nature genetics

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

The human noncoding genome defined by genetic diversity

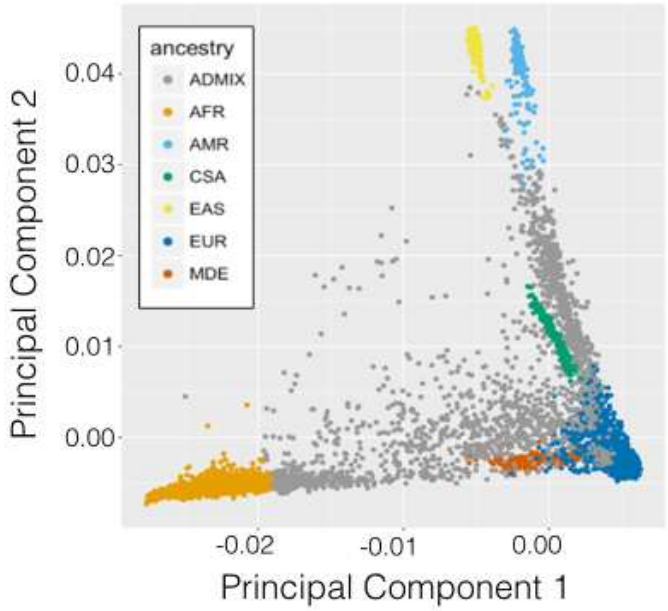
Julia di Iulio[1,5], Istvan Bartha[1,6], Emily H. M. Wong[1], Hung-Chun Yu[1], Victor Lavrenko[1], Dongchan Yang[2], Inkyung Jung[2], Michael A. Hicks[1], Naisha Shah[1], Ewen F. Kirkness[1], Martin M. Fabani[1,7], William H. Biggs[1], Bing Ren[3], J. Craig Venter[1,4] and Amalio Telenti[4,5]*

[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

A

Ancestry	Number of genomes	Number of unrelated genomes
EUR	6,725	4,436
AFR	1,234	1,087
CSA	280	137
EAS	236	171
MDE	265	94
AMR	132	106
ADMIX	2,385	1,763
Total	11,257	7,794

