases in a trinucleotide context, and we scaled this factor so that the weighted genome-wide average is the human per-base, per-generation mutation rate ($1.2 \cdot 10^{-8}$) to obtain the absolute mutation rates $\int$. The $\int$ estimates were well correlated with the proportion of possible variants observed in the 76,156 gnomAD genomes (R^2=0.999, Extended Fig. 1a,b; Supplementary Data 1), a dataset of 6,079,733,538 possible variants at 2,026,577,846 autosomal sites with 30-32X coverage, where 390,393,900 high-quality rare (AF≤1%) variants were observed. These fitted proportion observed (mutabilities) were then used for establishing the context-dependent expected variation in construction of constraint Z scores.

DNA methylation analysis

To correct for the effect of DNA methylation on 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 (from averaging different tissues[1]), we analyzed methylation data from germ cells across 14 developmental stages, comprising eight from preimplantation embryos (sperm, oocyte, pronucleus, two-cell-, four-cell-, eight-cell-, morula-, and blastocyst-stage embryos)[79] and six from primordial germ cells (7Wk, 10Wk, 11Wk, 13Wk, 17Wk, and 19Wk)[80]. 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. In order 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.

Construction of constraint Z score

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

$$\chi^2 = (Obs - Exp)^2/Exp$$
$$Z = \{\sqrt{\chi^2}\ \ if\ Obs < Exp\ \ -\sqrt{\chi^2}\ \ if\ Obs \geq Exp$$

Here, the observed variant count (Obs) is the number of unique rare (AF<0.1%) variants in a 1kb window identified in the gnomAD dataset. To establish the expected number of variants (Exp), we first created a context-dependent mutability for each window by summing all trinucleotide mutabilities (proportion observed modeled by ⌈) within the 1kb sequence. At the same time, we computed for each window 17 genomic features that have been shown to influence mutation rate (e.g., replication timing and recombination rate; see Collection of genomic features). To determine the relative contribution of the 18 features (trinucleotide context plus 17 genomic features), we trained these features to predict the occurrence of *de novo* mutations (DNMs), as a proxy of spontaneous mutations, using a random forest regression model. A set of 413,304 unique DNMs were compiled from two large-scale family-based whole-genome sequencing studies{Halldorsson, 2019 #17;An, 2018 #18}. Given the data sparsity, we counted the incidence of DNMs per 1Mb and regressed it against the 18 features computed correspondingly (R^2=0.746 by a 90/10 train/test split, Extended Fig. 1c). The fitted model was then applied to establish the expected variation (Exp) in the gnomAD dataset.

Constraint Z scores were created for 2,442,347 non-overlapping 1kb windows across the human genome (2,330,252 on autosomes and 112,095 on chromosome X). Due to the lack of DNM data on chromosome X, a model was trained using autosomal regions and was extrapolated to chromosome X for computing constraint scores. We performed downstream analyses separately for autosomes and chromosome X and presented the former as primary, with the latter provided in Extended Fig. 8. In 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, and 3) the mean coverage in the gnomAD genomes was between 25-35X (or 20-25X for chromosome X). This resulted in 1,797,153 windows (73.6% of initial) for the primary analyses (Supplementary Data 2), with 131,554 encompassing coding sequences and 1,665,599 exclusively non-coding.

Collection of genomic features

The 17 regional genomic features used for predicting the expected variation are 1) replication timing[81], 2) male and 3) female recombination rate[20], 4) GC content[82], 5) low-complexity region[83], 6) short and 7) long interspersed nuclear element[82], distance from the 8) telomere and the 9) centromere[82], 10) CpG island[82], 11) nucleosome density[81], 12) maternal and 13) paternal DNM cluster[84], DNA methylation in 14) sperm[79], 15) oocyte[79], 16) preimplantation embryo[79], and 17) primordial germ cell[80]. Data were downloaded from the referenced resources, lifted over to GRCh38 coordinates when needed using CrossMap[85], and files in .bed or .BigWig format were processed using bedtools[86] and bigWigAverageOverBed[81] to obtain feature values within specific genomic windows.

Correlation between constraint Z and APS

As an internal validation, we compared our constraint Z score against the SV constraint score APS[23]. For each SV from the original study (gnomAD-SV), we assessed its constraint by assigning the highest Z score among all overlapping 1kb windows. The correlation between constraint Z and APS was evaluated across 205,699 high-quality autosomal SVs 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 Z bins, with 95% confidence intervals computed from 100-fold bootstrapping.

Correlation of constraint Z with candidate functional non-coding elements

We validated the constraint metric using a number of external functional annotations, including 926,535 ENCODE cCREs[24], 63,285 FANTOM5[25] enhancers, 19,175 FANTOM6[28] lncRNAs, 331,601 super enhancers (SEdb[26]), 111,308 GWAS Catalog[31] variants (with an association *P*-value$\leq 5.0 \cdot 10^{-8}$; 32,708 with an internal replication and 6,748 with an external replication), 601,191 UKB GWAS variants fine-mapped from 94 heritable traits (95% credible set)[42], and 90,423 unique CNVs from a CNV morbidity map of developmental delay[58,59].

To assess the correlation between constraint Z scores and the collected candidate functional elements, we intersected each annotation with the scored 1kb windows binned by constraint Z (<-4, [-4,-3), [-3,-2), [-2,-1), [-1,-0), [0,1), [1,2), [2,3), [3,4), $\geq$4), and counted the frequency of overlapping windows within each bin. The enrichment of a given annotation (except CNVs) at a constraint level was evaluated by comparing the corresponding frequency to the genome-wide average using a Fisher's exact test. In the analysis of CNVs, we assessed their enrichment in constrained regions by assigning each CNV the highest Z score among its overlapping windows and comparing the proportions of constrained CNVs (Z$\geq$4) from cases of developmental delay and healthy controls. The enrichment was further examined using a logistic regression model to adjust for the size and gene content (gene constraint[1] 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 our constraint metric in characterizing the non-coding genome.

Coordination between non-coding constraint and gene constraint

To examine whether constraint of non-coding regulatory elements implicates the constraint of their target genes, we compared constraint Z scores of enhancers linked to constrained genes and unconstrained genes. The former included well-established gene sets of 189 ClinGen[87] haploinsufficient genes, 2,454 MGI[88] essential genes mapped to human orthologs, 1, 771 OMIM[89] autosomal dominant genes, and 1,920 LOEUF[1] first-decile genes; and the later 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). The LOEUF underpowered list included 1,117 genes with $\leq$5 expected LoF variants. Enhancers associated with 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[90,91]. For each gene, we aggregated and merged enhancers predicted from all 127 reference epigenomes and assigned the most constrained enhancer to each gene for analysis of enhancer-gene constraint coordination (Supplementary Data 3).

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[64]). For each gene and tissue type, we searched for tissue-specific gene expression in the Genotype-Tissue Expression (GTEx[65]) database (RNASeQCv1.1.9) and computed a normalized median expression for each gene ($\log_2$(TPM+1)). Enhancer constraint and gene expression levels were calculated for 11 matched tissue types and their correlation within each tissue type was evaluated by regressing gene expression on the enhancer constraint of corresponding gene. Importantly, we included gene constraint (LOEUF score) as a covariate to test the additional contribution from enhancer constraint. Similarly, we used a logistic regression model conditioning on LOEUF to examine the value of brain-tissue enhancer constraint in identifying genes functional in brain. As a proof-of-principle, we regressed genes encoding the mRNA targets of Fragile X Mental Retardation Protein (FMRP[70]; 0 or 1) against the constraint of enhancers derived from brain tissue, including LOEUF score as a covariate.

Correlation of constraint Z and other predictive scores with MAVE

We compared our metric with other four genome-wide predictive scores – Orion[13], CDTS[14], gwRVIS[17], and JARVIS[17]) – in their correlation with experimental measurements on 11 enhancers tested by MAVE[77]. Each predictive score was downloaded from the original study, lifted over to GRCh38 (for Orion), and applied to score enhancers by taking the average over corresponding genomic regions. Enhancers were ranked by the proportion of mutations identified by MAVE as causing significant changes in gene expression (high to low: *SORT1*, *IRF4*, *IRF6*, *ZRS*, *ZFAND3*, *RET*, *TCF7L2*, *MYC* rs11986220 [*MYCs2*], *BCL11A*, *UC88*, *MYC* rs11986220 [*MYCs1*]), and the rank correlation between MAVE and each predictive score was evaluated using a Spearman's rank correlation test. The correlation was computed on all scorable enhancers for each score, as well as enhancers that are scorable by all metrics (n=7; *RET* and *IRF4* were not scored by constraint Z due to feature missingness near the telomere/centromere and *MYC*s1 and *SORT1* were excluded due to quality filtering yet their scores are provided in the "Unfiltered Z" penal. In addition, we excluded *UC88* in favor of other scores as it appeared to be an "outlier" deflecting their ranks (Extended Fig. 6).

Figure legends

Fig. 1: Distribution of constraint Z scores across the genome. **a**, Histograms of constraint Z scores for 1,797,153 1kb windows across the human autosomes. Windows encompassing coding regions (n=131,554 with $\geq$ 1bp coding sequence; red) overall exhibit a higher constraint Z (stronger negative selection) than windows that are exclusively non-coding (n=1,665,599; blue). **b**, The correlation between constraint Z score and the adjusted proportion of singletons (APS) score developed for structural variation (SV) constraint. A collection of 205,699 autosomal SVs were assessed using constraint Z score by assigning each SV the highest Z among all overlapping 1kb windows, which shows a strong positive correlation with the SV constraint metric APS. Error bars indicate 100-fold bootstrapped 95% confidence intervals of the mean values.

Fig. 2: Evaluation of constraint in putatively functional non-coding regions. **a**, Distributions of candidate regulatory elements along the spectrum of non-coding constraint. Enrichment was evaluated by comparing the proportion of non-coding 1kb windows, binned by constraint Z, that overlap with a given functional annotation to the genome-wide average. Error bars indicate 95% confidence intervals of the odds ratios. **b**,**c**, The level of constraint of the 9p21 locus, in comparison to coding regions (**b**) and in concordance with external functional annotations (**c**). For **b**, the kernel density estimate (KDE) plots display the distribution of constraint Z scores of the 9p21 locus (blue) and regions encompassing coding sequences (red); the scatter plot shows the enrichment of 9p21 locus across constraint Z bins, with error bars indicating 95% confidence intervals of the odds ratios. For **c**, each bar shows the constraint Z score of a 1kb window within the locus; gaps indicate windows removed by quality filters.

Fig. 3: Prioritization of non-coding variants for genome-wide association studies (GWAS). **a**, The distribution of GWAS Catalog variants on the non-coding constraint spectrum. Variants were further partitioned by the degree of experimental replication (Int/Ext repl: internal/external replication), with the latter exhibiting the highest constraint. **b**, Enrichment of UKB GWAS variants in constrained non-coding regions (Z$\geq$4). Analysis was performed separately for 94 complex diseases and traits, and results with nominal significance are shown, ordered by the lower bound of 95% confidence interval. For both panels, enrichment was evaluated by comparing the proportion of non-coding 1kb windows (at a given constraint level) that contain at least one GWAS variant to the genome-wide average. Error bars indicate 95% confidence intervals of the odds ratios; for **b**, only lower bounds are shown for presentation purposes.

Fig. 4: Contribution of non-coding constraint in evaluating copy number variants (CNVs). **a**, The proportion of constrained CNVs (Z$\geq$4) identified in individuals with developmental delay (DD cases) in comparison to healthy controls. The non-coding constraint of a given CNV was assessed by the highest constraint Z score among all non-coding windows being affected. CNVs were categorized and the frequency of constrained CNVs (Z$\geq$4) was compared based on pathogenic potential: constrained CNVs are more common in DD cases than controls (6,631/15,717=42.2% versus 8,107/74,706=10.9%) and are most frequent for CNVs previously implicated as pathogenic (16/18=88.9% by DD and 3,219/3,886=82.8% by ClinVar). **b**, The contribution of non-coding constraint in predicting CNVs in DD cases versus controls. Non-coding constraint remains a strong predictor for the case/control status of CNVs after adjusting for gene constraint (LOEUF score), gene number, and size of CNVs, using a logistic regression test. Error bars indicate 95% confidence intervals of the log odds ratios. **c**, CNVs at the *IHH* locus associated with synpolydactyly and craniosynostosis. The four implicated duplications (grey bars) overlap in a ~10kb region that exhibit high non-coding constraint (blue), with the highest Z score coinciding with the major

IHH enhancers (dark blue). Each blue bar shows the constraint Z score of a 1kb window within the locus; gaps indicate windows removed by quality filters.

