nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0062-7

The human noncoding genome defined by genetic diversity

Julia di Iulio[1,5], Istvan Bartha[1,6], Emily H. M. Wong[1], Hung-Chun Yu[1], Victor Lavrenko[1], Dongchan Yang[2], Inkyung Jung[2], Michael A. Hicks[1], Naisha Shah[1], Ewen F. Kirkness[1], Martin M. Fabani[1,7], William H. Biggs[1], Bing Ren[3], J. Craig Venter[1,4] and Amalio Telenti[4,5]*

Understanding the significance of genetic variants in the noncoding genome is emerging as the next challenge in human genomics. We used the power of 11,257 whole-genome sequences and 16,384 heptamers (7-nt motifs) to build a map of sequence constraint for the human species. This build differed substantially from traditional maps of interspecies conservation and identified regulatory elements among the most constrained regions of the genome. Using new Hi-C experimental data, we describe a strong pattern of coordination over 2 Mb where the most constrained regulatory elements associate with the most essential genes. Constrained regions of the noncoding genome are up to 52-fold enriched for known pathogenic variants as compared to unconstrained regions (21-fold when compared to the genome average). This map of sequence constraint across thousands of individuals is an asset to help interpret noncoding elements in the human genome, prioritize variants and reconsider gene units at a larger scale.

There is a good understanding of the functional impact of protein-coding variants owing to historical studies of Mendelian disorders, the predictable consequences of amino acid changes and the recent availability of exome sequencing data[1]. However, protein-coding regions represent less than 2% of the total genome, and relatively little is known about the functional consequence of variation in the remaining 98% of the genome. The non-protein-coding sequences of the genome (hereafter referred to as 'noncoding') have been annotated through the ENCODE project, which relies on identification of biochemically active elements in the human genome, with attention paid to regulatory elements that control gene activity. Regulatory control is also influenced by higher-order chromatin structure[2]. In support of a role for noncoding variants in human disease and phenotypic traits, most of the over 16,000 common variants identified through genome-wide association studies (GWAS) are in noncoding regions of the genome (http://www.ebi.ac.uk/gwas). GWAS variants are increasingly being recognized as acting through changes in the regulatory circuitry[3–5]. Petrovski et al. used a combination of species conservation and human variation data to identify regions of proximal (UTRs) noncoding sequence to infer gene dosage sensitivity[6]. However, despite recent progress in the study of noncoding variants, it remains a substantial challenge to characterize the noncoding variants in the human genome, which grow by over 8,000 with each additional genome sequenced[7].

To better characterize the population variation in noncoding regions, we performed a comprehensive analysis of 11,257 whole-genome sequences (Supplementary Fig. 1 and Methods). We applied an approach[7] that exploits the contribution of thousands of elements in thousands of genomes. Metaprofiles integrate sequence variation and frequency across genomic landmarks with the same sequence, structure or function. Here we generated massive alignments of *k*-mers to determine the probabilities of variation of each nucleotide genome wide in the context of the surrounding nucleotides. Specifically, we exploited heptamers (7-mers) for the analysis; the heptanucleotide context was shown recently to explain more than 81% of variability in substitution probabilities[8]. The 16,384 unique heptamers present in the human genome vary greatly in abundance, between 1,941 and 6,332,326 counts per genome (Supplementary Fig. 2a). Heptamers were not evenly distributed across the genome, and some showed clear association with genomic elements (Fig. 1a). Each heptamer is characterized by unique rates of variation. To capture this property, we computed the rate and frequency of variation at the fourth nucleotide of each heptamer (Methods). The metric varied 95-fold across heptamers (between 0.0046 and 0.438; Supplementary Fig. 2b–d). The computed score was used to define the expectation of variation for each nucleotide in the genome.

A given heptamer or region may have rates of observed variation that are higher or lower than the rates estimated across the genome. We defined the context-dependent tolerance score (CDTS) as the absolute difference of the observed variation from the expected variation. Thereafter, we divided the genome into equally sized regions using a sliding window of 550 bp to study context-dependent constraint without consideration of existing annotation. On the basis of the CDTS, we ranked every region in the genome from the most context-dependent constrained regions (1st percentile) to the least context-dependent constrained regions (100th percentile) (Fig. 1b and Supplementary Fig. 3a). We identified patterns of enrichment and depletion for specific genomic elements across the spectrum of CDTS values (Fig. 1c; see Methods for the categorization of the genomic elements). As expected, protein-coding exons were strongly enriched in the 1st percentile of CDTS (49-fold when compared to the 100th percentile; 12-fold when compared to the genome average). The context-dependent correction also identified a notable enrichment for promoters in the most constrained regions of the genome (66-fold when compared to the 100th percentile; 23-fold when compared to the genome average). Super-enhancers were enriched at lower CDTS values, in a magnitude proportional to the number of cell types in which they were present (Supplementary Fig. 3b). In contrast, marked depletion was observed for territory associated

[1]Human Longevity, Inc., San Diego, CA, USA. [2]Department of Biological Science, Korea Advanced Institute of Science & Technology, Daejeon, Korea. [3]Ludwig Institute for Cancer Research, La Jolla, CA, USA. [4]J. Craig Venter Institute , La Jolla, CA, USA. [5]Present address: Scripps Research Institute , La Jolla, CA, USA. [6]Present address: Swiss Federal Institute of Technology, Lausanne, Switzerland. [7]Present address: Verogen , San Diego, CA, USA. *e-mail: atelenti@scripps.edu