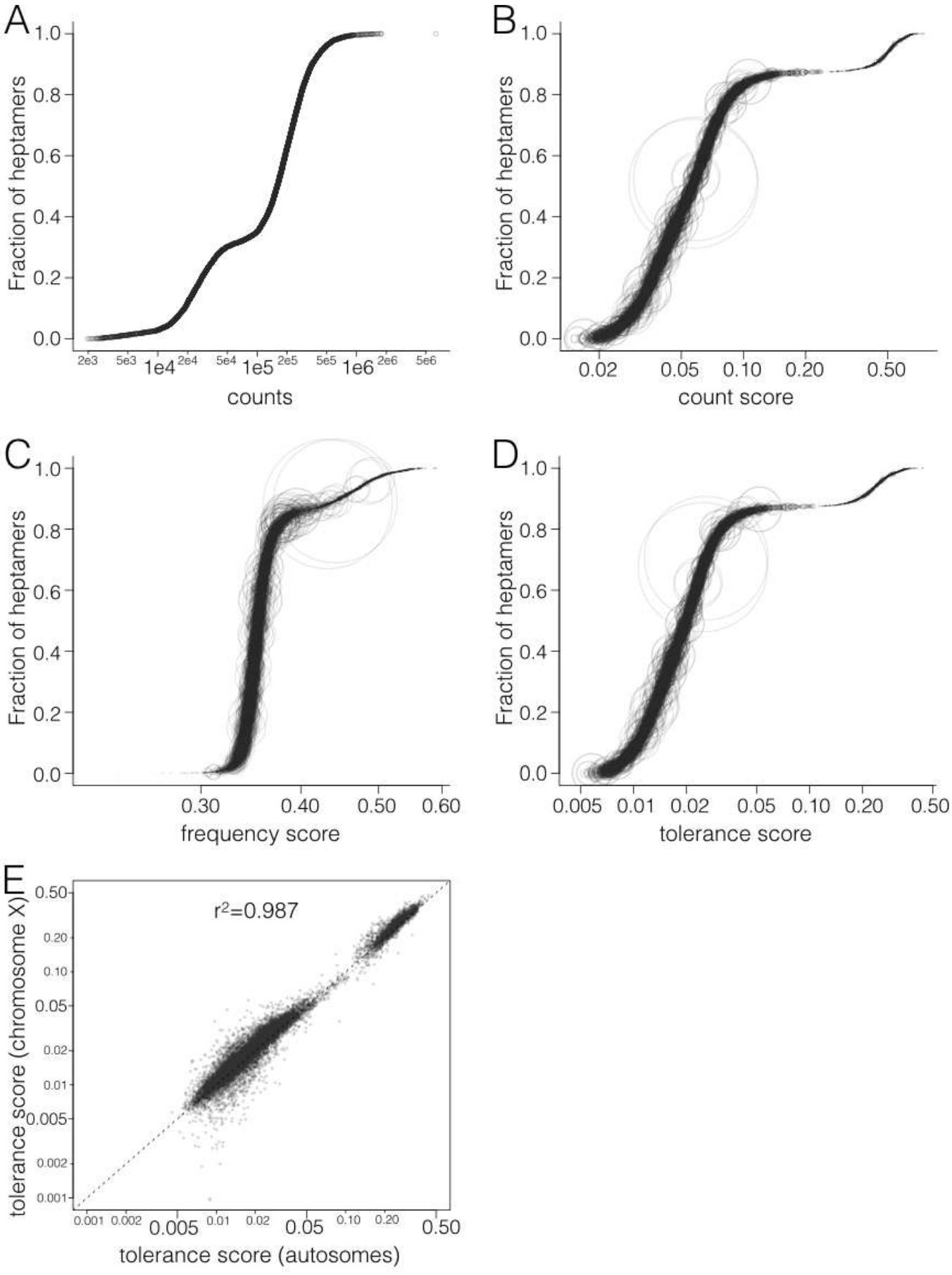
B



Supplementary Figure 1

Genetic ancestry of the study population

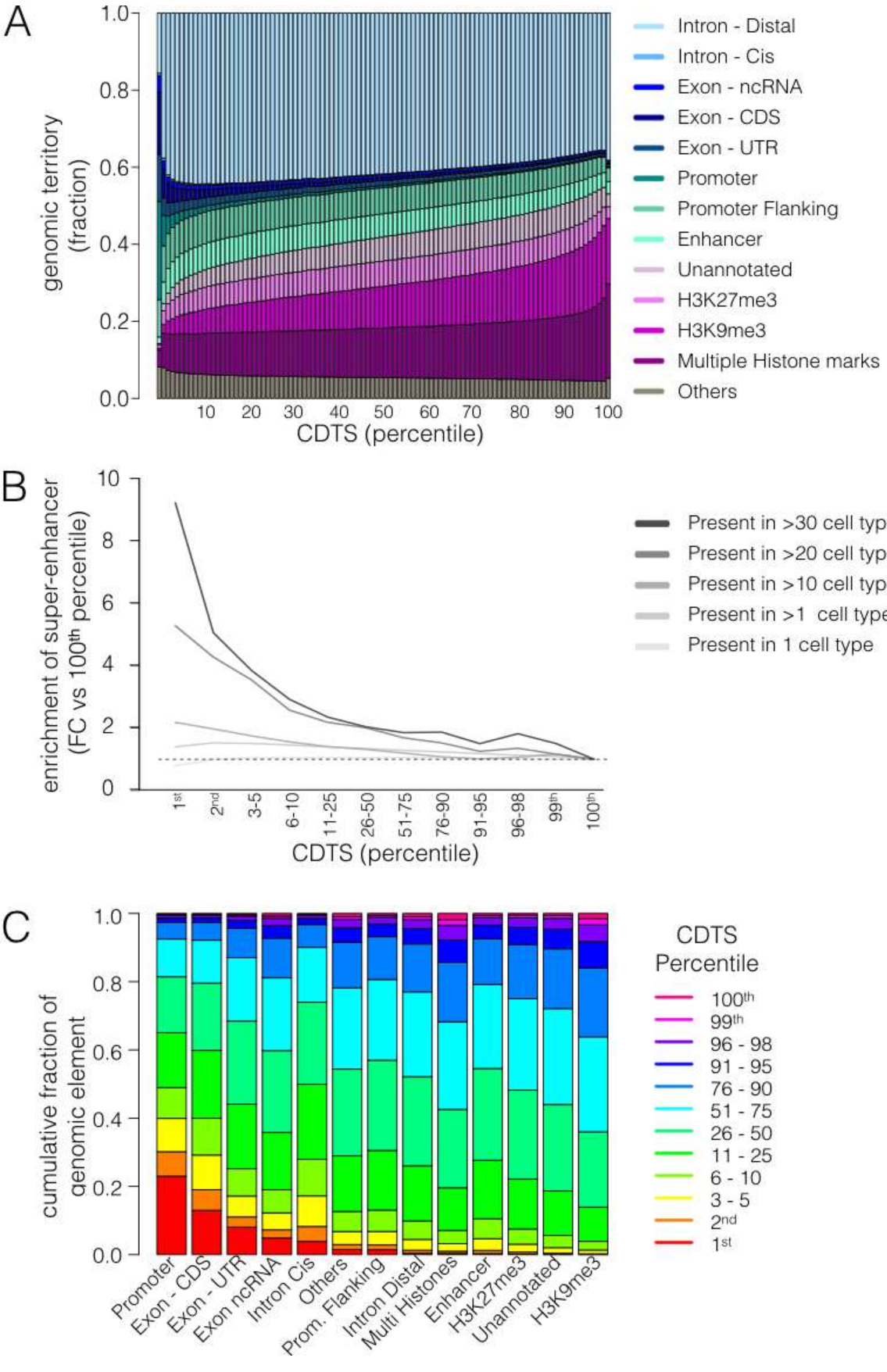
a, Number of genomes sharing each ancestry. **b**, Principal-component analysis (PCA) of the study population. PCA was performed using PLINK (1.9) on 162,997 ancestry-informative markers. Genomes are colored based on their major ancestries. EUR, European; AFR, African; EAS, East Asian; CSA, Central South Asian; ARM, Native American; ADMIX, admixed population group.


A
Fraction of heptamers
1.0
0.8
0.6
0.4
0.2
0.0
2e3
5e3
1e4
2e4
5e4
1e5
2e5
5e5
1e6
2e6
5e6
counts
B
Fraction of heptamers
0.02
0.05
0.10
0.20
0.50
count score
C
Fraction of heptamers
0.30
0.40
0.50
0.60
frequency score
D
Fraction of heptamers
0.005
0.01
0.02
0.05
0.10
0.20
0.50
tolerance score
E
r²=0.987
tolerance score (chromosome X)
0.50
0.20
0.10
0.05
0.02
0.01
0.005
0.002
0.001
tolerance score (autosomes)