Fig. 5: Relationship between non-coding constraint and gene constraint. **a**, The proportion of non-coding 1kb windows overlapping with enhancers that were predicted to regulate specific genes, as a function of their constraint Z scores. More constrained non-coding regions are more frequently linked to a gene. Error bars indicate standard errors of the proportions. **b**, Comparison of the constraint Z scores of enhancers linked to constrained and unconstrained genes. Enhancers of established sets of constrained genes (four blue boxes) are more constrained than enhancers of presumably less constrained genes (two grey boxes). Enhancers of genes that are underpowered for gene constraint detection ("LOEUF underpowered") present a higher constraint than those powered yet unconstrained genes ("LOUEF unconstrained"). **c**, Gene sets enriched for the LOUEF underpowered genes. Enrichment analysis was performed for a set of 506 genes underpowered by LOEUF while being linked with a constrained enhancer. Red dashed line indicates FDR=0.05. **d**, Comparison of evolutionary constraint for the LOUEF underpowered genes with and without a constrained enhancer. PhastCons, a phylogeny-based conservation metric, supports the functional importance of genes with constrained enhancers. **e**, Correlations between enhancer constraint and gene expression in specific tissue types. Enhancer constraint was reprocessed in a tissue-specific manner and was modeled to predict the expression level of target genes in matched tissue types; a linear regression test suggests significant contribution from enhancer constraint even conditioning on gene constraint (LOEUF score). Error bars indicate 95% confidence intervals of the beta coefficient estimates.

Extended Data Fig. 1: Construction of mutational model and constraint Z score. **a**,**b**, Estimating the relative mutability for each trinucleotide context across the genome. The proportion of possible variants observed in 76,156 gnomAD genomes (mutability; y-axis) is exponentially correlated with the absolute mutation rate estimated from 1,000 downsampled genomes (mu; x-axis). Fit lines were modeled separately for human autosomes (**a**) and chromosome X (**b**). **c**, Estimating the relative contribution of local sequence context and regional genomic features in predicting the expected variation. Mutational model incorporating trinucleotide context and 17 genomic features was trained on DNMs using a random forest regression model (with a 90/10 train/test split). Trinucleotide mutability appears the most important feature in predicting the expected variation. **d**,**e**, The distribution of constraint Z score as a function of expected and observed variation. Each point represents Z score for a 1kb window on the genome (n= 1,797,153 on autosomes (**d**) and n= 49,936 on chromosome X (**e**)), which quantifies the deviation of observed variation from expectation; positive Z (red) indicates deletion of variation (obs<exp) and the higher the Z the stronger the depletion (i.e., more constrained).

Extended Data Fig. 2: Comparison of constraint Z score between coding and non-coding regions. **a**, Exonic regions (1kb windows created by directly concatenating coding exons, n=26,987; purple) exhibit a significantly higher constraint Z than windows that are exclusively non-coding (n=1,665,599; blue). **b**, The proportion of highly constrained (Z≥4) windows is positively correlated with the percentage of coding sequences in a window and is substantially higher for the exonic windows. **c**, The proportion of highly constrained (Z≥4) windows increases linearly as more exonic windows are included in the analysis through random sampling . **d**, Constraint Z score percentiles of non-coding versus exonic windows. About 0.5% (100-99.5) and 13.5% (100-86.5) of the non-coding windows exhibit similar constraint to top 10% (90th percentile) and top 50% of exonic regions.

Extended Data Fig. 3: Distributions of GWAS variants in non-coding regions with regard to non-coding constraint and ENCODE cCRE annotations. **a**,**b**, The enrichment of GWAS catalog (**a**) and UKB GWAS (**b**) variants in constrained non-coding regions persists after excluding candidate cis-regulatory elements

(cCREs) annotated by ENCODE. **c**, GWAS variants occur more frequently in cCREs under higher constraint, suggesting the value of non-coding constraint in prioritizing existing functional annotations.

Extended Data Fig. 4: Enrichment of enhancers (**a**) and GWAS variants (**b**) across the spectrum of constraint Z and conservation (PhastCons) score. Non-coding 1kb windows were binned by their constraint Z and PhastCons conservation scores, from the least constrained/conserved (1st decile) to the most constrained/conserved (10th decile), and enrichment within each bin was evaluated by comparing the proportion of 1kb windows that overlap with an enhancer annotation (a) or a GWAS hit (b) to the genome-wide average. Red color indicates enriched and blue color indicates depleted; odds ratio is shown for nominally significant (*p* 0.05) enrichment/depletion. As expected, enrichment increases as both scores increase (upper right), yet each score captures independent signals – for instance, regions with high constraint Z scores present significant enrichment across the deciles of conservation score (rightmost column).

Extended Data Fig. 5: The enrichment of candidate regulatory elements (**a**) and GWAS variants (**b**, GWAS Catalog; **c**, UKB GWAS) in constrained non-coding regions persists when restricting to regions farther away from protein-coding genes ($\pm$10kb).

Extended Data Fig. 6: Correlation of constraint Z and other predictive scores with experimental data on enhancers tested by multiplex assays of variant effect (MAVE). **a**, Predictive scores on MAVE-tested enhancers, the importance of which was ranked by the proportion of mutations MAVE identified as causing significant changes in gene expression. Predictive scores of enhancers were aligned below in the same order and were modified when necessary (multiplied by -1 for CDTS and gwRVIS) such that a higher value represents higher importance for all scores. An increasing trend (dotted line) indicates positive correlation between the predictive score and the experimental measurement from MAVE (Spearman's rank correlations are shown in 1b). Gaps indicate unavailable values; *RET* and *IRF4* were not scored by constraint Z due to feature missingness near the telomere/centromere and *MYC*s1 and *SORT1* were excluded due to quality filtering (yet they are provided in the "Unfiltered Z" panel). **b**, Spearman's rank correlation between MAVE and each predictive score. Constraint Z score exhibits the highest correlation with the experimental measurements, either on all scorable enhancers (All) or the subset of seven enhancers scored by constraint Z (Z scored); the high performance persists even when excluding *UC88* (Exl. UC88) in favor of other scores as it appeared to be an outlier deflecting their ranks. *Of note, JARVIS and LINSIGHT employed enhancers (as well as other functional annotations) in constructing their scores, which may introduce circularity in favor of their performance in scoring enhancers.

Extended Data Fig. 7: The sample size required for well-powered non-coding constraint detection. The proportion of non-coding regions powered to detect constraint (Z>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. For a given level of depletion of variation, the minimum number of expected variants to achieve a Z>4 was determined, and the number of individuals required to achieve a given expected number of variants was extrapolated using a linear model of log(number of expected variants) ~ log(number of individuals), computed from downsamplings of the gnomAD datasets. Lighter color indicates milder deletion of variation (weaker selection), which requires a larger sample size to detect constraint; the grey dashed vertical line indicates current sample size. 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 80% detection power is labeled at corresponding benchmarks.

Extended Data Fig. 8: Analyses of constraint for chromosome X. **a**, Histograms of constraint Z scores for 49,936 1kb windows on chromosome X. Windows encompassing coding sequences (n=2,346; red) overall exhibit a higher constraint Z (stronger negative selection) than windows that are exclusively non-coding (n=47,590; blue). **b**, Distributions of candidate regulatory elements on the non-coding constraint spectrum of chromosome X. Enrichment was evaluated by comparing the proportion of non-coding 1kb windows, binned by constraint Z (<-4, [-4,-3), [-3,-2), [-2,-1), [-1,-0), [0,1), [1,2), [2,3), [3,4), ≥4), that overlap with a given functional annotation to the chromosome-wide average. Error bars indicate 95% confidence intervals of the odds ratios.

References

1 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443, doi:10.1038/s41586-020-2308-7 (2020).

2 Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611-616, doi:10.1038/nature25983 (2018).

3 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, doi:10.1016/j.cell.2019.12.036 (2020).

4 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, doi:10.1038/ng.3903 (2017).

5 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, doi:10.1016/j.ajhg.2018.05.002 (2018).

6 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

7 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).

8 Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends Genet* **31**, 67-76, doi:10.1016/j.tig.2014.12.003 (2015).

9 Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum Mol Genet* **25**, R157-R165, doi:10.1093/hmg/ddw205 (2016).

10 Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110, doi:10.1093/hmg/ddv259 (2015).

11 Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat Rev Genet* **22**, 672-686, doi:10.1038/s41576-021-00376-2 (2021).

12 Seplyarskiy, V. B. *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030-1035, doi:10.1126/science.aba7408 (2021).

13 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, doi:10.1371/journal.pone.0181604 (2017).

14 di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet* **50**, 333-337, doi:10.1038/s41588-018-0062-7 (2018).

15 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

16 Yousefian-Jazi, A., Jung, J., Choi, J. K. & Choi, J. Functional annotation of noncoding causal variants in autoimmune diseases. *Genomics* **112**, 1208-1213, doi:10.1016/j.ygeno.2019.07.006 (2020).

17 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, doi:10.1038/s41467-021-21790-4 (2021).

18 Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49**, 618-624, doi:10.1038/ng.3810 (2017).

19 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

20 Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, doi:10.1126/science.aau1043 (2019).

21 An, J. Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, doi:10.1126/science.aat6576 (2018).

22 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-950, doi:10.1038/ng.3050 (2014).

23 Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451, doi:10.1038/s41586-020-2287-8 (2020).

24 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710, doi:10.1038/s41586-020-2493-4 (2020).
25 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461, doi:10.1038/nature12787 (2014).
26 Jiang, Y. *et al.* SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* **47**, D235-D243, doi:10.1093/nar/gky1025 (2019).
27 Pott, S. & Lieb, J. D. What are super-enhancers? *Nat Genet* **47**, 8-12, doi:10.1038/ng.3167 (2015).
28 Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Res* **30**, 1060-1072, doi:10.1101/gr.254219.119 (2020).
29 Kung, J. T., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193**, 651-669, doi:10.1534/genetics.112.146704 (2013).
30 Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264-268, doi:10.1038/nature09753 (2011).
31 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-1006, doi:10.1093/nar/gkt1229 (2014).
32 Wakil, S. M. *et al.* A genome-wide association study reveals susceptibility loci for myocardial infarction/coronary artery disease in Saudi Arabs. *Atherosclerosis* **245**, 62-70, doi:10.1016/j.atherosclerosis.2015.11.019 (2016).
33 AlRasheed, M. M. *et al.* The role of CDKN2B in cardiovascular risk in ethnic Saudi Arabs: A validation study. *Gene* **673**, 206-210, doi:10.1016/j.gene.2018.06.024 (2018).
34 Silander, K. *et al.* Worldwide patterns of haplotype diversity at 9p21.3, a locus associated with type 2 diabetes and coronary heart disease. *Genome Med* **1**, 51, doi:10.1186/gm51 (2009).
35 Yu, J. H. *et al.* The transcription factors signal transducer and activator of transcription 5A (STAT5A) and STAT5B negatively regulate cell proliferation through the activation of cyclin-dependent kinase inhibitor 2b (Cdkn2b) and Cdkn1a expression. *Hepatology* **52**, 1808-1818, doi:10.1002/hep.23882 (2010).
36 Grange, M. *et al.* Control of CD8 T cell proliferation and terminal differentiation by active STAT5 and CDKN2A/CDKN2B. *Immunology* **145**, 543-557, doi:10.1111/imm.12471 (2015).
37 Almontashiri, N. A. M. The 9p21.3 risk locus for coronary artery disease: A 10-year search for its mechanism. *J Taibah Univ Med Sci* **12**, 199-204, doi:10.1016/j.jtumed.2017.03.001 (2017).
38 Jarinova, O. *et al.* Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. *Arterioscler Thromb Vasc Biol* **29**, 1671-1677, doi:10.1161/ATVBAHA.109.189522 (2009).
39 Motterle, A. *et al.* Functional analyses of coronary artery disease associated variation on chromosome 9p21 in vascular smooth muscle cells. *Hum Mol Genet* **21**, 4021-4029, doi:10.1093/hmg/dds224 (2012).
40 Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409-412, doi:10.1038/nature08801 (2010).
41 Almontashiri, N. A. *et al.* 9p21.3 Coronary Artery Disease Risk Variants Disrupt TEAD Transcription Factor-Dependent Transforming Growth Factor beta Regulation of p16 Expression in Human Aortic Smooth Muscle Cells. *Circulation* **132**, 1969-1978, doi:10.1161/CIRCULATIONAHA.114.015023 (2015).
42 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).
43 Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv*, 2021.2009.2003.21262975, doi:10.1101/2021.09.03.21262975 (2021).
44 Jung, R. G. *et al.* Association between plasminogen activator inhibitor-1 and cardiovascular events: a systematic review and meta-analysis. *Thromb J* **16**, 12, doi:10.1186/s12959-018-0166-4 (2018).
45 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**, doi:10.1161/JAHA.116.004918 (2017).
46 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, doi:10.1161/CIRCGENETICS.114.000554 (2015).
47 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, doi:10.1371/journal.pone.0033511 (2012).