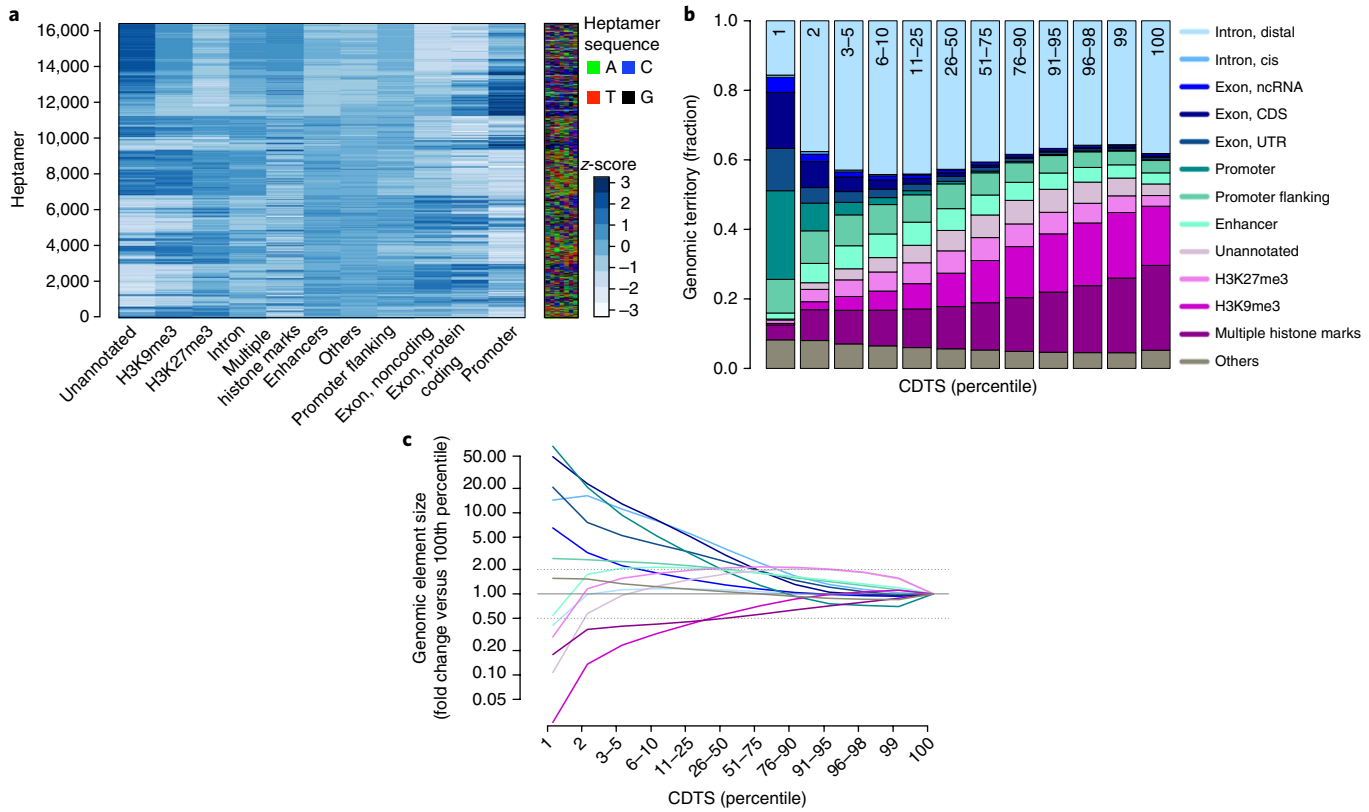

Fig. 1 | *k*-mer structure of the genome and composition of the constrained human genome. a, A blue-shades heat map representing the relative composition of *k*-mers for the different genomic elements. Each row corresponds to a heptamer, with the corresponding nucleotide sequence displayed on the heat map to the right. The relative abundance of a heptamer should be compared horizontally across the genomic element, with the shades of blue reflecting the *z* score. Before standardization, the counts of heptamers were normalized for each genomic element, to consider the different territory sizes of the element families. The order of the rows was obtained by hierarchical clustering. **b**, Bar plot displaying the cumulative territory fraction covered by each element family in the different percentile slices (indicated at the top of the bars). Here, and in other figures, we purposefully emphasize the patterns at the lowest 1st, 2nd and 3rd–5th percentiles (full display in Supplementary Fig. 3a). The percentiles are based on the rank of CDTS values. "Others" refers to Ensembl element families that did not cover a substantial part of the genome individually (such as transcription factor binding sites; Methods). The elements appear in the bar plot in the same order as in the legend. **c**, The enrichment and depletion of each percentile slice as compared to the value for the 100th percentile. The fold change is normalized by the size of the slice. Element families are colored as in **b**. The *y* axis is displayed in logarithmic scale. CDS, coding sequence; ncRNA, noncoding RNA.

with trimethylation of lysine 9 of histone H3 (H3K9me3) (39-fold when compared to the 100th percentile; 11-fold when compared to the genome average). However, it is important to underscore that the most context-dependent constrained regions of the genome contain representative examples of all families of genomic elements, as well as of unannotated genomic sequences (Fig. 1b, Supplementary Fig. 3c and Supplementary Table 1). Some chromosomes were characterized by larger content of constrained sequence (for example, chromosome 19, which has the highest gene density; Supplementary Fig. 4). Of note, the genomic element distribution was robust to changes in the study population (Supplementary Figs. 5 and 6). Therefore, further analyses were based on CDTSs computed with the subset of unrelated individuals ($n = 7{,}794$; Methods and Supplementary Fig. 1a). To compare these findings in the larger context of interspecies conservation, we evaluated the extent of overlap of constrained regions assessed with CDTS and genomic evolutionary rate profiling (GERP) across 34 mammalian species (i.e., interspecies conservation)[9]. From the 1st- to 10th-percentile levels, the overlap between the scores was limited and heavily biased for protein-coding regions (Supplementary Fig. 7). Schrider and Kern indicated that an important class of human-specific elements could be missed by searching for sequence conservation across species[10]. Our results suggest that constrained noncoding, regulatory regions in human populations can be identified by the CDTS.