Supplementary Figure 2

Heptamer metrics in the human genome

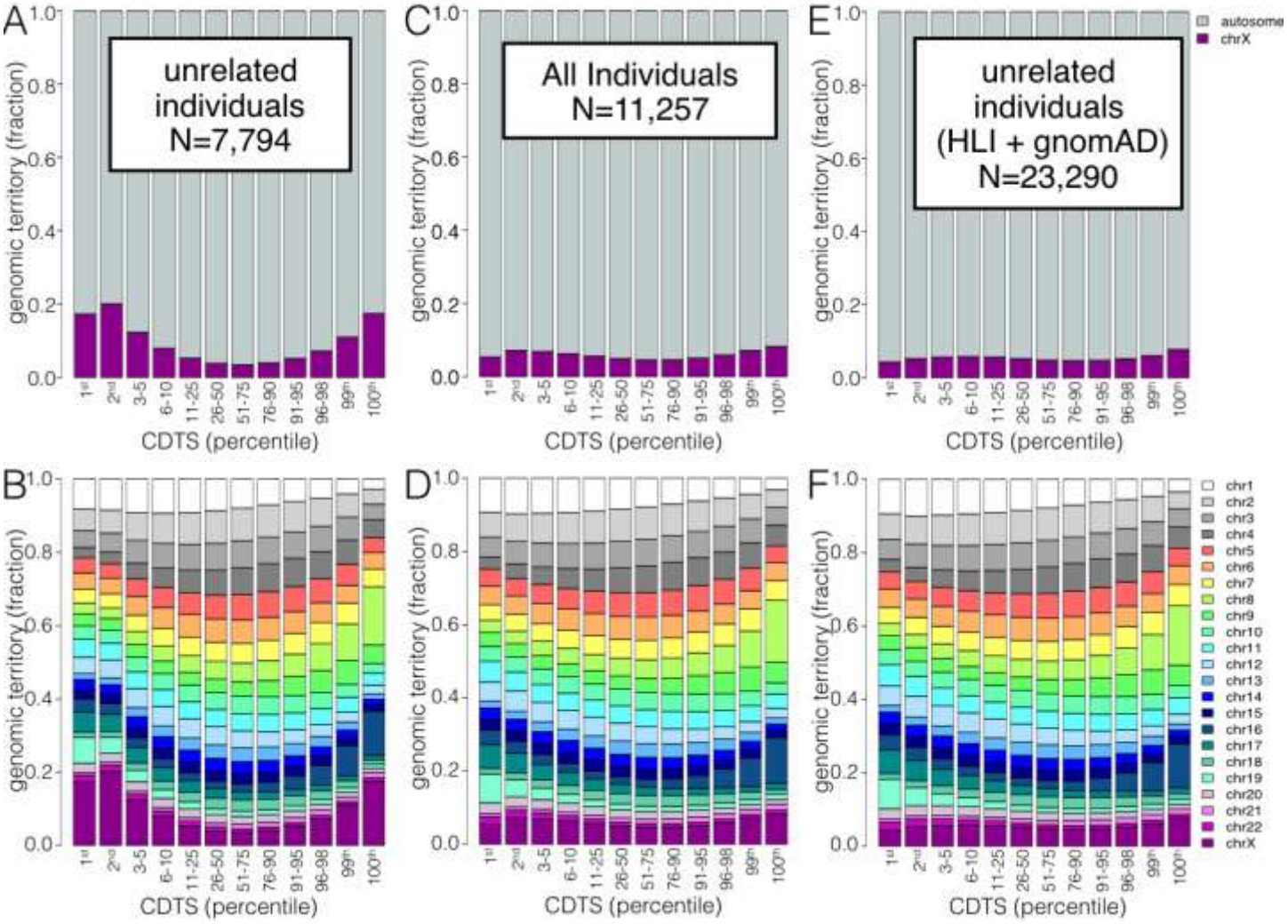
a, Cumulative distribution function of the total number of occurrence of each heptamer in the genome. Each dot (*n* = 16,384) represents a heptamer. **b**, Cumulative distribution function of the autosomal count scores. The count score represents the fraction of the middle nucleotide in a heptameric sequence that varies. Every circle (*n* = 16,384) represents a heptameric sequence. The size of the circles is proportional to the number of occurrences of the heptamer in the genome (plotted in **a**). **c**, Cumulative distribution function of the autosomal frequency scores. The frequency score represents the fraction of SNV at the middle nucleotide in a heptamer that varies with an allelic frequency >0.0001. Every circle (*n* = 16,384) represents a heptameric sequence. The size of the circles is proportional to the number of occurrences of the heptamer in the genome (plotted in **a**). **d**, Cumulative distribution function of the autosomal tolerance scores. The tolerance score represents the probability of the middle nucleotide in a heptamer varying with an AF >0.0001. Every circle (*n* = 16,384) represents a heptameric sequence. The size of the circles is proportional to the number of occurrences of the heptamer in the genome (plotted in **a**). **e**, Comparison of tolerance score separately computed on autosomes versus chromosome X. Each dot (*n* = 16,384) represents a heptamer. The r^2 represents the fraction of the variation explained by a linear regression model. The dashed line represents *x* = *y*. AF, allelic frequency; SNV, single-nucleotide variant.


A
genomic territory (fraction)
1.0
0.8
0.6
0.4
0.2
0.0
10 20 30 40 50 60 70 80 90 100
CDTS (percentile)
Intron - Distal
Intron - Cis
Exon - ncRNA
Exon - CDS
Exon - UTR
Promoter
Promoter Flanking
Enhancer
Unannotated
H3K27me3
H3K9me3
Multiple Histone marks
Others
B
enrichment of super-enhancer (FC vs 100th percentile)
10
8
6
4
2
0
1st 2nd 3-5 6-10 11-25 26-50 51-75 76-90 91-95 96-98 99th 100th
CDTS (percentile)
Present in >30 cell types
Present in >20 cell types
Present in >10 cell types
Present in >1 cell type
Present in 1 cell type
C
cumulative fraction of genomic element
1.0
0.8
0.6
0.4
0.2
0.0
Promoter
Exon - CDS
Exon - UTR
Exon ncRNA
Intron Cis
Others
Prom. Flanking
Intron Distal
Multi Histones
Enhancer
H3K27me3
Unannotated
H3K9me3
CDTS Percentile
100th
99th
96 - 98
91 - 95
76 - 90
51 - 75
26 - 50
11 - 25
6 - 10
3 - 5
2nd
1st