48 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, doi:10.1111/j.1538-7836.2006.01979.x (2006).
49 Lowe, G. D. *et al.* Tissue plasminogen activator antigen and coronary heart disease. Prospective study and meta-analysis. *Eur Heart J* **25**, 252-259, doi:10.1016/j.ehj.2003.11.004 (2004).
50 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, doi:10.1038/ng.415 (2009).
51 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, doi:10.1086/522591 (2007).
52 Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449, doi:10.1126/science.1138659 (2007).
53 Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232-236, doi:10.1038/nature07229 (2008).
54 Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-543, doi:10.1126/science.1155174 (2008).
55 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, doi:10.1016/S0140-6736(14)61705-0 (2015).
56 Rice, A. M. & McLysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat Commun* **8**, 14366, doi:10.1038/ncomms14366 (2017).
57 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-481, doi:10.1146/annurev.genom.9.081307.164217 (2009).
58 Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**, 1063-1071, doi:10.1038/ng.3092 (2014).
59 Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-846, doi:10.1038/ng.909 (2011).
60 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062-D1067, doi:10.1093/nar/gkx1153 (2018).
61 Klopocki, E. *et al.* Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am J Hum Genet* **88**, 70-75, doi:10.1016/j.ajhg.2010.11.006 (2011).
62 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, doi:10.1002/ajmg.a.36811 (2015).
63 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, doi:10.1038/ng.3939 (2017).
64 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).
65 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585, doi:10.1038/ng.2653 (2013).
66 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).
67 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, doi:10.1038/sj.onc.1209382 (2006).
68 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, doi:10.1074/jbc.M114.598383 (2014).
69 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).
70 Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261, doi:10.1016/j.cell.2011.06.013 (2011).
71 Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**, 185-196, doi:10.1016/j.cell.2013.05.056 (2013).

72 Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381-1392, doi:10.1016/j.cell.2012.04.029 (2012).
73 Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007-1012, doi:10.1126/science.1246426 (2014).
74 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).
75 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159, doi:10.1126/science.1260943 (2015).
76 Young, R. S. *et al.* The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* **25**, 1546-1557, doi:10.1101/gr.190546.115 (2015).
77 Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**, 3583, doi:10.1038/s41467-019-11526-w (2019).
78 Hail Team. Hail 0.2.62-84fa81b9ea3d. https://github.com/hail-is/hail/commit/84fa81b9ea3d.
79 Zhu, P. *et al.* Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat Genet* **50**, 12-19, doi:10.1038/s41588-017-0007-6 (2018).
80 Tang, W. W. *et al.* A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell* **161**, 1453-1467, doi:10.1016/j.cell.2015.04.053 (2015).
81 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801, doi:10.1093/nar/gkx1081 (2018).
82 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).
83 Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843-2851, doi:10.1093/bioinformatics/btu356 (2014).
84 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, doi:10.1038/s41588-018-0071-6 (2018).
85 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007, doi:10.1093/bioinformatics/btt730 (2014).
86 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
87 Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**, 2235-2242, doi:10.1056/NEJMsr1406261 (2015).
88 Blake, J. A. *et al.* The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res* **39**, D842-848, doi:10.1093/nar/gkq1008 (2011).
89 McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588-604, doi:10.1086/514346 (2007).
90 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).
91 Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol* **18**, 193, doi:10.1186/s13059-017-1308-x (2017).

Genome Aggregation Database Consortium

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

[1]University of Miami Miller School of Medicine, Gastroenterology, Miami, USA
[2]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico
[3]Peninsula College of Medicine and Dentistry, Exeter, UK
[4]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA
[5]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

[8]Department of Cardiology University Hospital, Parma, Italy
[9]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK
[10]Department of Biology Faculty of Natural Sciences, University of Haifa, Haifa, Israel
[11]Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA
[12]Data Science Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[13]Department of Quantitative Health Sciences, Lerner Research Institute Cleveland Clinic, Cleveland, OH, USA
[14]Sorbonne Université, APHP, Gastroenterology Department Saint Antoine Hospital, Paris, France
[15]NHLBI and Boston University's Framingham Heart Study, Framingham, MA, USA
[16]Department of Medicine, Boston University School of Medicine, Boston, MA, USA
[17]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
[18]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
[19]National Human Genome Research Institute, National Institutes of Health Bethesda, MD, USA
[20]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[21]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA
[22]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
[23]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
[24]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
[25]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
[26]John Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[27]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[28]Harvard School of Public Health, Boston, MA, USA
[29]Central American Population Center, San Pedro, Costa Rica
[30]Department of Epidemiology and Biostatistics, Imperial College London, London, UK
[31]Department of Cardiology, Ealing Hospital, NHS Trust, Southall, UK
[32]Imperial College, Healthcare NHS Trust Imperial College London, London, UK
[33]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China
[34]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[35]Department of Medicine, Harvard Medical School, Boston, MA, USA
[36]Northwestern University, Evanston, IL, USA
[37]University of Cambridge, Cambridge, England
[38]Departments of Cardiovascular, Medicine Cellular and Molecular Medicine Molecular Cardiology, Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA
[39]Department of Pediatrics, Columbia University Irving Medical Center, New York, NY, USA
[40]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA
[41]Department of Medicine, Columbia University Medical Center, New York, NY, USA
[42]McLean Hospital, Belmont, MA, USA
[43]Division of Medical Sciences, Harvard Medical School, Boston, MA, USA
[44]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[45]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA
[46]Department of Epidemiology Colorado School of Public Health Aurora, CO, USA
[47]Institute for Molecular Medicine Finland, (FIMM) Helsinki, Finland
[48]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA
[49]Vanderbilt University Medical Center, Nashville, TN, USA
[50]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
[51]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[52]National Heart Lung and Blood Institute's Framingham Heart Study, Framingham, MA, USA
[53]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA
[54]Cardiovascular Epidemiology and Genetics Hospital del Mar Medical Research Institute, (IMIM) Barcelona Catalonia, Spain

[55]CIBER CV Barcelona, Catalonia, Spain
[56]Departament of Medicine, Medical School University of Vic-Central, University of Catalonia, Vic Catalonia, Spain
[57]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
[58]German Research Centre for Cardiovascular Research, Hamburg/Lübeck/Kiel, Lübeck, Germany
[59]University Heart Center Lübeck, Lübeck, Germany
[60]Estonian Genome Center, Institute of Genomics University of Tartu, Tartu, Estonia
[61]Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia
[62]Faculty of Medicine, UNSW Sydney, Kensington, NSW, Australia
[63]Cardiology Department, St Vincent's Hospital, Darlinghurst, NSW, Australia
[64]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[65]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[66]Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[67]Institute of Clinical Molecular Biology, (IKMB) Christian-Albrechts-University of Kiel, Kiel, Germany
[68]Helsinki University and Helsinki University Hospital Clinic of Gastroenterology, Helsinki, Finland
[69]Bioinformatics Program MGH Cancer Center and Department of Pathology, Boston, MA, USA
[70]Cancer Genome Computational Analysis, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[71]Department of Psychiatry and Behavioral Sciences, Boston Children's Hospitaland Harvard Medical School, Boston, MA, USA
[72]Harvard Medical School Teaching Hospital, Boston, MA, USA
[73]Department of Endocrinology and Metabolism, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Israel
[74]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA
[75]Institute for Genomic Medicine, Columbia University Medical Center Hammer Health Sciences, New York, NY, USA
[76]Department of Genetics & Development Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA
[77]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Mexico
[78]Lund University Sweden, Sweden
[79]Institute for Molecular Medicine Finland, (FIMM) HiLIFE University of Helsinki, Helsinki, Finland
[80]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA
[81]Lund University Diabetes Centre, Malmö, Skåne County, Sweden
[82]Washington School of Medicine, St Louis, MI, USA
[83]Human Genetics Center University of Texas Health Science Center at Houston, Houston, TX, USA
[84]Department of Neurology Columbia University, New York City, NY, USA
[85]Institute of Genomic Medicine, Columbia University, New York City, NY, USA
[86]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
[87]Department of Psychiatry, Helsinki University Central Hospital Lapinlahdentie, Helsinki, Finland
[88]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[89]Icahn School of Medicine at Mount Sinai, New York, NY, USA
[90]Bonei Olam, Center for Rare Jewish Genetic Diseases, Brooklyn, NY, USA
[91]Department of Neurology, Helsinki University, Central Hospital, Helsinki, Finland
[92]Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[93]Charles Bronfman Institute for Personalized Medicine, New York, NY, USA
[94]Division of Genome Science, Department of Precision Medicine, National Institute of Health, Republic of Korea
[95]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, Wales
[96]Imperial College, Healthcare NHS Trust, London, UK
[97]National Heart and Lung Institute Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK
[98]Department of Health THL-National Institute for Health and Welfare, Helsinki, Finland
[99]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, Center for Outcomes Research and Evaluation Yale-New Haven Hospital, New Haven, CT, USA

[100]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA
[101]Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[102]The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland
[103]Kuopio University Hospital, Kuopio, Finland
[104]Department of Genetics, Yale School of Medicine, New Haven, CT, USA
[105]Department of Clinical Chemistry Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere Faculty of Medicine and Health Technology, Tampere University, Finland
[106]The Mindich Child Health and Development, Institute Icahn School of Medicine at Mount Sinai, New York, NY, USA
[107]National Autonomous University of Mexico, Mexico City, Mexico
[108]Salvador Zubirán National Institute of Health Sciences and Nutrition, Mexico City, Mexico
[109]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
[110]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China
[111]Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, Australia
[112]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia
[113]University of California San Francisco Parnassus Campus, San Francisco, CA, USA
[114]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute, (IMIM) Barcelona, Catalonia, Spain
[115]Department of Genetics, Harvard Medical School, Boston, MA, USA
[116]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital Old Road Headington, Oxford, OX, LJ, UK
[117]Welcome Centre for Human Genetics, University of Oxford, Oxford, OX, BN, UK
[118]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals, NHS Foundation Trust, John Radcliffe Hospital, Oxford, OX, DU, UK
[119]John P. Hussman Institute for Human Genomics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA
[120]The Dr. John T. Macdonald Foundation Department of Human Genetics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA
[121]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute Cedars-Sinai Medical Center, Los Angeles, CA, USA
[122]Atherogenomics Laboratory University of Ottawa, Heart Institute, Ottawa, Canada
[123]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA
[124]Department of Clinical Sciences University, Hospital Malmo Clinical Research Center, Lund University, Malmö, Sweden
[125]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
[126]University of Arizona Health Science, Tuscon, AZ, USA
[127]University of Maryland School of Medicine, Baltimore, MD, USA
[128]Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA
[129]International Centre for Diarrhoeal Disease Research, Bangladesh
[130]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[131]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[132]Lund University, Dept. Clinical Sciences, Skåne University Hospital, Malmö, Sweden
[133]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan
[134]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Suita, Japan
[135]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan
[136]Instituto Nacional de Medicina Genómica, (INMEGEN) Mexico City, Mexico
[137]Medical Research Institute, Ninewells Hospital and Medical School University of Dundee, Dundee, UK
[138]Wake Forest School of Medicine, Winston-Salem, NC, USA
[139]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea

[140]Department of Psychiatry Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA
[141]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[142]Children's Hospital of Philadelphia, Philadelphia, PA, USA
[143]Division of Genetics and Epidemiology, Institute of Cancer Research, London, SM, NG
[144]University of Washington, Seattle, WA, USA
[145]Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[146]Medical Research Center, Oulu University Hospital, Oulu Finland and Research Unit of Clinical Neuroscience Neurology University of Oulu, Oulu, Finland
[147]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[148]Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA
[149]Research Center Montreal Heart Institute, Montreal, Quebec, Canada
[150]Department of Medicine, Faculty of Medicine Université de Montréal, Québec, Canada
[151]Department of Public Health Faculty of Medicine, University of Helsinki, Helsinki, Finland
[152]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[153]Department of Biomedical Informatics Vanderbilt, University Medical Center, Nashville, TN, USA
[154]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[155]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
[156]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[157]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[158]Center for Non-Communicable Diseases, Karachi, Pakistan
[159]National Institute for Health and Welfare, Helsinki, Finland
[160]Deutsches Herzzentrum, München, Germany
[161]Technische Universität München, Germany
[162]Institute of Genetic Epidemiology, Department of Genetics and Pharmacology, Medical University of Innsbruck, 6020 Innsbruck, Austria
[163]Duke Molecular Physiology Institute, Durham, NC
[164]Division of Cardiovascular Medicine, Nashville VA Medical Center, Vanderbilt University School of Medicine, Nashville, TN, USA
[165]Division of Endocrinology, National University Hospital, Singapore
[166]NUS Saw Swee Hock School of Public Health, Singapore
[167]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA
[168]Harvard Medical School, Boston, MA, USA
[169]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[170]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[171]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[172]The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University and the Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden
[173]Department of Cardiology, Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Clinical Sciences, Lund University and Skåne University Hospital, Lund, Sweden
[174]Institute of Clinical Medicine Neurology, University of Eastern Finad, Kuopio, Finland
[175]Sorbonne Université, INSERM, Centre de Recherche Saint-Antoine, CRSA, AP-HP, Saint Antoine Hospital, Gastroenterology department, F-75012 Paris, France
[176]INRA, UMR1319 Micalis & AgroParisTech, Jouy en Josas, France
[177]Paris Center for Microbiome Medicine, (PaCeMM) FHU, Paris, France
[178]Department of Twin Research and Genetic Epidemiology King's College London, London, UK
[179]The McDonnell Genome Institute at Washington University, Seattle, WA, USA
[180]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
[181]Saw Swee Hock School of Public Health National University of Singapore, National University Health System, Singapore
[182]Department of Medicine, Yong Loo Lin School of Medicine National University of Singapore, Singapore