A large proportion of the constrained human noncoding genome is associated with regulatory elements such as promoters, enhancers, transcription factor binding sites and regions associated with active chromatin marks (Fig. 1b and Supplementary Table 1). We hypothesized that the most constrained regulatory regions serve to regulate the most functionally important genes. To test the hypothesis, we used the notion of gene essentiality as a surrogate of 'functional importance'. A gene can be defined as essential when its loss of function compromises the viability of the individual or when it results in profound loss of fitness[11]. At the population level, identification of essential genes is supported by observing intolerance to loss-of-function variants[1,12–17]. We assigned the essentiality score of the gene to the corresponding upstream regulatory elements. This step confirmed that promoters in the constrained part of the genome associate with essential genes (Fig. 2a). We then observed that cis enhancer regions also shared sequence constraint with genes (within 15 kb) that were putatively regulated by those elements (Fig. 2a). A similar pattern was observed for other cis noncoding elements (for example, chromatin histone marks and transcription factor binding sites) and for unannotated and intronic regions, and it consistently identified coordination between the CDTS of noncoding or regulatory regions and gene essentiality (Fig. 2a). Overall, our data are consistent with previous evidence of purifying selection in regulatory regions in humans[18].

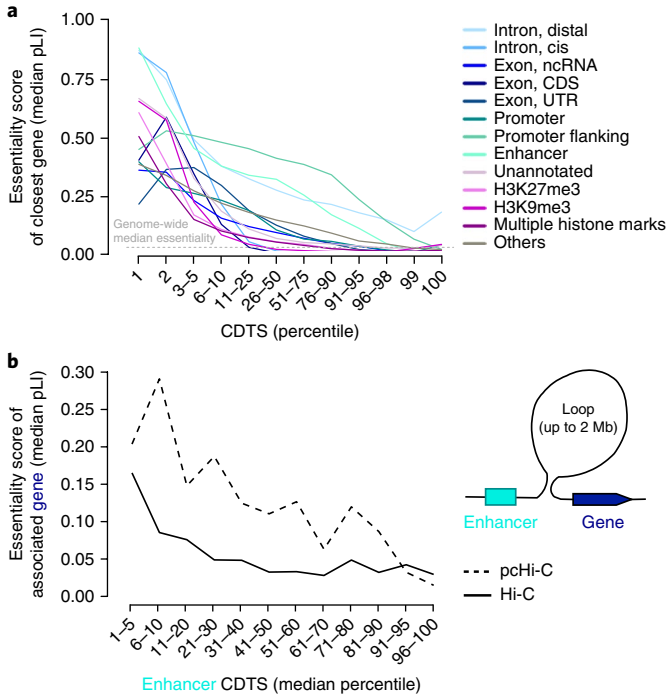

Fig. 2 | Coordinated constraint of genes and cis or distal regulatory elements. a, Coordination of cis elements. Each genomic bin within 15 kb of a gene (cis) was attributed the essentiality score (pLI score) of the closest gene. The median essentiality score of the closest genes is depicted on the *y* axis for each genomic element family across the CDTS spectrum (*x* axis). The gray horizontal dashed line represents the median gene essentiality score across the genome (0.028). pLI, probability of loss-of-function intolerance[1]. **b**, The CDTS–essentiality coordination of distal regulatory regions and their putative target gene. Gene–enhancer pairs were defined by in situ Hi-C[22] (solid line; $n=7{,}791$) or pcHi-C (dashed line; $n=2{,}658$) (Methods). Both approaches identified an association between the essentiality of the associated gene (*y* axis, pLI score) and context-dependent conservation of the distal regulatory element (*x* axis, CDTS percentile) up to 2 Mb.

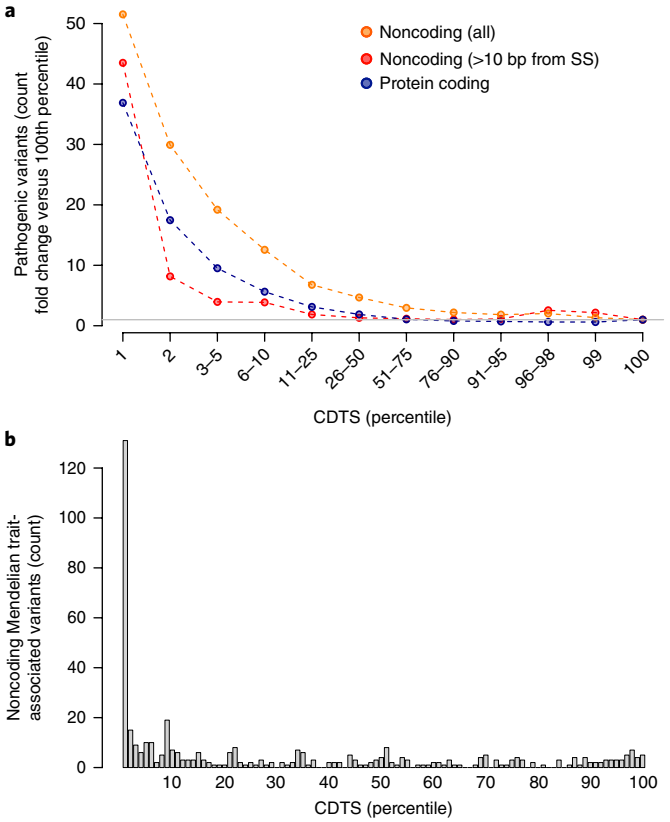

Fig. 3 | Distribution of pathogenic variants across the genome. a, The distribution of pathogenic variants across the different percentile slices, which identified a strong enrichment at lower CDTS percentiles. The relative enrichment was calculated relative to the value for the 100th percentile. The total numbers of pathogenic variants were as follows: $n=120{,}608$ protein-coding variants (dark blue) and $n=15{,}741$ noncoding variants (orange), including $n=1{,}369$ noncoding variants that are located > 10 bp from a splice-site (SS) position (red). The gray horizontal line indicates no enrichment or depletion (onefold). **b**, Noncoding pathogenic variants associated with Mendelian traits. Shown are 427 manually curated noncoding Mendelian pathogenic variants. Pathogenic variants were enriched at the lowest percentiles.