Supplementary Figure 3

Distribution of genomic elements within the CDTS spectrum

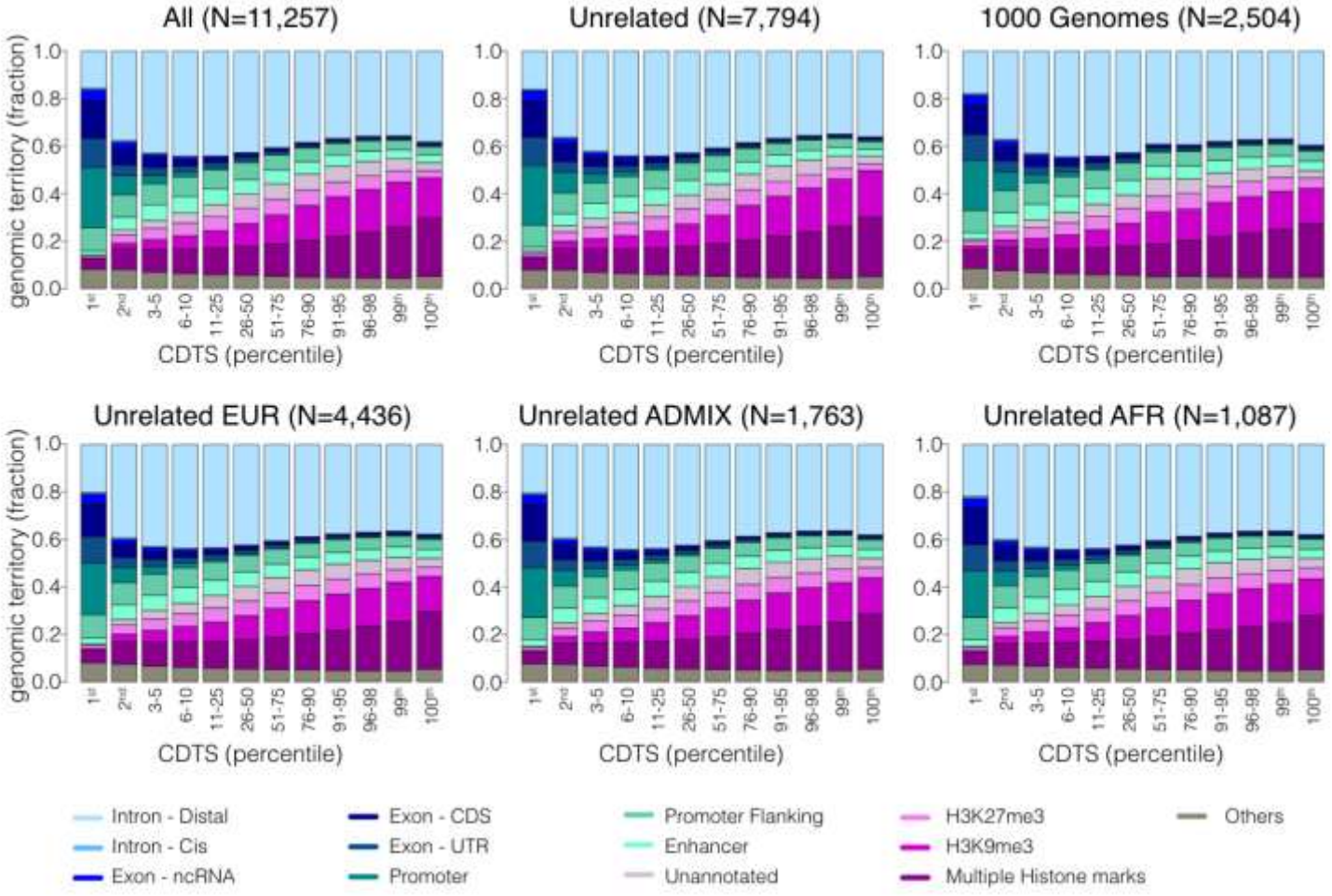
a, The bar plot displays the cumulative territory fraction covered by each element family at different percentiles (1 to 100). “Others” refers to ENCODE element families that did not cover a substantial part of the genome individually (such as transcription factor binding sites; Methods). The elements appear in the same order as in the legend. **b**, Size-normalized distribution of super-enhancer annotation. The relative enrichment of the fraction of enhancer bins overlapping with super-enhancer annotation is calculated with regard to the 100th percentile. Super-enhancers were subcategorized depending on the number of cell types in which they were annotated, represented by the lines of multiple shades of gray. **c**, The bar plot displays the distribution of the total number of nucleotides within the percentile slices for each element family. The boxes within a bar indicate the fraction of elements in each percentile slice (e.g., 23% of the promoters are within the 1st percentile). The element families are ordered on the *x* axis by the fraction of elements within the 1st-percentile slice. The coloring of the boxes is in the same order as in the legend. CDS, coding sequence; ncRNA, noncoding RNA; Prom., promoter; FC, fold change; CDTS, context-dependent tolerance score.



Supplementary Figure 4

Distribution of chromosomes within the CDTS spectrum

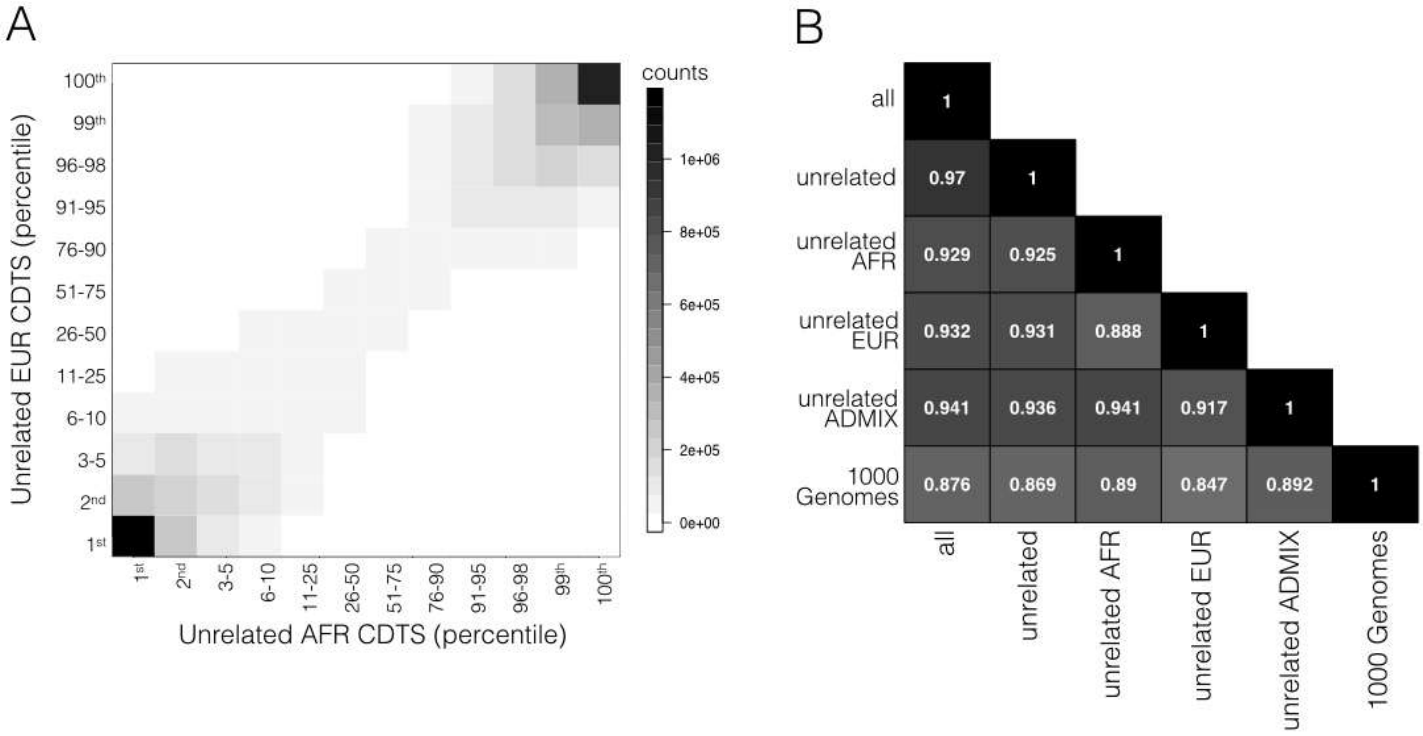
a, The bar plot displays the cumulative territory fraction covered by autosomes and chromosome X throughout the CDTS spectrum for unrelated individuals in the study (*n* = 7,794). **b**, The bar plot displays the cumulative territory fraction covered by each chromosome throughout the CDTS spectrum for unrelated individuals in the study. **c**, The bar plot displays the cumulative territory fraction covered by autosomes and chromosome X throughout the CDTS spectrum for all individuals in the study (*n* = 11,257). **d**, The bar plot displays the cumulative territory fraction covered by each chromosome throughout the CDTS spectrum for all individuals. **e**, The bar plot displays the cumulative territory fraction covered by autosomes and chromosome X throughout the CDTS spectrum for unrelated individuals (merged from this study and the gnomAD Consortium; *n* = 23,290). **f**, The bar plot displays the cumulative territory fraction covered by each chromosome throughout the CDTS spectrum for unrelated individuals merged from this study and the gnomAD Consortium. The coloring of the boxes is in the same order as in the legend. The difference in chromosome X distribution for the smaller population reflects the lack of power to discriminate variation at the allelic frequency threshold used. The distribution of chromosome X in “all individuals” and “unrelated individuals (merged from this study and gnomAD Consortium” is very similar and indicates that the distribution stabilizes after reaching a sufficient number of chromosome X alleles. The autosome distribution is not subject to the same noise in the smaller study population, as both males and females provide two allele counts each. CDTS, context-dependent tolerance score; HLI, Human Longevity, Inc.,; gnomAD, genome aggregation database (http://gnomad.broadinstitute.org/).