[183]Duke-NUS Graduate Medical School, Singapore
[184]Life Sciences Institute, National University of Singapore, Singapore
[185]Department of Statistics and Applied Probability, National University of Singapore, Singapore
[186]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA
[187]Institute of Genomic Medicine, University of California San Diego, San Diego, CA, USA
[188]Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland
[189]Institute of Genetics, Folkhalsan Research Center, Helsinki, Finland
[190]Juliet Keidan Institute of Pediatric Gastroenterology Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel
[191]Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico
[192]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico
[193]Department of Public Health Faculty of Medicine University of Helsinki, Helsinki, Finland
[194]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA
[195]National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College, London, UK
[196]Cardiovascular Research Centre Royal Brompton & Harefield Hospitals, London, UK
[197]Radcliffe Department of Medicine, University of Oxford, Oxford, UK
[198]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, Netherlands
[199]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland
[200]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK
[201]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK
[202]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA
[203]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[204]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA

Authors received funding as follows:
Matthew J. Bown: British Heart Foundation awards CS/14/2/30841 and RG/18/10/33842
Josée Dupuis: National Heart Lung and Blood Institute's Framingham Heart Study Contract (HHSNI); National Institute for Diabetes and Digestive and Kidney Diseases (NIDDK) R DK
Martti Färkkilä: State funding for university level health research
Laura D. Gauthier: Intel, Illumina
Stephen J. Glatt: U.S. NIMH Grant R MH
Leif Groop: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312063 and 336822), Sigrid Jusélius Foundation; IMI 2 (grant No 115974 and 15881)
Mikko Hiltunen: Academy of Finland (grant 338182) Sigrid Jusélius Foundation the Strategic Neuroscience Funding of the University of Eastern Finland
Chaim Jalas: Bonei Olam
Jaakko Kaprio: Academy of Finland (grants 312073 and 336823)
Jacob McCauley: National Institute of Diabetes and Digestive and Kidney Disease Grant R01DK104844
Yukinori Okada: JSPS KAKENHI (19H01021, 20K21834), AMED (JP21km0405211, JP21ek0109413, JP21gm4010006, JP21km0405217, JP21ek0410075), JST Moonshot R&D (JPMJMS2021)
Michael J. Owen: Medical Research Council UK: Centre Grant No. MR/L010305/1, Program Grant No. G0800509
Aarno Palotie: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant numbers 312074 and 336824) and Sigrid Jusélius Foundation
John D. Rioux: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK; DK062432), from the Canadian Institutes of Health (CIHR GPG 102170), from Genome Canada/Génome Québec (GPH-129341), and a Canada Research Chair (#230625)
Samuli Ripatti: the Academy of Finland Center of Excellence for Complex Disease Genetics (grant number) Sigrid Jusélius Foundation
Jerome I. Rotter: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support

including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center

Edwin K. Silverman: NIH Grants U01 HL089856 and U01 HL089897

J. Gustav Smith: The Swedish Heart-Lung Foundation (2019-0526), the Swedish Research Council (2017-02554), the European Research Council (ERC-STG-2015-679242), Skåne University Hospital, governmental funding of clinical research within the Swedish National Health Service, a generous donation from the Knut and Alice Wallenberg foundation to the Wallenberg Center for Molecular Medicine in Lund, and funding from the Swedish Research Council (Linnaeus grant Dnr 349-2006-237, Strategic Research Area Exodiab Dnr 2009-1039) and Swedish Foundation for Strategic Research (Dnr IRC15-0067) to the Lund University Diabetes Center

Kent D. Taylor: Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for MESA and TOPMed. JSK was supported by the Pulmonary Fibrosis Foundation Scholars Award and grant K23-HL-150301 from the NHLBI. MRA was supported by grant K23-HL-150280, AJP was supported by grant K23-HL-140199, and AM was supported by R01-HL131565 from the NHLBI. EJB was supported by grant K23-AR-075112 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center

Tiinamaija Tuomi: The Academy of Finland and University of Helsinki: Center of Excellence for Complex Disease Genetics (grant number 312072 and 336826), Folkhalsan Research Foundation, Helsinki University Hospital, Ollqvist Foundation, Liv och Halsa foundation; NovoNordisk Foundation

Teresa Tusie-Luna: CONACyT Project 312688

James S. Ware: Wellcome Trust [107469/Z/15/Z], Medical Research Council (UK), NIHR Imperial College Biomedical Research Centre

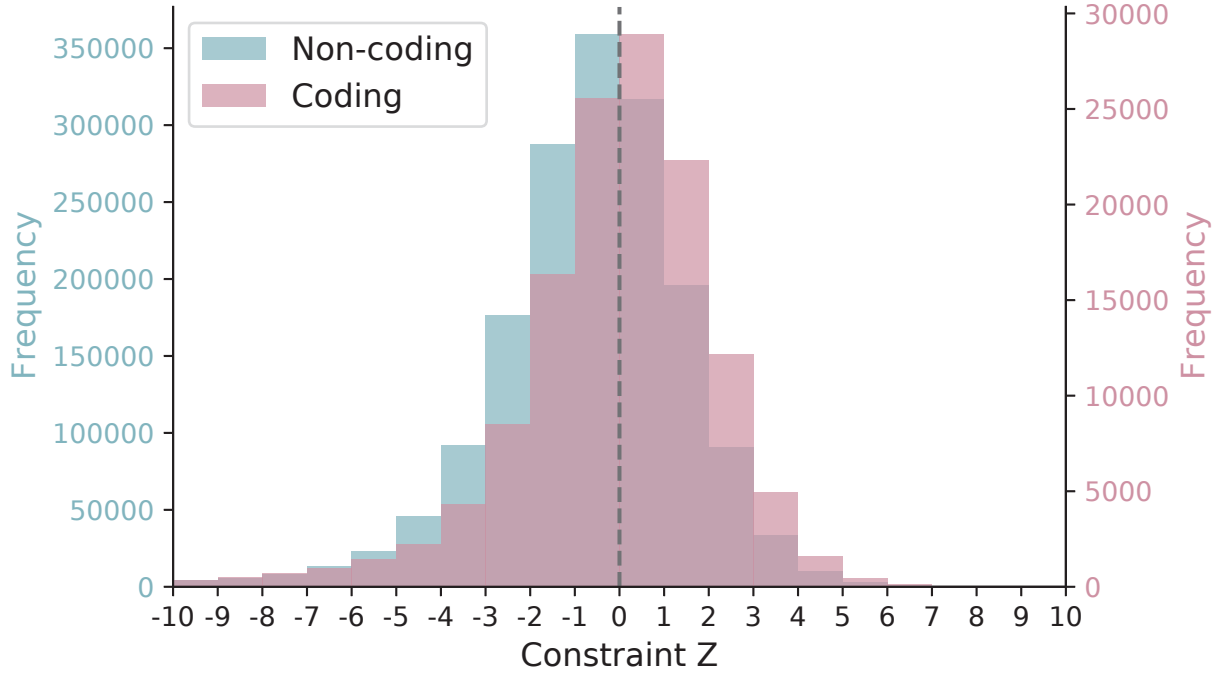
Rinse K. Weersma: The Lifelines Biobank initiative has been made possible by subsidy from the Dutch Ministry of

Health Welfare and Sport the Dutch Ministry of Economic Affairs the University Medical Centre Groningen (UMCG the Netherlands) the University of Groningen and the Northern Provinces of the Netherlands

No conflicts of interest to declare

Figure 1

a


Non-coding
Coding
Frequency
Frequency
Constraint Z
350000
300000
250000
200000
150000
100000
50000
0
30000
25000
20000
15000
10000
5000
0
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