We assessed whether the observed coordination of the CDTS between promoters and the dependent gene could merely reflect the effect of selective sweeps on linked, neutral variation. Hernandez et al.[19] suggested that reduced diversity near exons could be explained by this mechanism. Under that model, high constraint near genes would not reflect direct selection on those elements but rather would reflect the pattern of the nearby gene. The analysis of 15 kb of sequence surrounding the first exon of genes generated two complementary patterns: (i) a global increase in conservation or constraint around exons when all genomic element annotations were considered that is accentuated in the proximity of essential genes (Supplementary Fig. 8a; annotations are detailed in Supplementary Fig. 8d) and (ii) a profound asymmetry of signals for annotated promoters as compared to intronic regions (Supplementary Fig. 8b,c,e). This pattern of asymmetry was also detected by Eigen, an unsupervised approach to integrate annotations in the human genome into one measure of functional importance[20] (Supplementary Fig. 8g), but was not captured by mammalian conservation, as assessed by GERP[9] (Supplementary Fig. 8f). We interpret these patterns as indicative of coordination of promoter constraint with that of the cognate exon and that this asymmetric pattern sits in a larger region that globally reflects the essentiality of the gene. Therefore, the signal cannot be interpreted solely as the result of linked neutral variation.

Next, we searched for evidence that functional constraints could be shared over greater distances. Topologically associating domains (TADs) and chromatin loops were defined using information from 3D genome structure[21,22] and from new promoter-capture Hi-C (pcHi-C) data (Methods). TADs were more constrained than non-TAD regions of the genome, particularly for those observed in multiple cell types (Supplementary Fig. 9a). Within TADs, the regulatory regions and associated genes were more constrained than the intervening loops (Supplementary Fig. 9b). We also identified a correlation between conservation of the distal enhancer and the essentiality of the target gene (Fig. 2b). Overall, the data support the concept of constrained and coordinated regulatory and coding units in the genome over large genome distances.

We next assessed whether CDTS ranking was a good proxy to score functionality and the consequences of mutations. For this purpose, we investigated the distribution of annotated pathogenic variants across the noncoding genome. The pattern of enrichment was 52-fold in the 1st versus 100th percentile of CDTS values (or 21-fold when compared to the genome average) for the subset of noncoding pathogenic variants ($n=1{,}369$) present > 10 bp from any splice site (Fig. 3a). The pattern of enrichment was 44-fold in the 1st versus 100th percentile of CDTS values (or 9-fold when compared to the genome average) for all pathogenic noncoding variants ($n=15{,}741$) (Fig. 3a and Supplementary

Table 2). There was no benefit of using the CDTS for the identification of protein-coding pathogenic variants when taking into consideration the large exome territory in low-CDTS domains (Supplementary Fig. 10a). We believe that the powerful biological constraints of the coding region (codon rules, the nature of missense modifications (such as polar or bulk residues), truncation, impact on active or structural sites, and proximity to splice sites) overwhelm any signal derived from heptamers at the current resolution.

Because there is the possibility of misclassification of pathogenicity in databases[23], we further investigated 427 manually curated noncoding variants associated with Mendelian disorders[24–27]. Mendelian noncoding variants were highly enriched in the regions with the lowest CDTS values (Fig. 3b and Supplementary Tables 3 and 4). It is also notable that, although GERP (interspecies conservation) performed extremely well to capture variants at splice sites (Supplementary Fig. 10b), the most stringent set of noncoding pathogenic variants ($n = 1{,}369$) were preferentially identified by CDTS rather than by GERP (Supplementary Fig. 10c). This offers support to the notion of human-specific constraint of functionally important regulatory regions and underscores the negative effect of genetic variation at those privileged sites.

We then explored how the CDTS compared to other functional predictive scores[9,20,28–31] that are used to prioritize variants in the noncoding genome. We focused the analysis to the stringent set of 1,369 noncoding pathogenic variants that were found further than 10 bp from any splice site. Eigen had the best performance of the metrics at a low false positive rate threshold, as represented by receiver operating characteristic (ROC) curves (Fig. 4a). Of the set of 1,369 noncoding pathogenic variants that were scored by at least one of the metrics, 713 were identified by at least one of the metrics as being in their top 1st-percentile score. CDTS captured the highest proportion of variants uniquely detected by a single metric (Fig. 4b). Other metrics capture more redundant information because they were developed or trained on similar datasets. In contrast, CDTS requires no prior knowledge and thus captures a very