Supplementary Figure 5

Robustness of the approach with different study populations

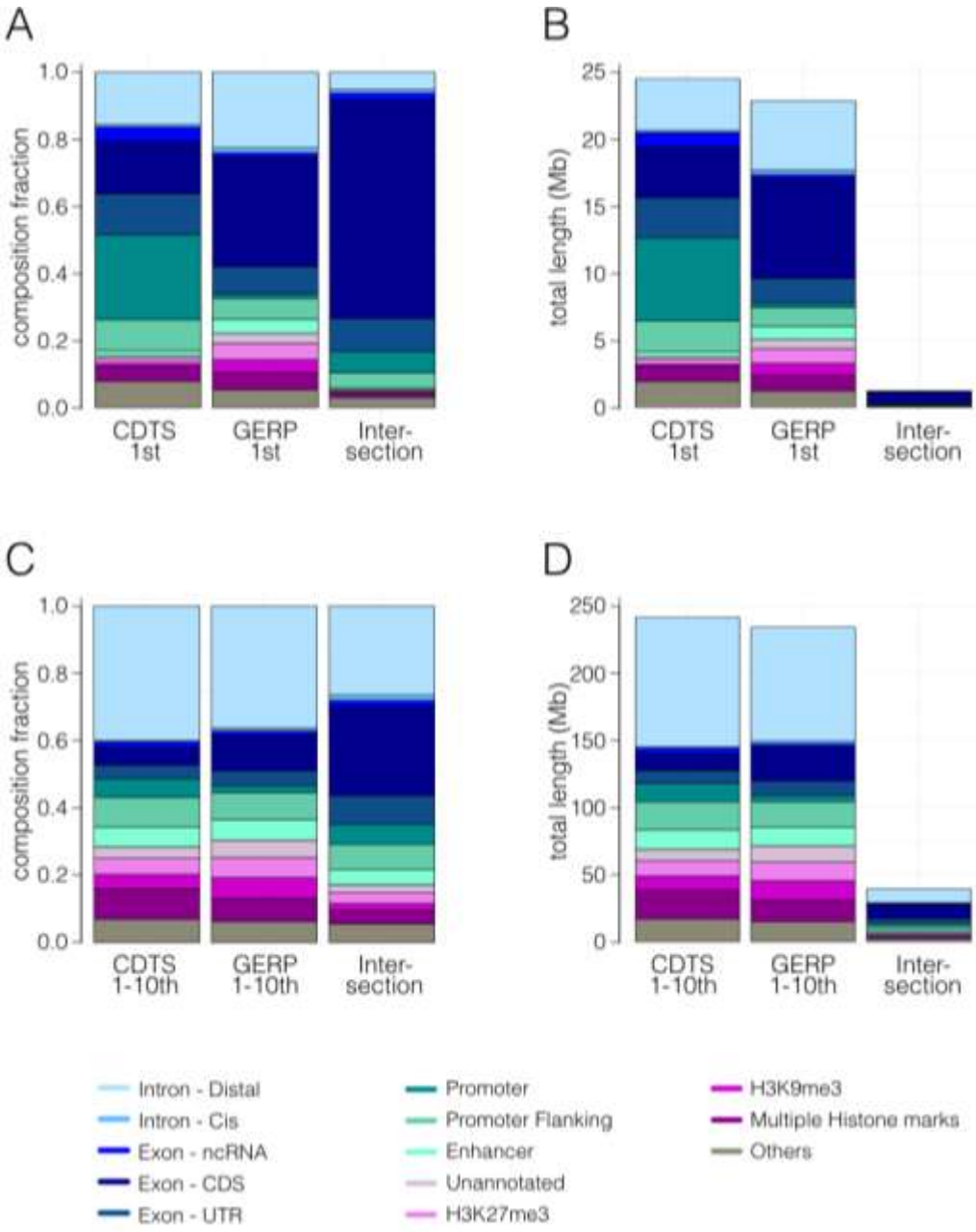
The bar plots display the cumulative territory fraction covered by each element family in the different percentile slices (indicated on the *x* axis). The percentiles are based on the rank of CDTS values. The similarity in distributions indicates that the CDTS metric is robust to downsampling or different population. “Others” refers to ENCODE 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 same order as in the legend depicted below the bar plots. Every bar plot was obtained by computing CDTS with a different study population or subset of study populations. Unrelated are a subset of All. Unrelated EUR, AFR and ADMIX are a subset of unrelated. CDS, coding sequence; ncRNA, noncoding RNA; EUR, European; AFR, African; ADMIX, admixed population group.