b

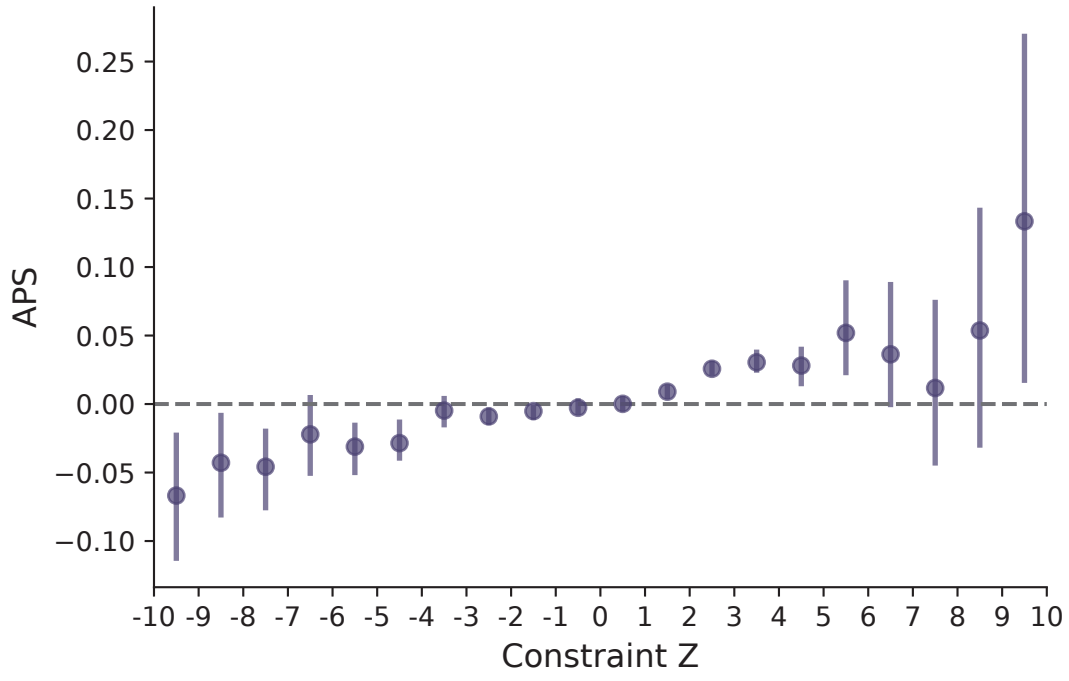

APS
Constraint Z
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

Figure 2

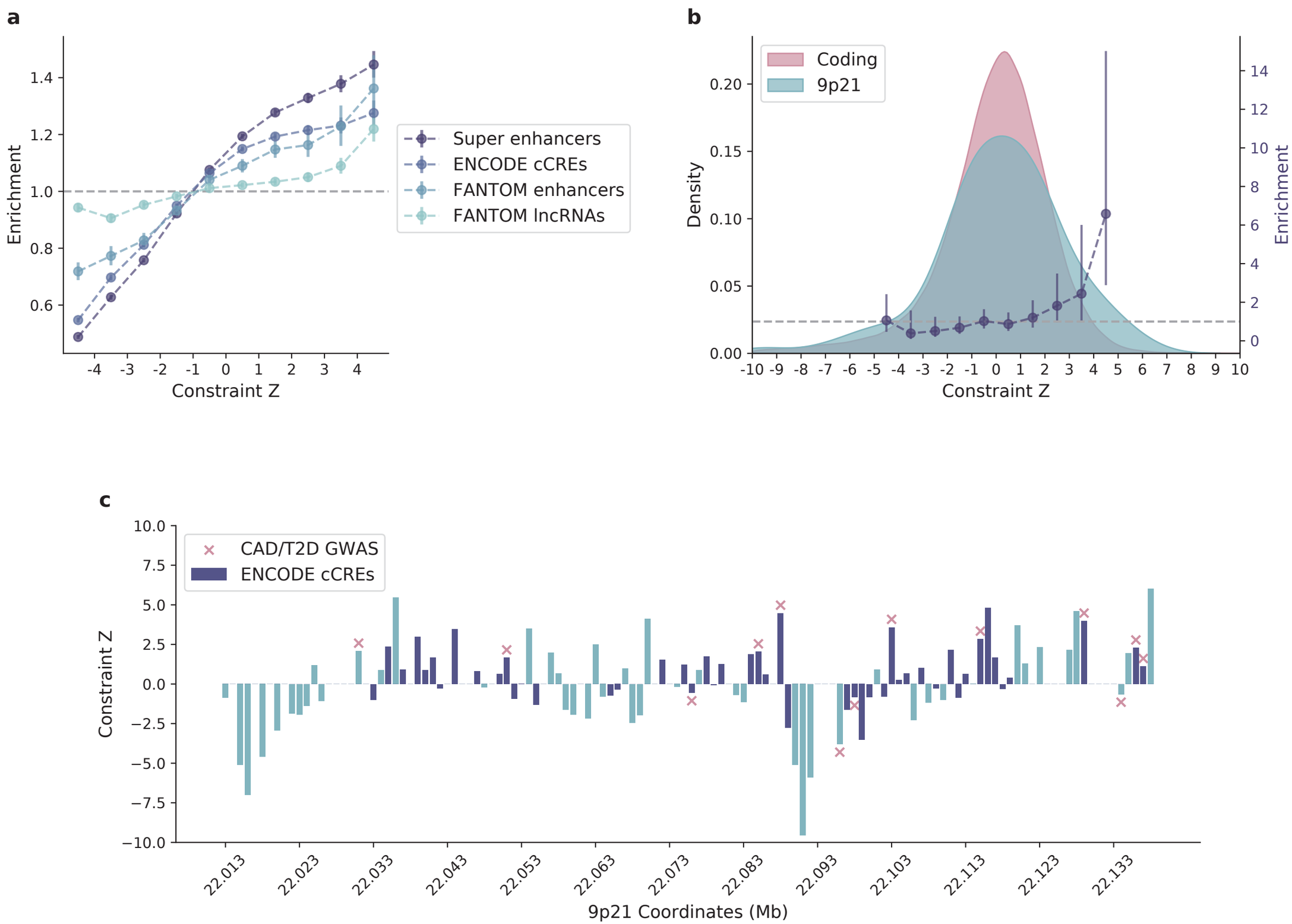
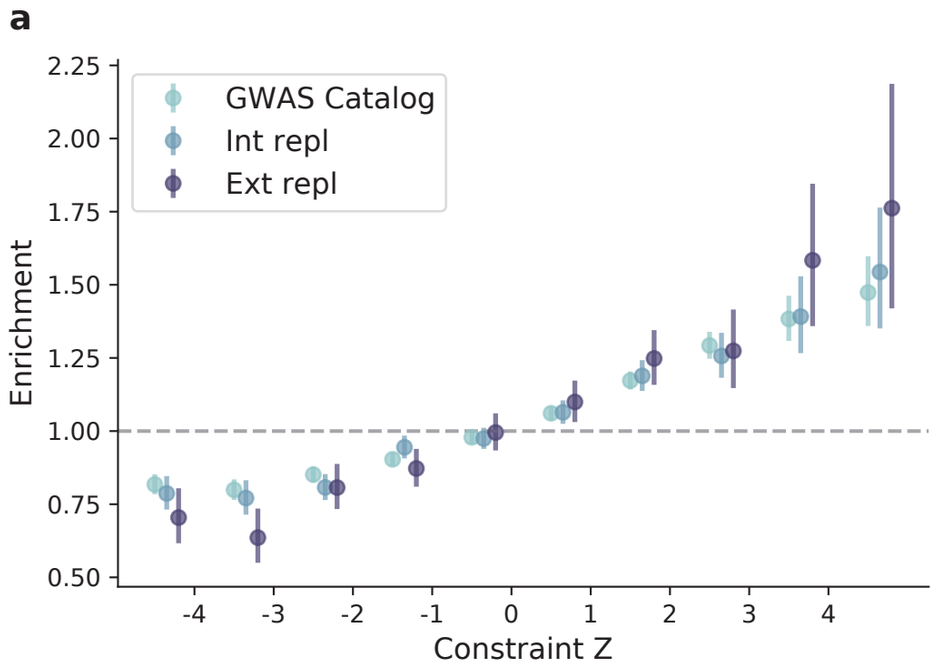

a
Enrichment
Constraint Z
Super enhancers
ENCODE cCREs
FANTOM enhancers
FANTOM lncRNAs
b
Coding
9p21
Density
Enrichment
Constraint Z
c
CAD/T2D GWAS
ENCODE cCREs
Constraint Z
9p21 Coordinates (Mb)

Figure 3


a
GWAS Catalog
Int repl
Ext repl
Enrichment
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
-4
-3
-2
-1
0
1
2
3
4
Constraint Z

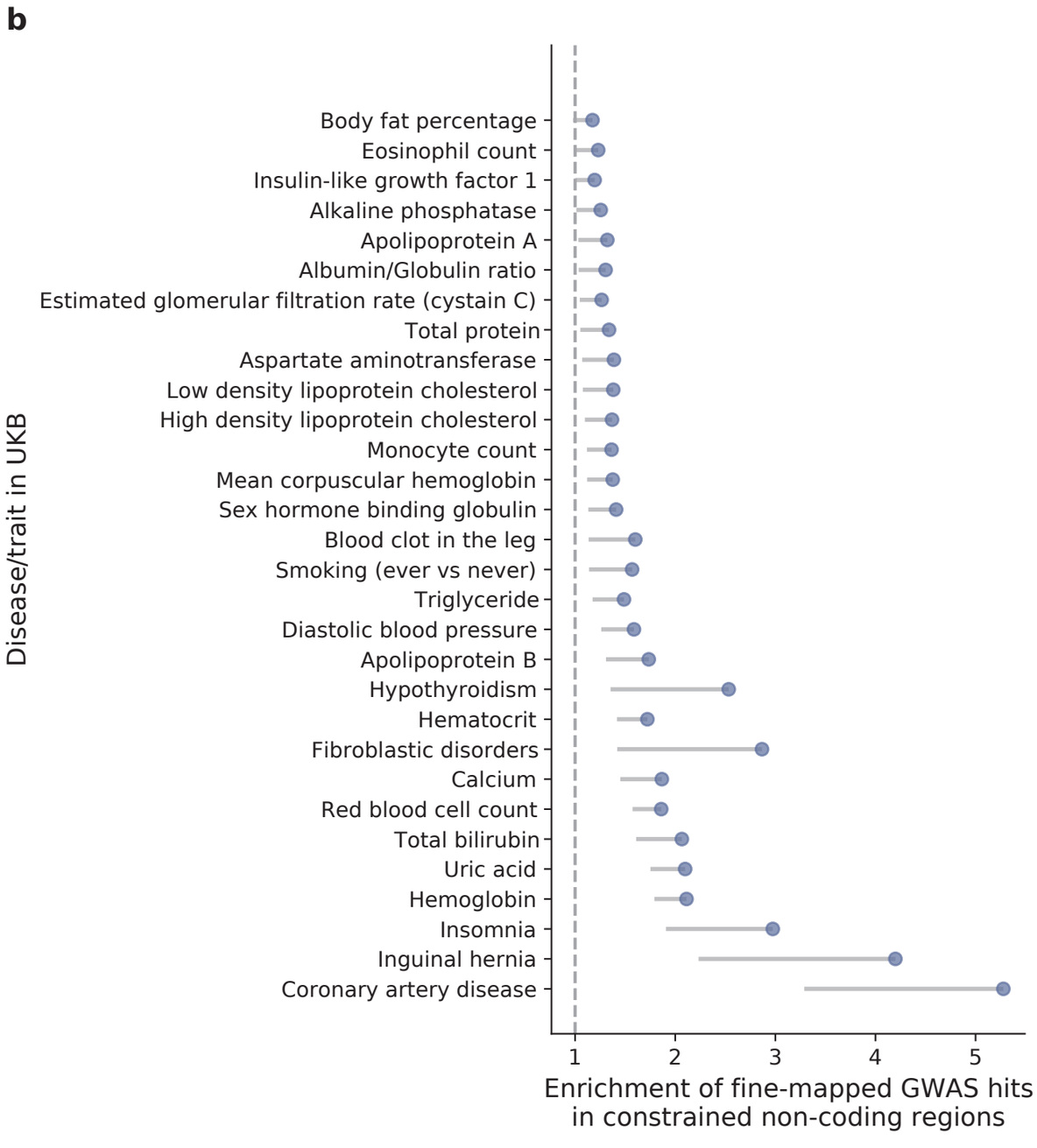

b
Disease/trait in UKB
Body fat percentage
Eosinophil count
Insulin-like growth factor 1
Alkaline phosphatase
Apolipoprotein A
Albumin/Globulin ratio
Estimated glomerular filtration rate (cystain C)
Total protein
Aspartate aminotransferase
Low density lipoprotein cholesterol
High density lipoprotein cholesterol
Monocyte count
Mean corpuscular hemoglobin
Sex hormone binding globulin
Blood clot in the leg
Smoking (ever vs never)
Triglyceride
Diastolic blood pressure
Apolipoprotein B
Hypothyroidism
Hematocrit
Fibroblastic disorders
Calcium
Red blood cell count
Total bilirubin
Uric acid
Hemoglobin
Insomnia
Inguinal hernia
Coronary artery disease
1
2
3
4
5
Enrichment of fine-mapped GWAS hits
in constrained non-coding regions