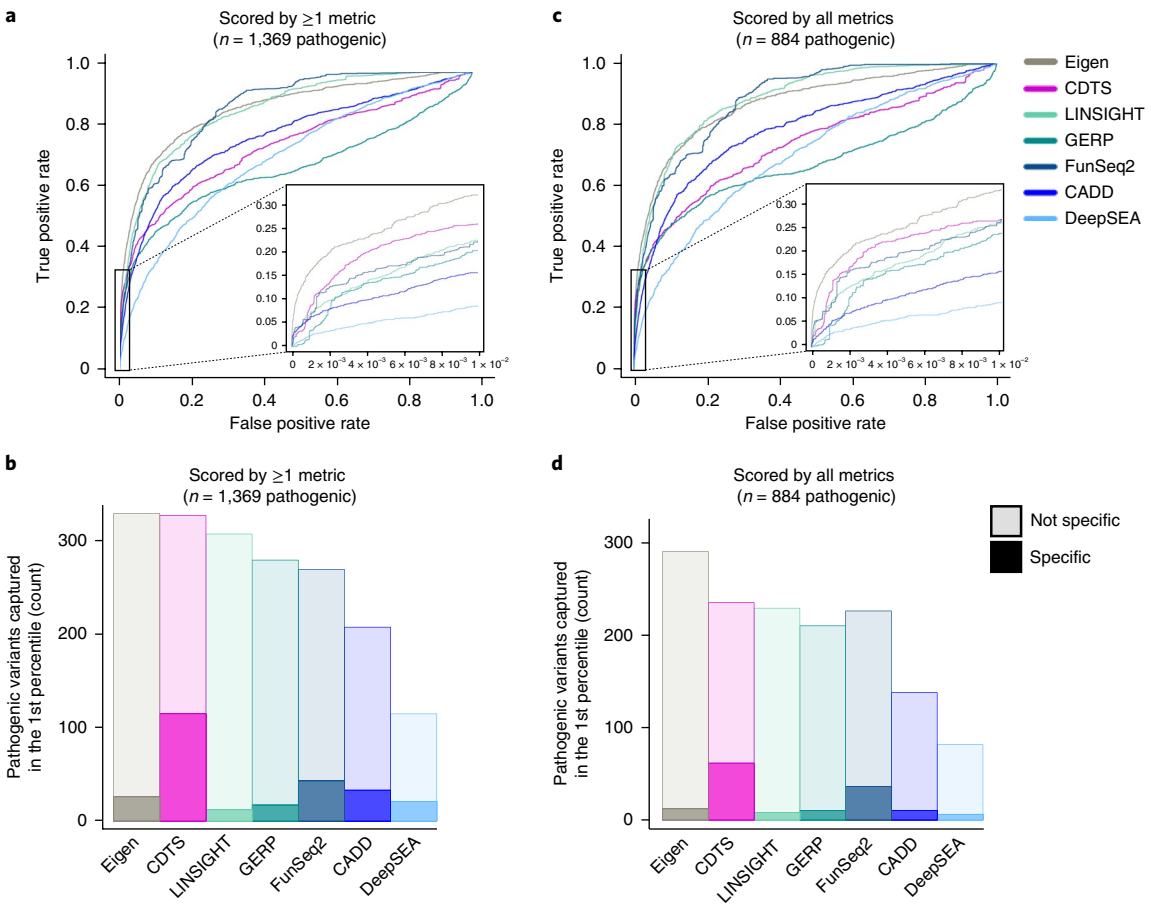

Fig. 4 | Performance and complementarity of CDTS and other metrics for noncoding variants. a, ROC curves for CDTS and six additional metrics. Noncoding variants scored by at least one metric were used for analysis ($n = 1{,}369$ pathogenic variants as cases, and >5 million variants with allelic frequency >0.05 as controls). The inset figure highlights the performance at the lowest false positive rates (*x* axis), which represents the most relevant segment for variant prioritization. **b**, Number of pathogenic variants identified by each metric at the 1st percentile. The same set of variants as in **a** was used. The darker hue represents the subset that was uniquely identified by a single metric, whereas the lighter hue represents the subset that was identified by at least two of the metrics. CDTS contributed a substantial number of uniquely identified variants, demonstrating its complementarity to the other metrics. The numbers of noncoding pathogenic variants scored per metric were as follows: CDTS ($n = 1{,}226$), Eigen ($n = 1{,}000$), CADD ($n = 1{,}283$), DeepSEA ($n = 1{,}324$), LINSIGHT ($n = 1{,}350$), GERP ($n = 1{,}354$) and FunSeq2 ($n = 1{,}203$). **c**, ROC curves for CDTS and six additional metrics. Only noncoding variants scored by all metrics were used for analysis ($n = 884$ pathogenic variants as cases, and >4 million variants with allelic frequency >0.05 as controls). The inset figure highlights the performance at the lowest false positive rate (*x* axis), which represents the most relevant segment for variant prioritization. **d**, Number of pathogenic variants identified by each metric at the 1st percentile. The same set of variants as in **c** was used. The darker hue represents the subset that was uniquely identified by a single metric, whereas the lighter hue represents the subset that was identified by at least two of the metrics. CADD, combined annotation-dependent depletion.

specific set of pathogenic variants that are not detected by other metrics. Similar results were obtained when analyzing only the subset of variants scored by all metrics (Fig. 4c,d). These data indicate that the CDTS complements other functional predictive scores in the analysis of the noncoding genome.

In summary, we assessed constraint of the human genome based solely on human variation. Its clinical relevance is manifested by the enrichment of known pathogenic variants in the constrained genome. A practical implementation of this observation is the targeting of sequencing efforts beyond the exome. Many exons could possibly be eliminated from targeted analysis while including an equivalent amount of sequence that represents the most constrained regions of the noncoding genome. The second notable observation is the complementarity of human conservation metrics with other analyses of the noncoding genome. Kellis et al.[32] reviewed the contribution of biochemical, evolutionary (interspecies conservation) and genetic approaches for defining the functional genome. They concluded that the combination of approaches was most informative. The final notable observation is the organization of functional units to share a conservation profile. The data indicate that an essential gene will use proximal and distant regulatory elements that are constrained. Use of this information supports the identification of cis or distal rare variants that regulate the expression of medically important genes.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0062-7.

Received: 16 November 2016; Accepted: 19 January 2018;
Published online: 26 February 2018

References

1. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
2. Bouwman, B. A. & de Laat, W. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biol.* **16**, 154 (2015).
3. Knight, J. C. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med.* **6**, 92 (2014).
4. GTEx Consortium. The Genotype–Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
5. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
6. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* **11**, e1005492 (2015).
7. Telenti, A. et al. Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. USA* **113**, 11901–11906 (2016).
8. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48**, 349–355 (2016).
9. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
10. Schrider, D. R. & Kern, A. D. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. *Genome Biol. Evol.* **7**, 3511–3528 (2015).
11. Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62 (2018).
12. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
13. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
14. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
15. Rackham, O. J., Shihab, H. A., Johnson, M. R. & Petretto, E. EvoTol: a protein-sequence-based evolutionary intolerance framework for disease gene prioritization. *Nucleic Acids Res.* **43**, e33 (2015).
16. Bartha, I. et al. The characteristics of heterozygous protein-truncating variants in the human genome. *PLoS Comput. Biol.* **11**, e1004647 (2015).
17. Fadista, J., Oskolkov, N., Hansson, O. & Groop, L. LoFtool: a gene intolerance score based on loss-of-function variants in 60,706 individuals. *Bioinformatics* **33**, 471–474 (2017).
18. Ward, L. D. & Kellis, M. Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". *Science* **340**, 682 (2013).
19. Hernandez, R. D. et al. Classic selective sweeps were rare in recent human evolution. *Science* **331**, 920–924 (2011).
20. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
21. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
22. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
23. Shah, N. et al. Identification of misclassified ClinVar variants using disease population prevalence. *Am. J. Hum. Genet.* (in the press).
24. Esteller, M. Noncoding RNAs in human disease. *Nat. Rev. Genet.* **12**, 861–874 (2011).
25. Makrythanasis, P. & Antonarakis, S. E. Pathogenic variants in non-protein-coding sequences. *Clin. Genet.* **84**, 422–428 (2013).
26. Gordon, C. T. & Lyonnet, S. Enhancer mutations and phenotype modularity. *Nat. Genet.* **46**, 3–4 (2014).
27. Smedley, D. et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am. J. Hum. Genet.* **99**, 595–606 (2016).
28. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
29. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep-learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
30. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
31. 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 (2017).
32. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. USA* **111**, 6131–6138 (2014).