Supplementary Figure 6

Comparison of CDTS between study populations

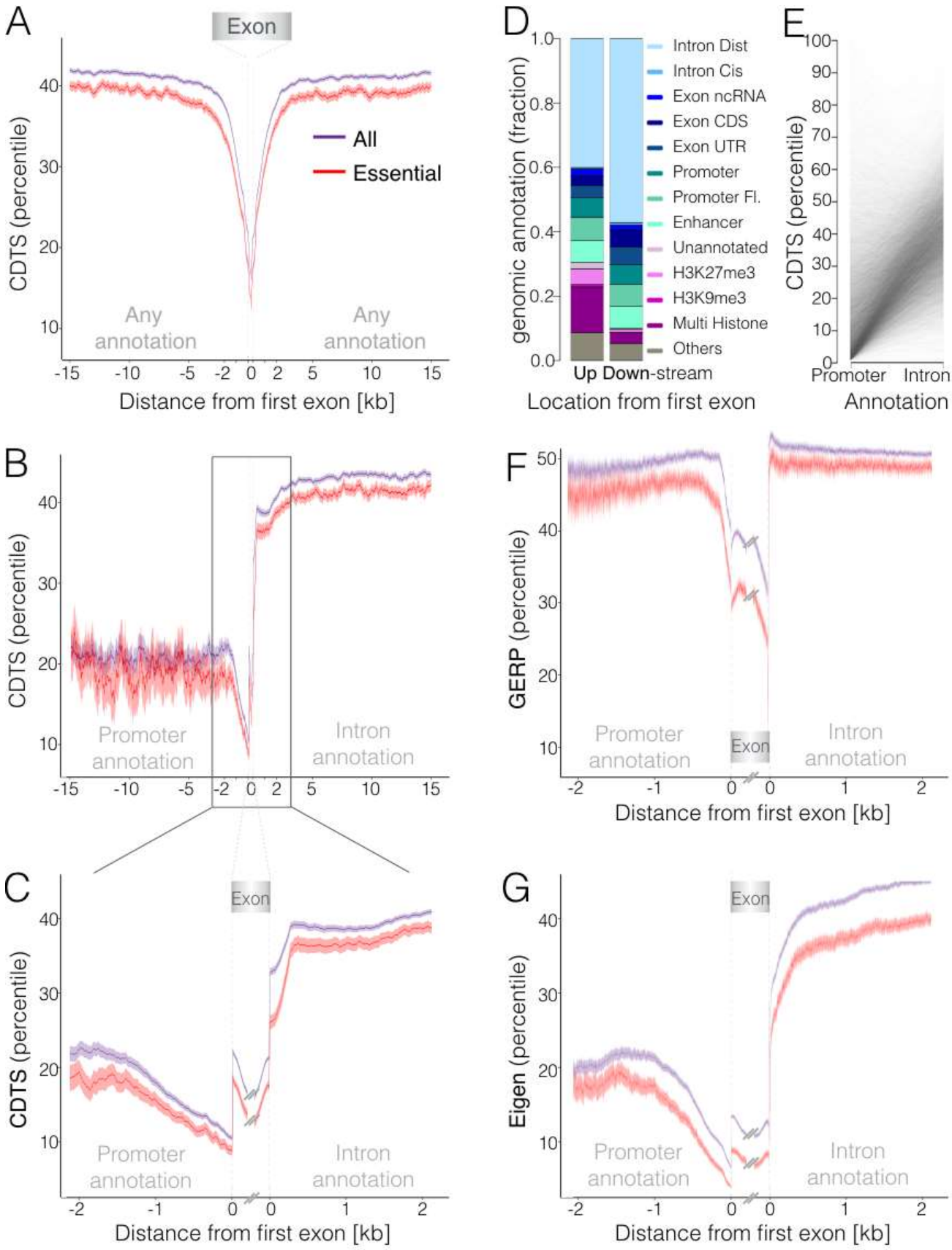
a, The heat map compares the CDTS percentiles computed with two different study populations: unrelated EUR (*n* = 4,436) and unrelated AFR (*n* = 1,087). The counts are normalized by the size of the respective percentile slices. The intensity of the coloring reflects the number of normalized counts. Overall matched CDTS percentiles are particularly dense at both ends of the spectrum. **b**, The figure illustrates the R^2 obtained through linear regression when comparing the CDTS percentiles of all study populations presented in Supplementary Fig. 5 (all, *n* = 11,257; unrelated, *n* = 7,794; unrelated AFR, *n* = 1,087; unrelated EUR, *n* = 4,436; unrelated ADMIX, *n* = 1,763; 1000 Genomes, *n* = 2,504). The linear regression for each comparison was computed with the percentile-slice-size-normalized counts, as depicted in **a**. There is strong agreement in genome domains that have high constraint across ancestries. However, we observed occasional differences among ancestry groups that will merit attention to separate technical noise (sequencing, alignment, limited data for some populations) from biologically relevant differences. One possibility is that recent population growth may have resulted in changes in the patterns of deleterious genetic variation and genome structure with consequences for fitness and disease architecture. Unrelated are a subset of All. Unrelated EUR, AFR and ADMIX are a subset of unrelated. EUR, European; AFR, African; ADMIX, admixed population group.