Figure 4

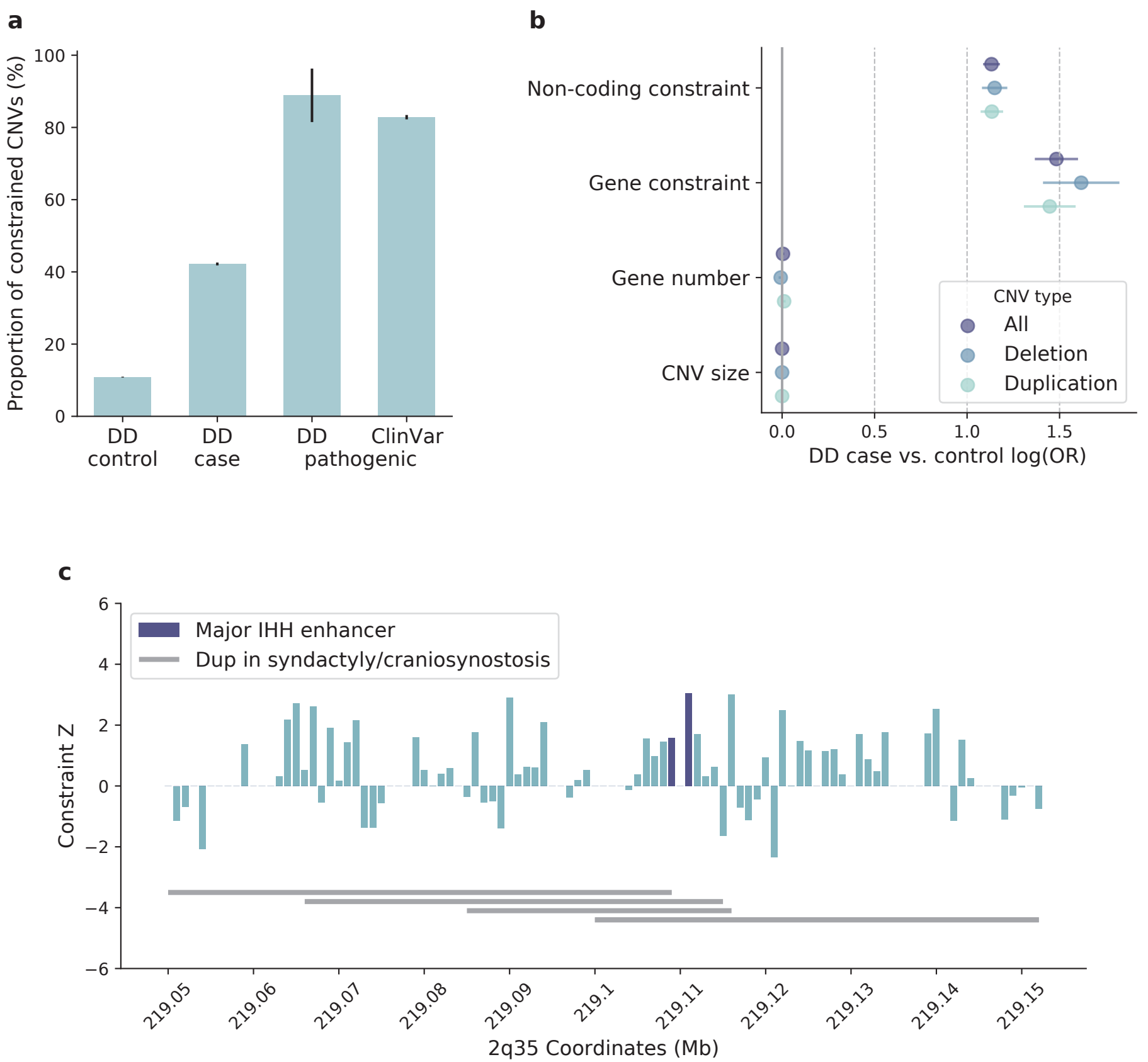
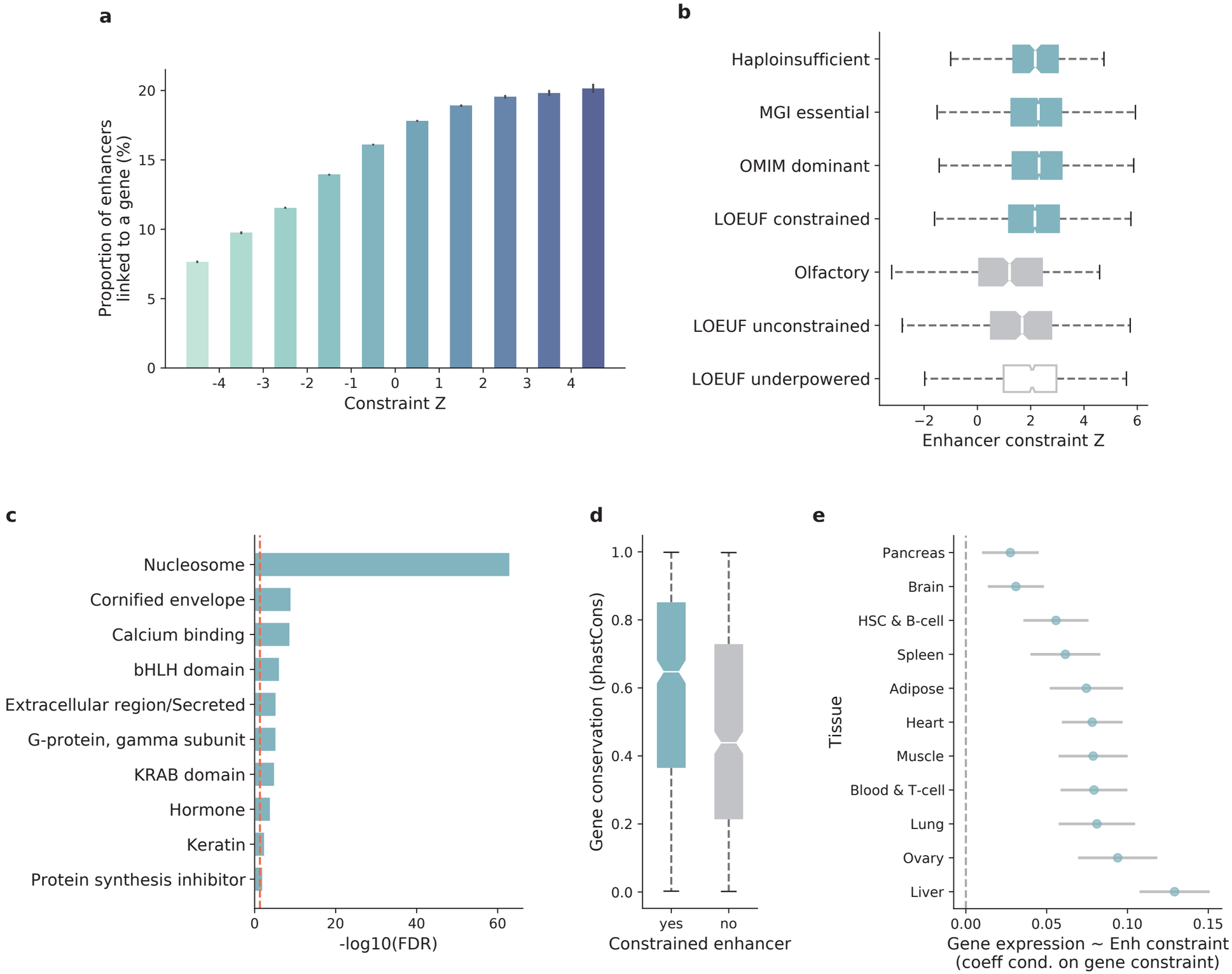
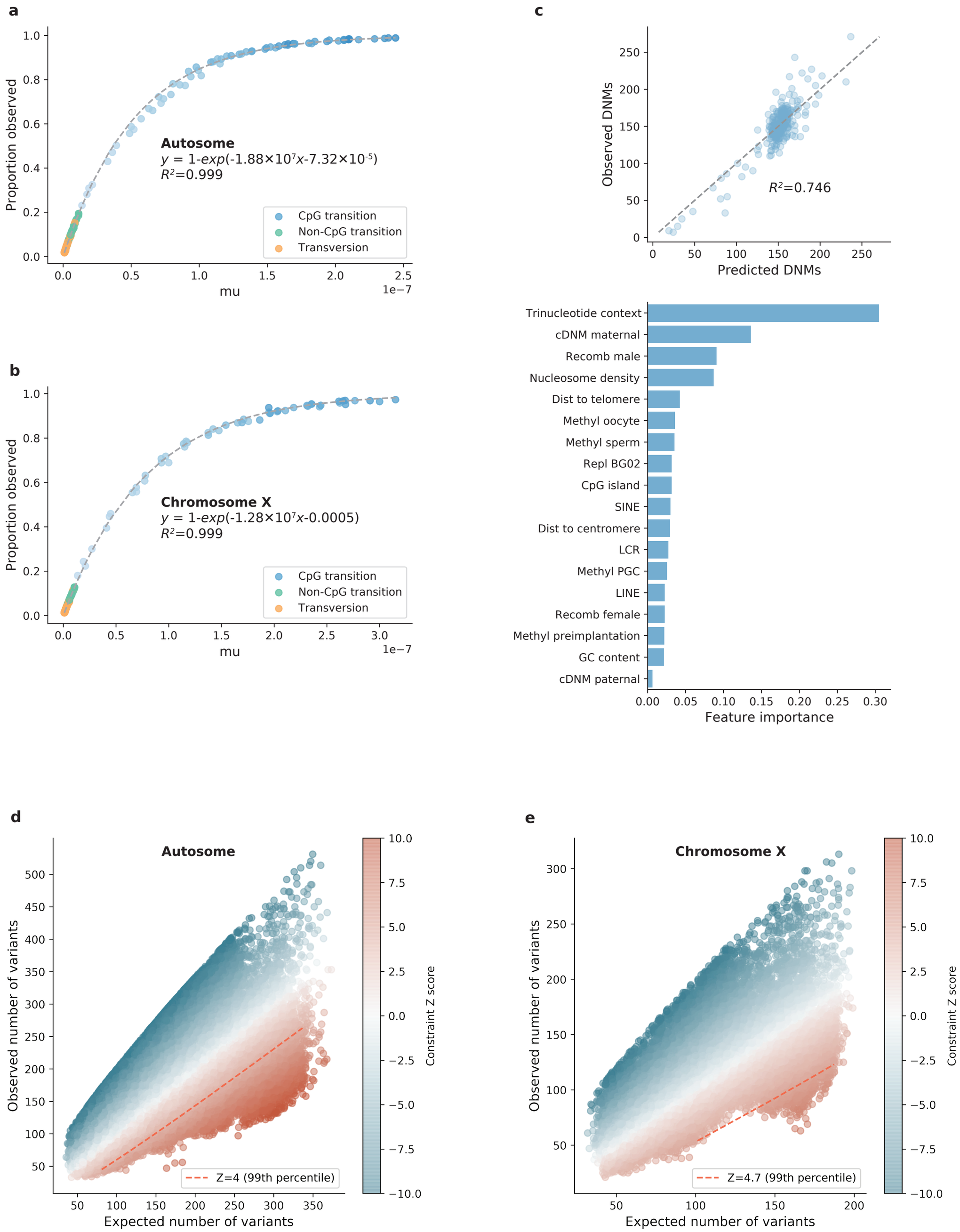

a
Proportion of constrained CNVs (%)
100
80
60
40
20
0
DD control
DD case
DD pathogenic
ClinVar pathogenic
b
Non-coding constraint
Gene constraint
Gene number
CNV size
CNV type
All
Deletion
Duplication
0.0
0.5
1.0
1.5
DD case vs. control log(OR)
c
Major IHH enhancer
Dup in syndactyly/craniosynostosis
Constraint Z
6
4
2
0
−2
−4
−6
219.05
219.06
219.07
219.08
219.09
219.1
219.11
219.12
219.13
219.14
219.15
2q35 Coordinates (Mb)

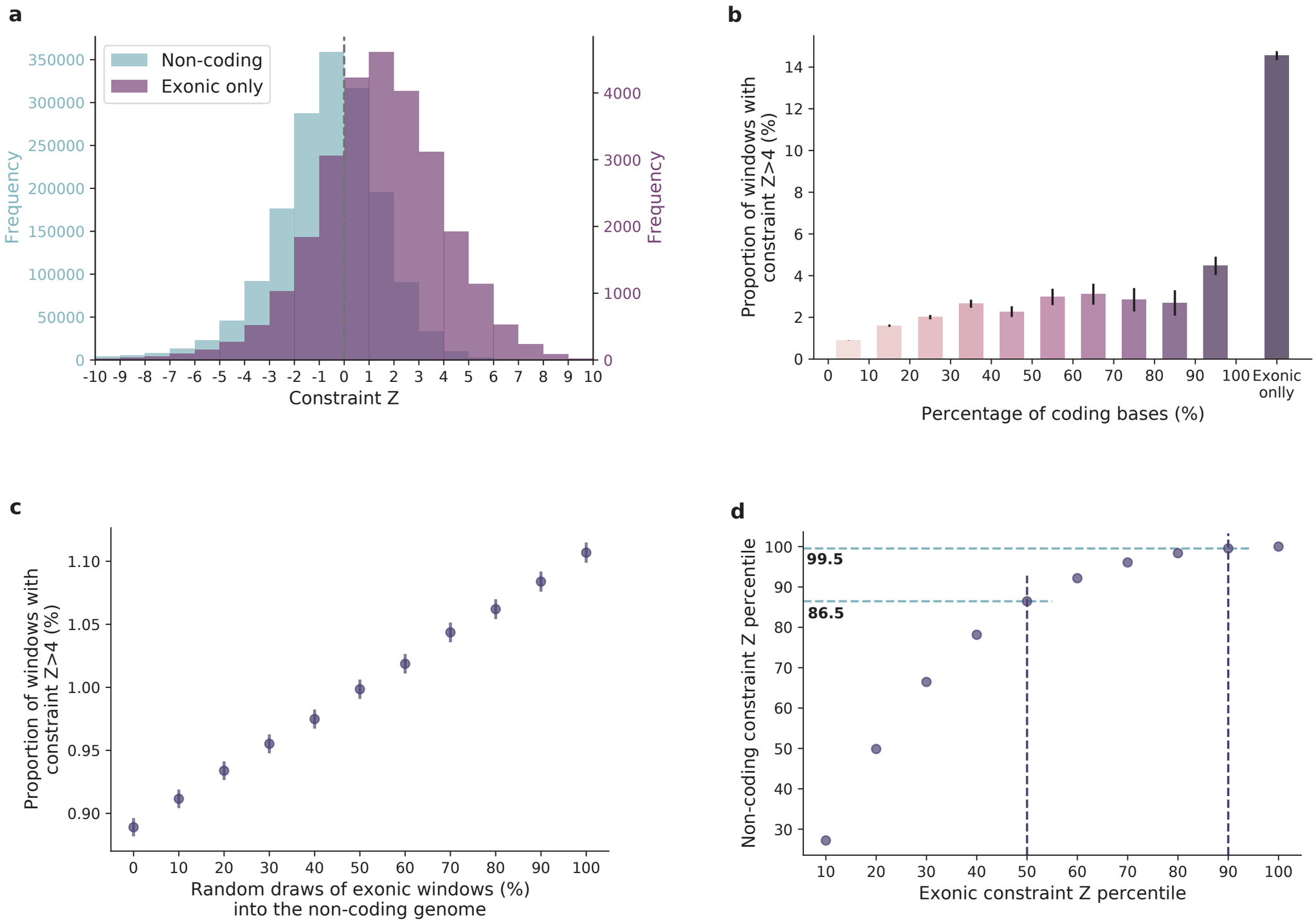
Figure 5


a
Proportion of enhancers linked to a gene (%)
0
5
10
15
20
-4
-3
-2
-1
0
1
2
3
4
Constraint Z
b
Haploinsufficient
MGI essential
OMIM dominant
LOEUF constrained
Olfactory
LOEUF unconstrained
LOEUF underpowered
−2
0
2
4
6
Enhancer constraint Z
c
Nucleosome
Cornified envelope
Calcium binding
bHLH domain
Extracellular region/Secreted
G-protein, gamma subunit
KRAB domain
Hormone
Keratin
Protein synthesis inhibitor
0
20
40
60
-log10(FDR)
d
Gene conservation (phastCons)
0.0
0.2
0.4
0.6
0.8
1.0
yes
no
Constrained enhancer
e
Tissue
Pancreas
Brain
HSC & B-cell
Spleen
Adipose
Heart
Muscle
Blood & T-cell
Lung
Ovary
Liver
0.00
0.05
0.10
0.15
Gene expression ~ Enh constraint
(coeff cond. on gene constraint)