Acknowledgements

We thank Human Longevity, Inc., for finnancial support.

Author contributions

J.d.I., J.C.V. and A.T. conceived and designed the study; J.d.I., I.B., E.H.M.W., H.-C.Y., M.A.H., N.S. and E.F.K. performed the analyses; V.L. established the search capability; M.M.F. and W.H.B. performed sequencing; D.Y., I.J. and B.R. performed pcHi-C; and J.d.I., E.H.M.W., B.R. and A.T. wrote the manuscript.

Competing interests

J.d.I., E.H.M.W., H.-C.Y., V.L., M.A.H., N.S., E.F.K., W.H.B. and J.C.V. are employees of Human Longevity, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0062-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Genomes. The analysis used deep-sequence genome data of 11,257 individuals ($n = 6,775$ females and $n = 4,482$ males), including 7,794 unrelated individuals ($n = 4,396$ females and $n = 3,398$ males). Read alignment (hg38 build) and variant calling was performed as previously described[7]. Analysis was limited to the high-confidence region of the genome as defined in Telenti et al.[7], a region covering approximately 84% of the genome and closely overlapping with the high-confidence region described in the most recent release of Genome in a Bottle (GiaB v3.2; ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.2.2/). The ancestry and population details are presented in Supplementary Fig. 1. The relatedness of individuals was established as previously reported[7]. All research involving human subjects was performed, and informed consent was obtained, under protocols approved by the Western Institutional Review Board (http://www.wirb.com)[7].

Metaprofiles. Metaprofiles consist of the massive alignment of elements of the same nature in the genome[7]. These genomic elements can be chosen based on their structure (for example, exonic, intronic, intergenic, etc.), function (for example, transcription factor binding sites, protein domains, etc.) or sequence composition (*k*-mers). Genetic diversity is assessed at each nucleotide position of the alignment of genomic elements by monitoring both the occurrence of variation in the population (reported as a binary—presence or absence) and the allelic frequency. More specifically, three metrics were computed at each position: (i) the fraction of elements with single-nucleotide variants (SNVs) (count score; Supplementary Fig. 2b), (ii) the fraction of SNVs with an allelic frequency > 0.0001 (frequency score; Supplementary Fig. 2c), and (iii) the product of both scores (tolerance score; Supplementary Fig. 2d). Each score was calculated using between 10^6 and 10^{10} values: the product of the number of aligned elements multiplied by the number of genomes sequenced. Therefore, the metaprofile strategy massively increased the power to compute the variation rate at nucleotide resolution with high precision. A priori knowledge of genomic landmarks is required for constructing metaprofiles based on similarity in structure or function. To remove potential biases using this a priori knowledge, we developed a strategy to construct metaprofiles based on all of the possible heptameric sequences found in the genome ($4^7 = 16,384$) and scored the middle nucleotide for each of these sequences as described above. Because every nucleotide in the genome was part of a heptamer, every single position could be attributed the corresponding genome-wide computed scores. Scores were computed separately for autosomes and for chromosome X. Because only the effective number of chromosomes was used to compute the allelic frequency, there was no need to include a normalizing factor to handle the surveyed number of chromosomes on chromosome X. To account for the difference in effective population size over history for chromosome X, the allelic frequency threshold was adjusted by a factor of 0.75 to have observation and expectation on the same scale for chromosome X and the autosomes. Given the high correlation of the tolerance score that was separately computed on autosomes and chromosome X (Supplementary Fig. 2e), chromosome X and autosomes were analyzed together genome wide. Insertions and deletions (indels) were not used to compute the score. When testing the score on smaller study populations consisting of <5,000 individuals (Supplementary Figs. 5 and 6), the allelic frequency threshold was adjusted to retain only non-singleton positions.

Expected versus observed variation. The variation rates computed through heptamer metaprofiles reflect the chemical propensity of a nucleotide to vary depending on its surrounding context and can be interpreted as an expectation of variation. We rationalized that functional regions would vary significantly less than they would be expected to, as assessed genome wide through the heptamer tolerance score. The observed regional tolerance score was the number of SNVs present at an allelic frequency >0.0001 in the studied population in a defined region. The expected regional tolerance score was the sum of the heptamer tolerance scores in the same region. To evaluate the departure from expectation, we compared the observed and expected tolerance scores obtained in defined genomic regions. The difference between the observed and expected scores was referred to as the context-dependent tolerance score (CDTS). Genomic regions were then ranked based on their CDTS. Regions with the lowest rank (1st percentile) have the lowest context-dependent tolerance to variation. Regions with the highest rank (100th percentile) have the highest context-dependent tolerance to variation.