Supplementary Figure 7

Comparison of conserved regions assessed with CDTS and GERP

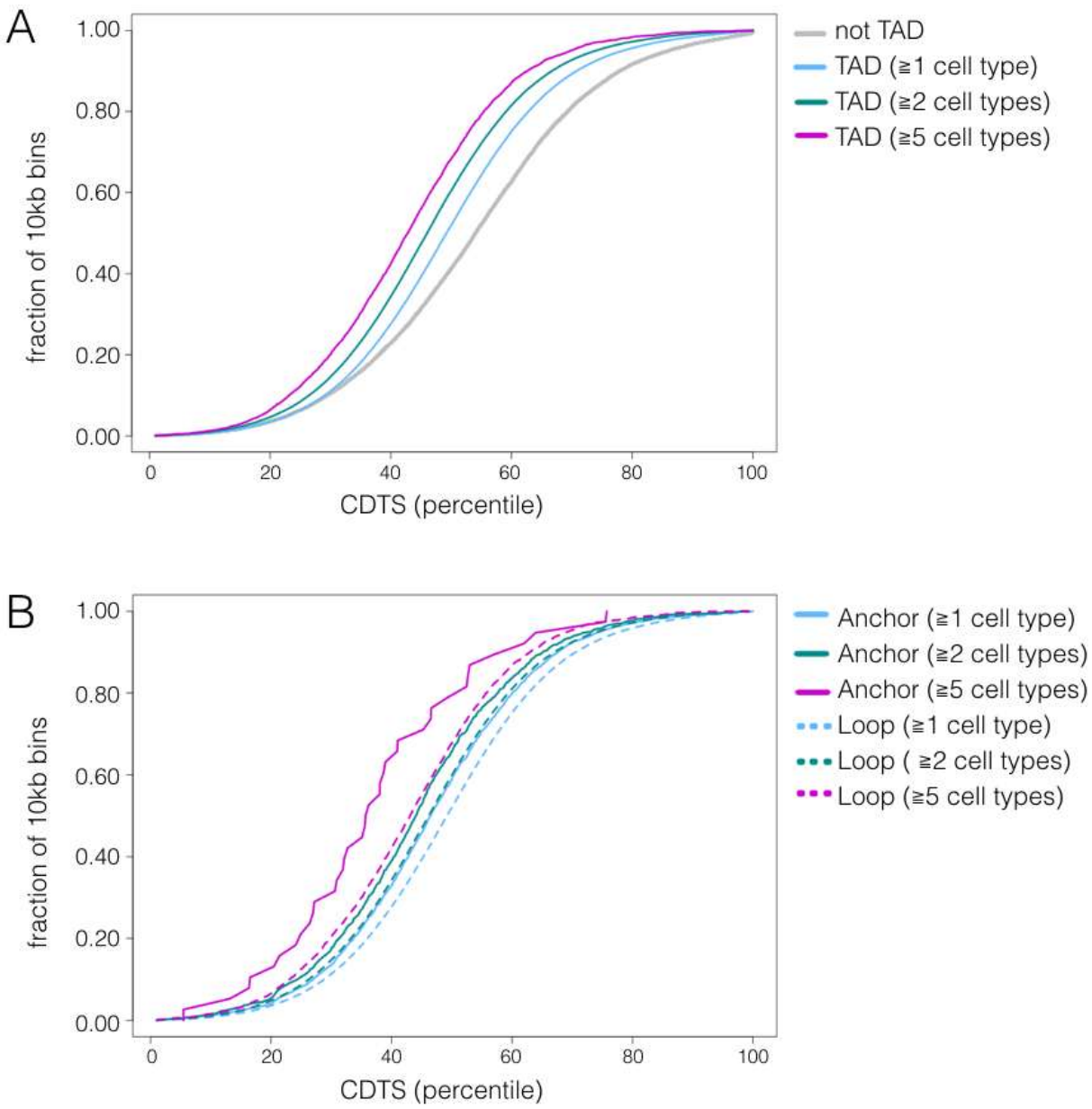
a, Element family composition in the 1st-percentile regions of CDTS (the bar labeled as “CTDS 1st”), GERP (“GERP 1st”) and the overlap region of CDTS and GERP (“Intersection”). Boxes in the bar correspond to different element families. “Others” refers to ENCODE element families that did not cover a substantial part of the genome individually (such as transcription factor binding sites; Methods). The coloring of the boxes is in the same order as in the legend. **b**, Absolute length of the 1st-percentile regions of CDTS, GERP and the overlap region of CDTS and GERP. Bins without GERP score, due to insufficient multiple-species alignments in the region, were not considered in the ranking process. This explains the total length difference between the 1st-percentile regions of CDTS and GERP. **c**, Element family composition in the first ten percentile regions of CDTS (the bar labeled as “CTDS 1–10th”), GERP (“GERP 1–10th”) and the overlap region (“Intersection”). **d**, Absolute length of the first ten percentile regions of CDTS, GERP and the overlap region of CDTS and GERP. CDS, coding sequence; ncRNA, noncoding RNA; CDTS, context-dependent tolerance score; GERP, Genomic Evolutionary Rate Profiling.


A
Exon
CDTS (percentile)
All
Essential
Any annotation
Any annotation
Distance from first exon [kb]
B
CDTS (percentile)
Promoter annotation
Intron annotation
C
Exon
CDTS (percentile)
Promoter annotation
Intron annotation
Distance from first exon [kb]
D
genomic annotation (fraction)
Intron Dist
Intron Cis
Exon ncRNA
Exon CDS
Exon UTR
Promoter
Promoter Fl.
Enhancer
Unannotated
H3K27me3
H3K9me3
Multi Histone
Others
Up Down-stream
Location from first exon
E
CDTS (percentile)
Promoter
Intron
Annotation
F
GERP (percentile)
Promoter annotation
Exon
Intron annotation
Distance from first exon [kb]
G
Exon
Eigen (percentile)
Promoter annotation
Intron annotation
Distance from first exon [kb]