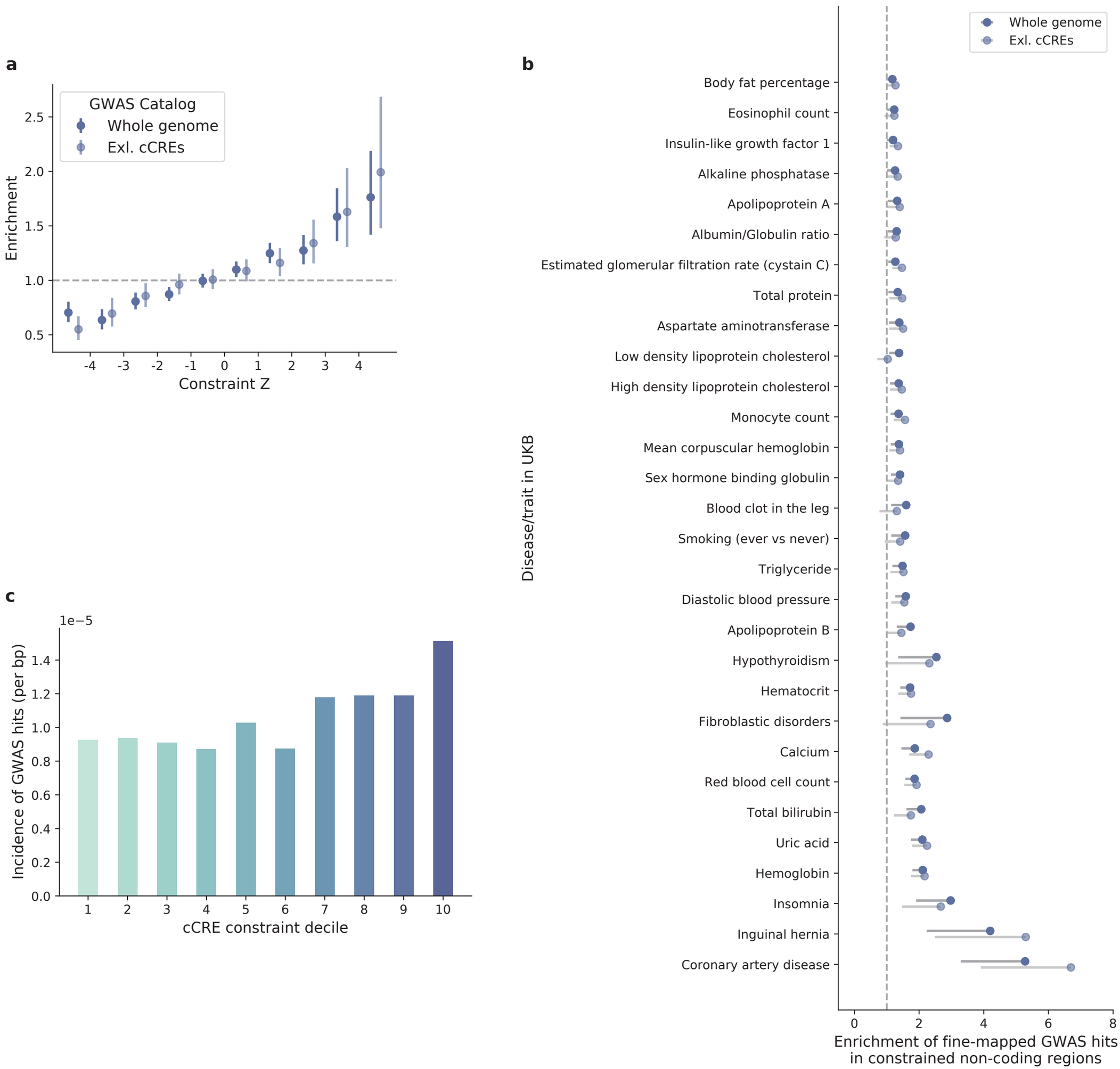
Extended Data Figure 1



Extended Data Figure 2

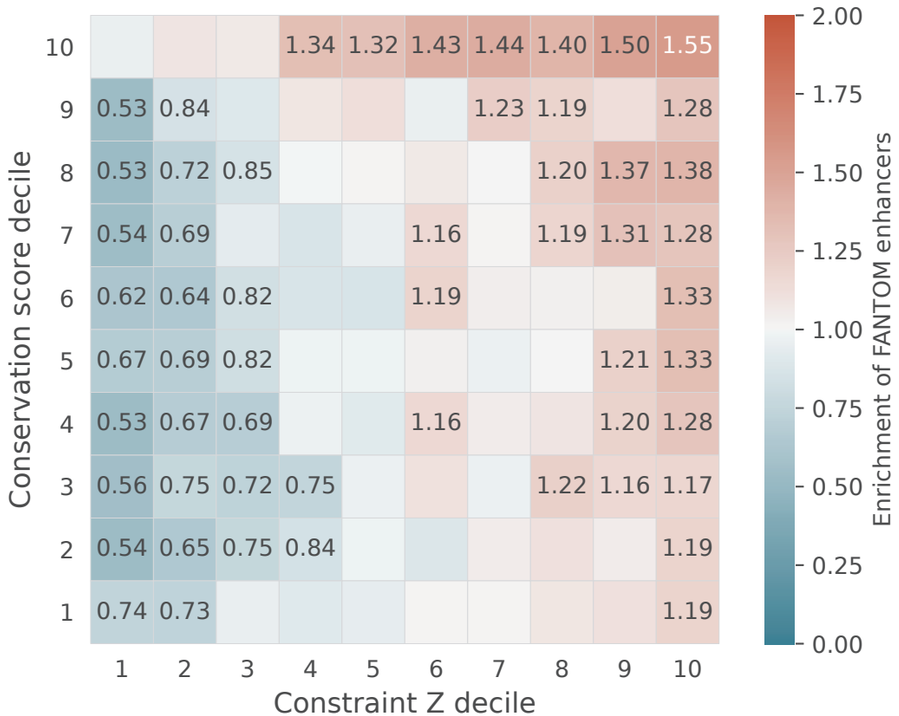


Extended Data Figure 3

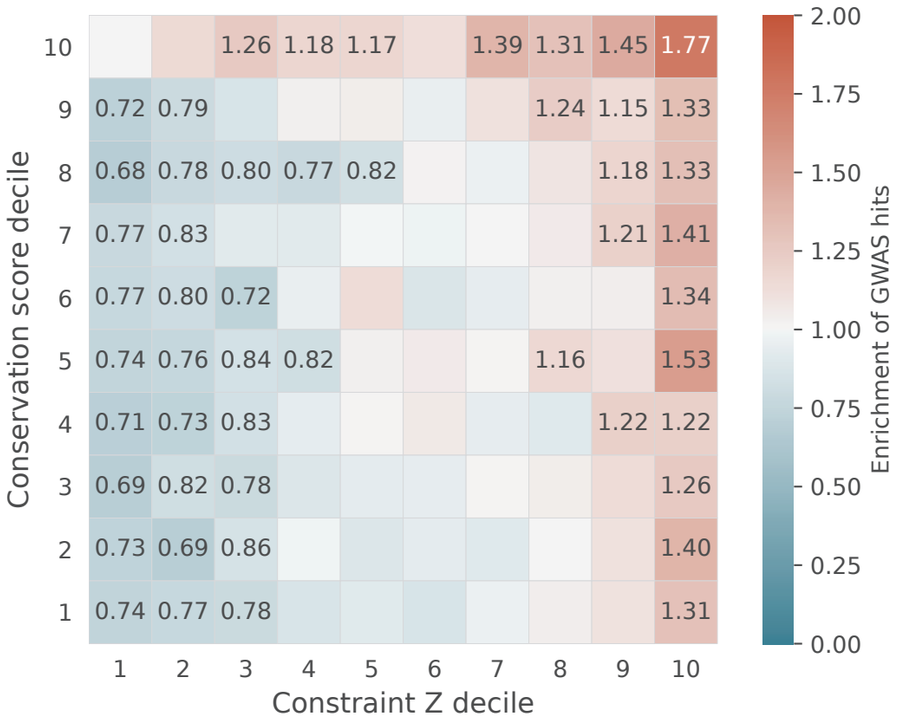


Extended Data Figure 4

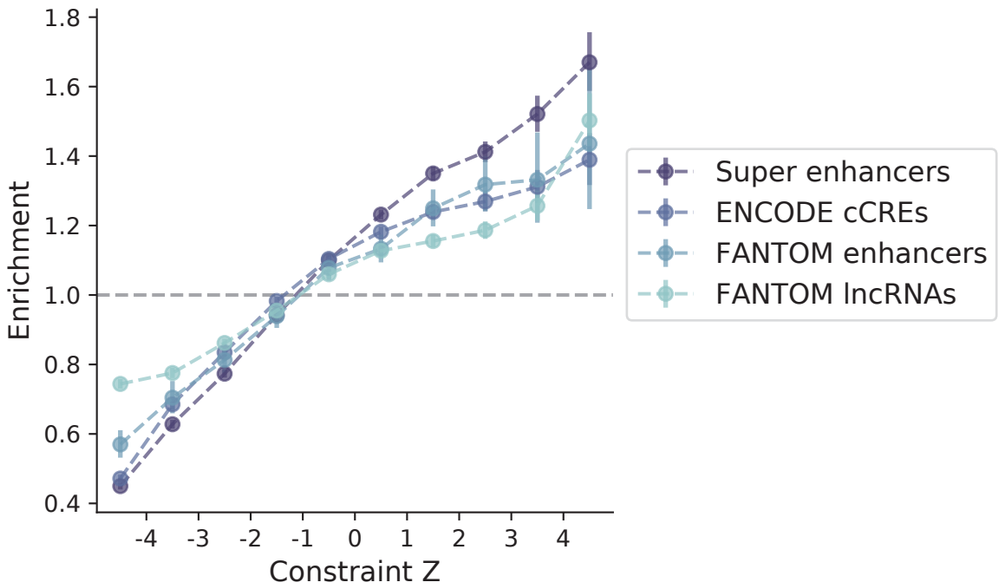
a

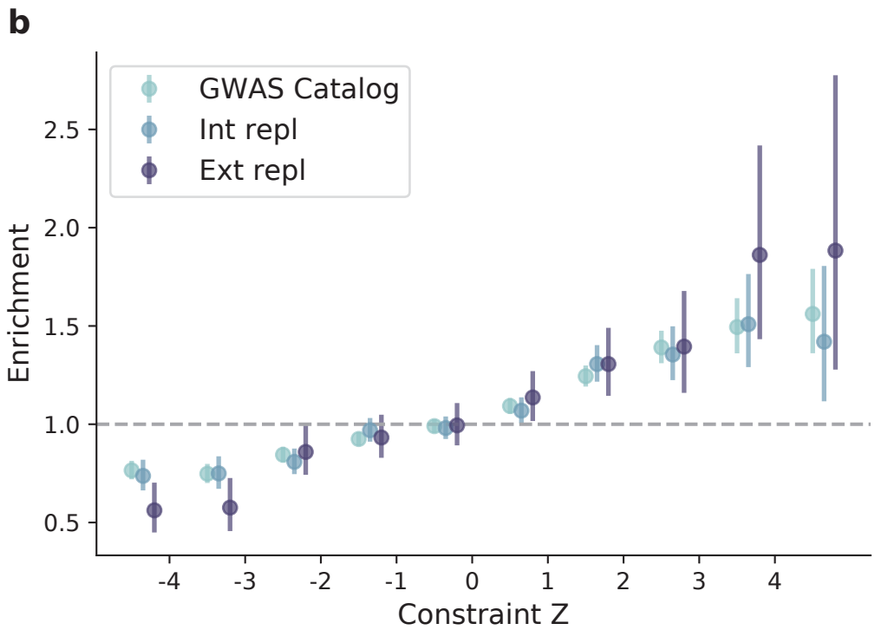