Region definition and annotation. To avoid the use of a priori knowledge and biases due to the differing sizes of the regions (for example, more power to detect differences between observation and expectation in longer elements), the genome was divided into sliding windows of the same size. The window size was 550 bp, sliding every 10 bp, and the calculated CDTS across the 550-bp window was attributed to the middle 10-bp bin. The size of the window was chosen as the best compromise between resolution and power. Only regions with at least 90% of the nucleotides in the 550-bp window present in high-confidence regions were used. To evaluate the element distribution across these size-defined windows, we built a new annotation model by combining sources of annotation from GenCode (v.23) and Ensembl (annotated features and multicell regulatory elements, Ensembl v84 Regulatory Build). To avoid conflicting and overlapping annotations from the two different sources and thereby use the score of the same region multiple times, we prioritized element annotation such that only the highest-order element would be used: exonic protein-coding sequence (CDS), then exonic noncoding RNA, then exonic protein-coding UTR, then multicell, then intronic and then annotated features. We assessed the element composition of the different percentiles, using the combined GenCode/Ensembl annotation, by computing the number of nucleotides of an element in each percentile. The following categories were used: "Exon - CDS", which referred to protein-coding nucleotides in exonic regions contained in protein-coding genes as annotated in GenCode; "Exon - UTR", which referred to non-protein-coding nucleotides in exonic regions contained in protein-coding genes as annotated in GenCode; "Exon - ncRNA", which referred to nucleotides in exonic regions contained in noncoding RNAs (for example, snRNA, snoRNA and lincRNA) as annotated in GenCode; "Intron - distal", which referred to nucleotides in intronic regions contained in either protein-coding or noncoding genes located >10 bp from a splice-site position, as annotated in GenCode; "Intron - cis", which referred to nucleotides in intronic regions contained in either protein-coding or noncoding genes located within 10 bp of a splice-site position, as annotated in GenCode; "Promoter", "Promoter flanking" and "Enhancer", which referred to the nucleotides contained in the respective elements as annotated in Ensembl multicell regulatory elements; "H3K9me3" and "H3K27me3", which referred to the nucleotides that overlapped with (and only) the respective elements as annotated in Ensembl annotated features; "Multiple histone marks", which referred to the nucleotides that overlapped with a combination of histone marks, as annotated in Ensembl annotated features; "Others", which referred to the remaining nucleotides with Ensembl multicell regulatory element or annotated features that did not cover a substantial part of the genome individually and notably encompassed transcription factor binding sites, as well as other regulatory element combinations (for example, nucleotides annotated as both Promoter and Enhancer); and "Unannotated", which referred to nucleotides in regions that had no annotation in either GenCode or Ensembl. Super-enhancer annotation (Supplementary Fig. 3b) was obtained from dbSUPER (http://bioinfo.au.tsinghua.edu.cn/dbsuper/index.php). In Supplementary Fig. 8, every protein-coding isoform with more than one exon was considered for analysis. When the first exons of distinct isoforms overlapped, they were joined in a single exon. First exons present in genes with an essentiality score (using pLI score from Lek et al.[1]) >0.9 were defined as essential.

Essentiality and CDTS coordination. We used gene essentiality (pLI score from Lek et al.[1]) as an orthogonal proxy for functionality to assess whether genomic bins, annotated with the same genomic element, have different biological importance depending on their CDTS ranking. Each genomic bin present within 15 kb of a gene was attributed the essentiality score of its closest or overlapping gene, except for genomic bins annotated as "Promoters" that had the mandatory constraint of being upstream of the closest gene. The median essentiality score was then assessed per genomic element annotation and per percentile slice. To assess the possible coordination of distal gene–enhancer pairs, we used two external datasets, a Hi-C dataset aggregating the results of multiple cell types[22] and a new pcHi-C dataset in lymphoblast cell lines that identifies promoter-centered long-range interactions (provided in Supplementary Table 5 in hg19 coordinates). Briefly, the pcHi-C library was constructed by enriching target promoter-centered proximity-ligation fragments from a Hi-C library using capture RNA probes. Unmapped reads, non-uniquely mapped reads, PCR duplicates, trans-chromosomal read pairs, putative self-ligated products (<15-kb read pairs) and off-target reads were removed. We further eliminated experimental biases by using 'capture' scores, which denoted the probability of a region being captured. We considered only promoter-centered long-range interactions within the distance of 2 Mb from the transcription start site (using GenCode v.23 annotation). After removing distance-dependent background signals, only significant pcHi-C chromatin interactions (Weibull distribution, $P < 0.002$) were retained for analysis. Only distal interacting regions that were overlapping with an annotated enhancer (from Ensembl multicell annotation release 82) were kept for analysis. The enhancer median CDTS of each remaining distal interacting region was then compared to the essentiality score of the associated gene. When a distal interacting genomic region was associated with a promoter region that could pertain to more than one gene (for example, divergent genes), both pairs were retained. When the interacting genomic region overlapped with more than one enhancer, the median of the individual enhancer score was used. To pair distal enhancers with their hypothetically associated genes using the Hi-C dataset[22], we extracted, for each identified loop that was <2 Mb, the genes and enhancers that were the closest to both loop-anchor points. We then kept only meaningful pairs, in which an enhancer was annotated in the upstream anchor and a gene was annotated in the downstream anchor, or vice versa. In addition, the 5′ end of the gene had to be facing the loop-anchor point. A maximum of one pair per gene was considered; in the cases of several possible pairs, the pair that had the smallest total distance between the enhancer and the gene after subtracting the loop size was kept. We computed the median CDTS of the enhancers associated in such a distal gene–enhancer pair and compared it to the essentiality score of the associated gene.

Interspecies conservation. We used 'genomic evolutionary rate profiling' (GERP++)[9] to capture the interspecies conservation. GERP++ provides

conservation scores through the quantification of position-specific constraint in multiple-species alignments. We calculated and attributed the mean GERP scores to the same set of 10-bp bins as mentioned in the section "Region definition and annotation." Bins were ranked based on the GERP score from the most (percentile 1) to the least (percentile 100) conserved. Bins without a GERP score, due to an insufficient number of species in the alignment, were not considered in the ranking process.

Pathogenic variants. We assessed the distribution of known annotated pathogenic variants—which were defined as either HGMD high DM[33] (version: HGMD_2016_R1) or ClinVar variants, that were consistently annotated as pathogenic or likely pathogenic, and had at least one entry with star 1 or more[34,35] (version: ClinVarFullRelease_2016-07.xml.gz)—by counting the number of variants present in each percentile of the genome. Indels were not used to compute the CDTS. However, once a CDTS was attributed to a genomic region, the score could be used for any event in this region, including indels. Pathogenic variants with conflicting annotations—defined here as variants having a high DM in HGMD and a consistent annotation of benign or likely benign with at least one entry being star 1 or more in ClinVar—were removed. Supplementary Table 2 contains the list of all pathogenic noncoding variants ($n=15,741$) used in Fig. 3a and Supplementary Fig. 10b, which includes the more stringent set of variants that were >10 bp from any splice sites ($n=1,369$; used in Figs. 3a and 4, and Supplementary Fig. 10c). The noncoding variants associated with Mendelian traits ($n=427$; used in Fig. 3b) are provided in Supplementary Table 3. Those variants were extracted from ClinVar (copy number variants were excluded from analysis) and manually curated, and additional variants were collected by literature review[24–27]. A filter of >10 bp from any splice acceptor or splice donor site was applied, as splice-site variants have a high likelihood of being pathogenic and would not be a good control to test the model.