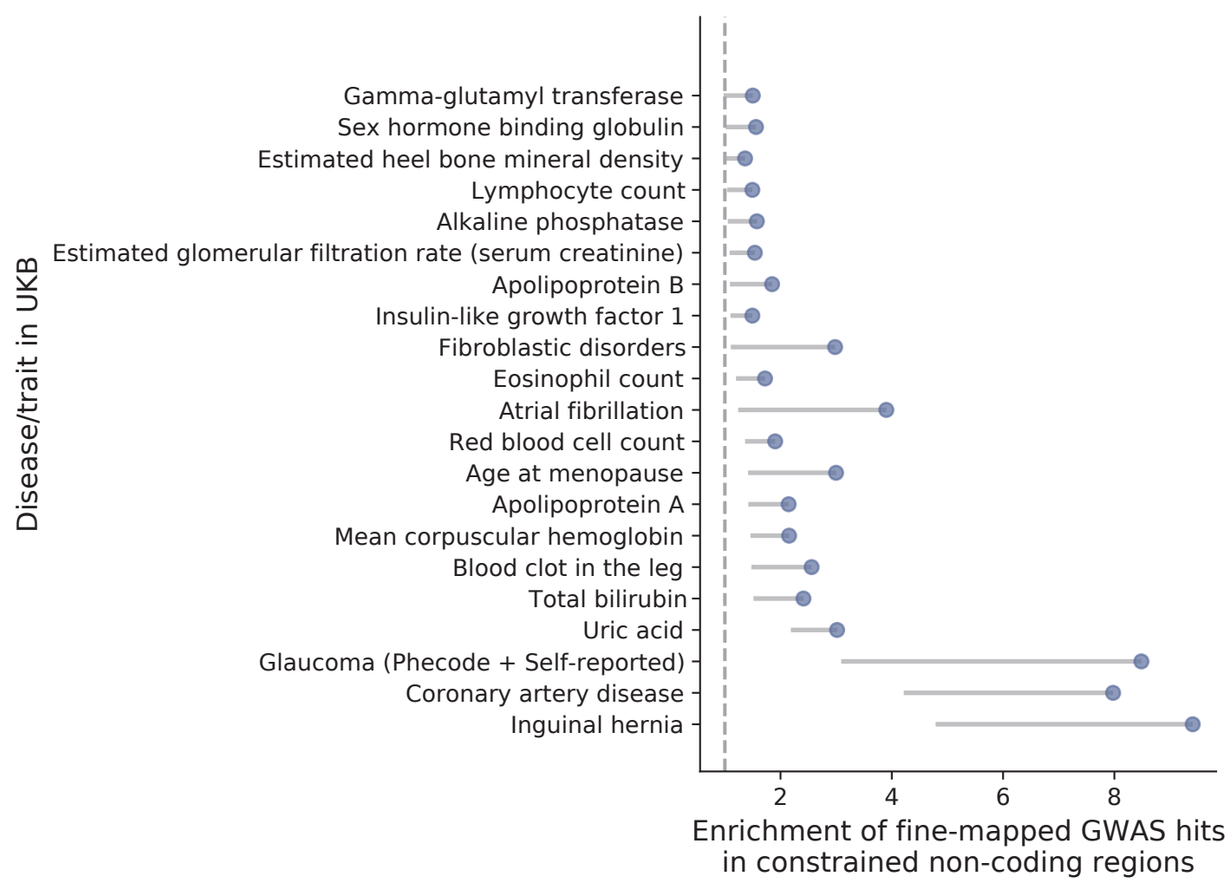
b



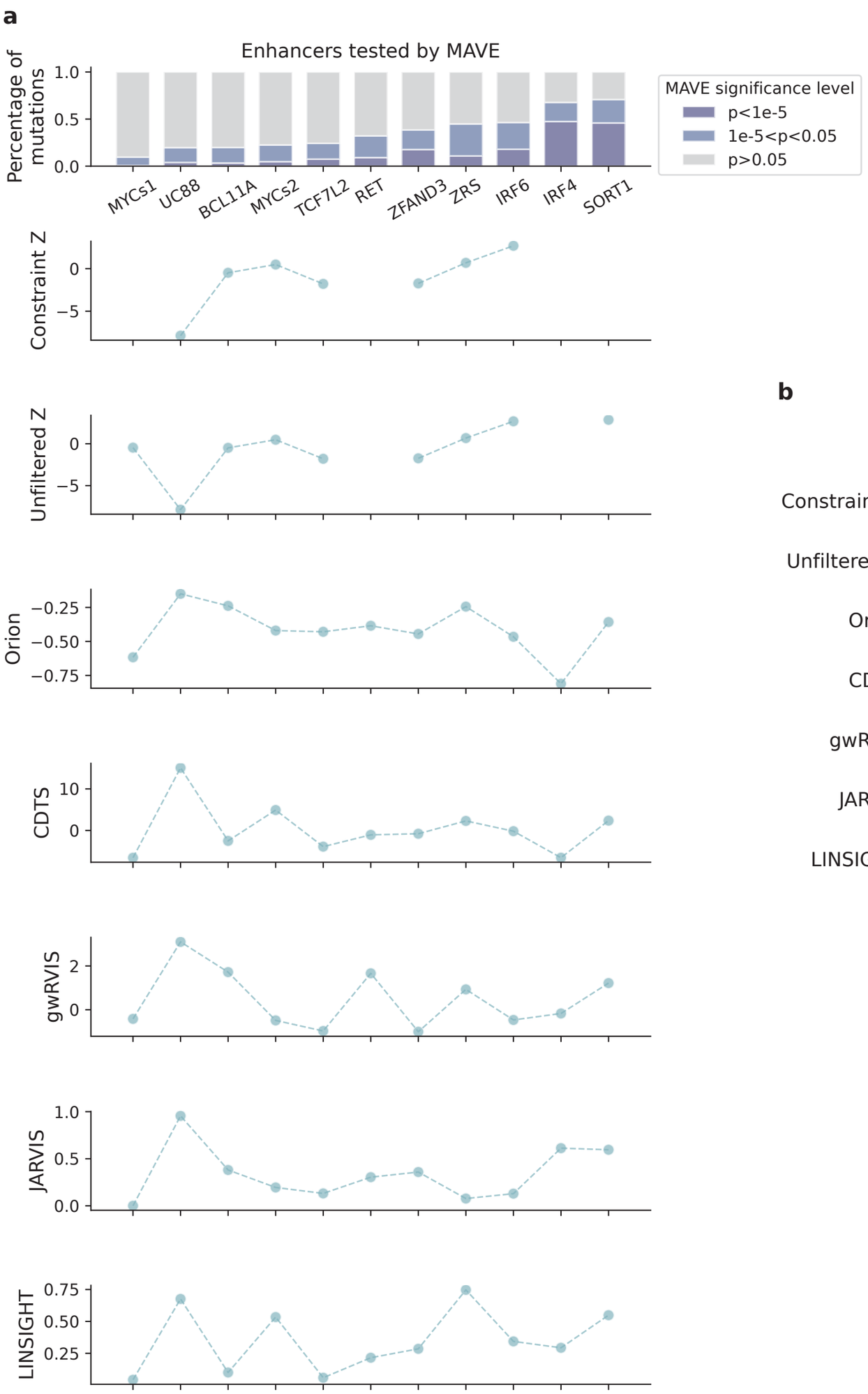
Extended Figure Data 5

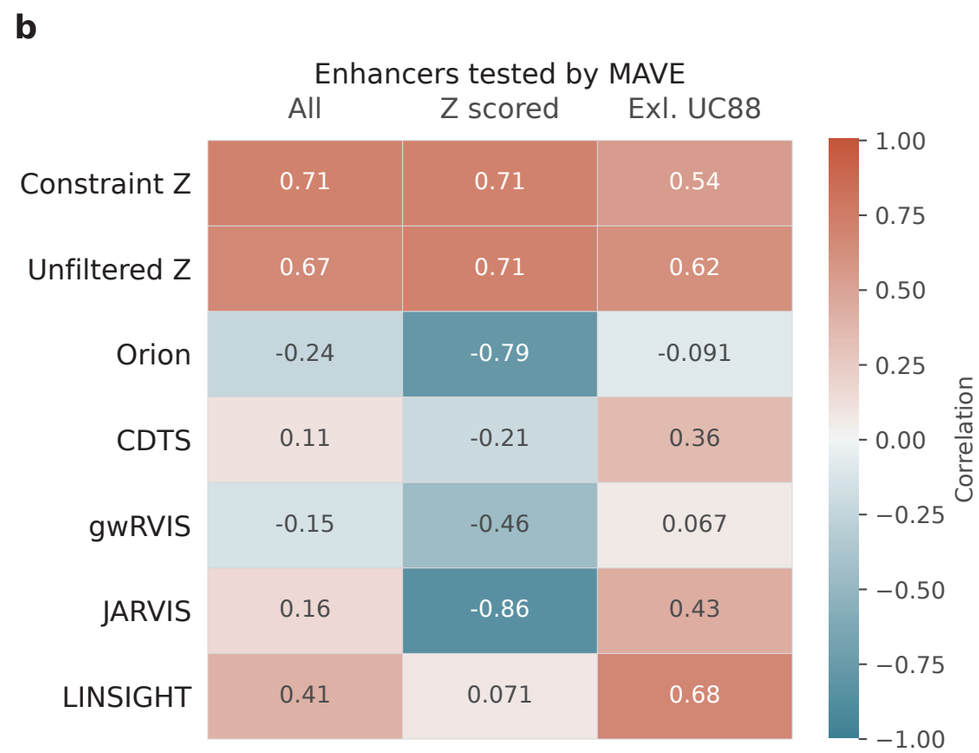




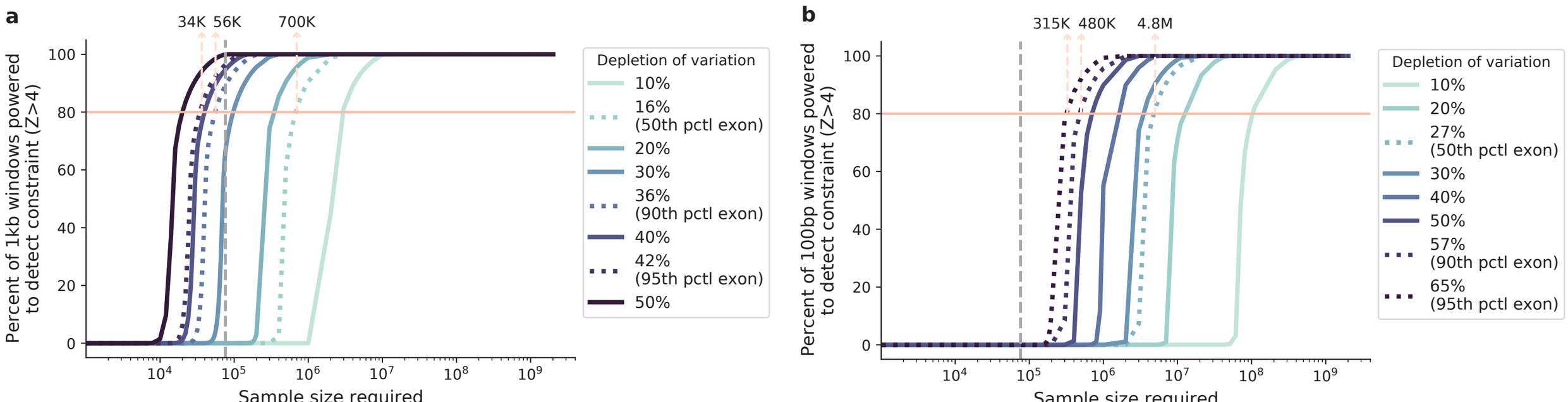
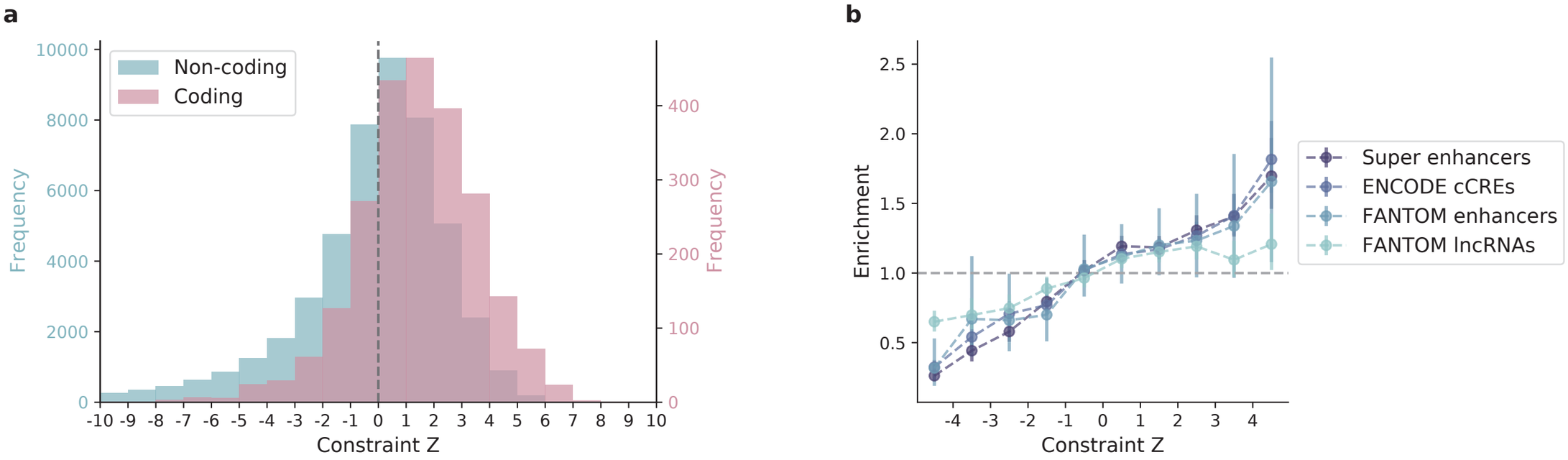


Extended Data Figure 6





Extended Data Figure 7



Extended Data Figure 8