Functional predictive scores. The CDTS metric was compared to other metrics used for variant prioritization, such as CADD[28], Eigen[20], GERP[9], DeepSEA[29], LINSIGHT[31] and FunSeq2[30]. A control set of variants relative to the previously defined pathogenic variants ($n=1,369$, detailed in the subsection "Pathogenic variants") was created using variants from the dbSNP[36] database (June 2015 release). The control variants were defined as having the 'COMMON' and 'G5A' tag (>5% minor allele frequency in each population and all populations overall), being in a high-confidence region[7] and, similar to the tested pathogenic variant set, not being present in an exonic region and being >10 bp from any splice site. The remaining working set of noncoding pathogenic and control variants were ranked according to their CDTS, CADD, Eigen, GERP, DeepSEA, LINSIGHT or FunSeq2 scores, and the ranking was normalized from 0 to 100 (the direction of the values of the scores was modified if necessary so that, for all metrics, the lower rank would represent the pathogenic state). Of note, the CDTS ranking might differ slightly as only variant positions (control + pathogenic) were used here. To compare the different metrics, the true positive rate (TP/(TP + FN)) and false positive rate (FP/(FP + TN)) were computed at each step (threshold) of the new ranking. TP is the true positives, in this case the number of pathogenic variants with a ranking below or equal to the threshold; FP is the false positives, in this case the number of control variants with rank below or equal to the threshold; FN is the false negatives, in this case the number of pathogenic variants with a ranking above the threshold; TN is the true negatives, in this case the number of control variants with rank above the threshold; where the threshold can be any step in the new ranking (from 0 to 100). Given the fact that the control set of variants ($n>5$ million) is orders of magnitude larger than the pathogenic set ($n=1,369$), a false positive rate of 0.01 (the threshold used in Fig. 4a for the zoomed-in view) corresponded approximately to the 1st percentile of the data. Of note, not all variants were scored by all of the metrics (for example, no scores on chromosome X, conversion conflicts from hg19 to hg38; indels were not scored by all metrics or not in a high-confidence region). The numbers of noncoding pathogenic variants scored per metric are the following: CDTS ($n=1,226$), Eigen ($n=1,000$), CADD ($n=1,283$), DeepSEA ($n=1,324$), LINSIGHT ($n=1,350$), GERP ($n=1,354$) and FunSeq2 ($n=1,203$).

Software and external datasets. The pipeline to build the CDTS is provided at http://www.hli-opendata.com/noncoding/ and included the use of bedops (v2.4.14)[37], bedtools (v2.25.0)[38] and tabix (v1.2.1)[39]. The figures were plotted with R v3.3.2 (https://www.R-project.org/), notably using the package ggplot2 (http://ggplot2.org/). Data mining was performed using Python (v.2.7.13). CADD[28]-precomputed scores were downloaded from http://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs.tsv.gz, Eigen[20]-precomputed scores were downloaded from https://xioniti01.u.hpc.mssm.edu/v1.1/, FunSeq2[29]-precomputed scores were downloaded from http://org.gersteinlab.funseq.s3-website-us-east-1.amazonaws.com/funseq2.1.2/hg19_NCscore_funseq216.tsv.bgz, GERP[9]-precomputed scores were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz, LINSIGHT[29]-precomputed scores were downloaded from http://genome-mirror.cshl.edu/ (Table Browser; group: All Tracks; track: LINSIGHT; table: LINSIGHT), and finally DeepSEA scores were computed on the set of tested pathogenic and control variants using the docker image for running DeepSEA[29] (v.0.94) provided at https://github.com/gifford-lab/deepsea-predict-docker. pLI scores were obtained from Supplementary Table of Lek et al.[1].

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Data availability. Files with the genome-wide CDTS scores and allelic frequencies of the variants used for computing the CDTS metric are available for download at http://www.hli-opendata.com/noncoding. Access to biological data on pcHi-C is provided in Supplementary Table 5. Extended access to genomic data is possible through managed access agreement (http://hli-opendata.com/docs/HLIDataAccessAgreement101717.docx). Facilitated simple CDTS queries are also possible through http://www.hli-opensearch.com/: genome-wide use of the query terms CDTS1 and CDTS10 will return all of the variants with CDTS scores within the 1st and the 10th percentiles, respectively.

References

33. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
34. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44** (D1), D862–D868 (2016).
35. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
36. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
37. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
39. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).

nature research

Corresponding Author: Amalio Telenti

Date: 2018/01/16

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   No sample size calculation was done. We used whole genome sequencing data from all individuals with written consent at the time of analysis. The sample size is sufficient as we obtained comparable results with external dataset from 1000 Genome project and GnomAD, indicating the robustness of the approach.

2. Data exclusions

   Describe any data exclusions.

   we excluded samples that did not pass sequencing quality filters.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   The results were reproducible across different study populations.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   There was no group allocation.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   There was no group allocation

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	☐	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	python, ggplot2 package in R, unix commands. All analytical procedures are described in Online Methods section. In addition, a new paragraph "Software and External Data Sets" with the software and version used was added in the Online Method section.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	no unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	no antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	GM12878 and GM19240. Coriell Institute, https://www.coriell.org
b. Describe the method of cell line authentication used.	The cell lines have not been authenticated
c. Report whether the cell lines were tested for mycoplasma contamination.	The cells were not tested for mycoplasma contamination
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	no commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	no animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	no phenotypic covariate were relevant for the analysis.