Supplementary Figure 8

CDTS distribution near coding regions

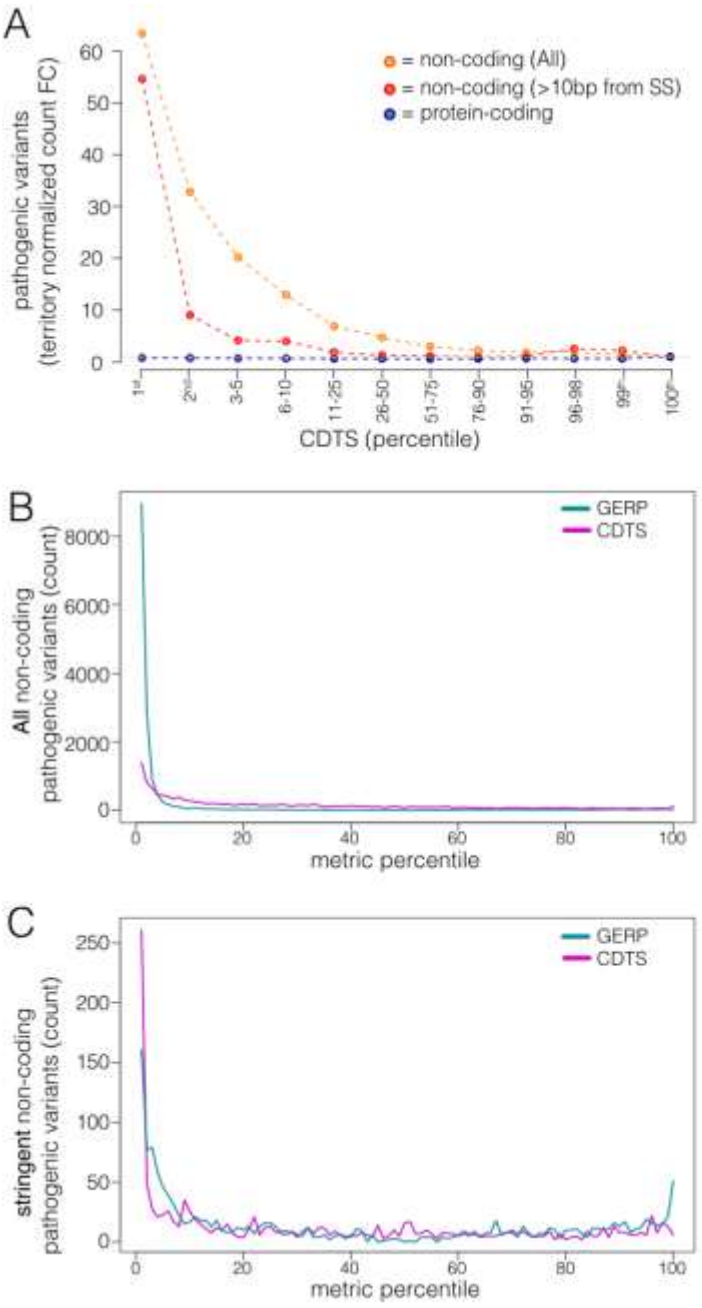
a, Mean CDTS values are depicted for a 15-kb window up- and downstream of first exons (*n* = 39,948 for “All genes/isoforms”, shown in purple; *n* = 9,176 for “Essential genes/isoforms”, shown in red; Methods). The regional profile is distinct, indicating a general pattern of constraint around exons with more profound constrain around exons of essential genes. “Any annotation” indicates any sequence surrounding the first exon. **b**,**c**, The apparent symmetry for regions up- and downstream of the first exons shown in **a** disappears when only regions annotated as promoters and introns (upstream and downstream, respectively, of the exons) are considered—in particular, in the immediate vicinity of the coding region (**c**). The asymmetric pattern supports the specific coordination between promoters and exons. **d**, The bar plots display the cumulative territory fraction covered by each element family upstream and downstream of the first exon (indicated on the *x* axis). As every protein-coding isoform was used to increase the power of the analysis, the annotation upstream/downstream of the first exon consists of a mixture of genomic elements. “Others” refers to ENCODE element families that did not cover a substantial part of the genome individually (such as transcription factor binding sites; Methods). The coloring of the boxes is in the same order as in the legend. **e**, Paired analysis of promoter:intron CDTS percentile. The upward signal indicates that the asymmetry in constrains surrounding the first exon is present in most genes/isoforms. **f**,**g**, GERP (**f**) and Eigen (**g**) mean percentile distributions in the vicinity of the first exon. The same set of exons were used as in **a**–**e**. Shaded regions represent 95% CIs. CDTS, context-dependent tolerance score; GERP, Genomic Evolutionary Rate Profiling.



Supplementary Figure 9

Properties of topologically associating domains

a, The plot depicts the cumulative distribution function of the mean CDTS values (in ≤10-kb windows) inside and outside TADs. TAD and non-TAD regions were divided into 10-kb windows (the overhang windows were discarded if smaller than 1 kb). The most constrained TAD windows are those identified by Hi-C as present in five or more cell types. TAD in at least one cell type versus no TAD: Kolmogorov–Smirnov two-sided test, $P = 2.2 \times 10^{-16}$. The total number of windows per group was as follows: non-TAD (n = 19,999 covering 139 Mb), TAD ≥1 cell type (n = 331,471 covering 2.4 Gb), TAD ≥2 cell types (n = 134,486 covering 911 Mb), TAD ≥5 cell types (n = 4,558 covering 29 Mb). **b**, The plot depicts the cumulative distribution function of the mean CDTS values (in ≤10-kb windows) of anchor and loop regions within TADs. Anchor and loop regions were divided into 10-kb windows (the overhang windows were discarded if smaller than 1 kb). The anchor regions are consistently more constrained than the loops within the same TADs. Anchor in at least one cell type versus loop in at least one cell type: Kolmogorov–Smirnov two-sided test, $P = 2.7 \times 10^{-14}$. The total number of windows per group is as follows: anchor ≥1 cell type (n = 2,954 covering 17 Mb), anchor ≥2 cell types (n = 1,321 covering 7 Mb), anchor ≥5 cell types (n = 38 covering 0.2 Mb), loop ≥1 cell type (n = 271,356 covering 1.8 Gb), loop ≥2 cell types (n = 117,581 covering 753 Mb), loop ≥5 cell types (n = 4,020 covering 24 Mb). TAD, topologically associating domain.



Supplementary Figure 10

The distribution of pathogenic variants

a, The distribution of pathogenic variants across the different percentile slices is normalized by the size of protein-coding and noncoding regions in the respective percentile slices. The relative enrichment is calculated with regard to the 100th percentile. The total number of pathogenic variants was as follows: *n* = 120,608 protein-coding variants (dark blue) and *n* = 15,741 noncoding variants (orange), including *n* = 1,369 variants that are located more than 10 bp from a splice-site position (red) **b**, The distribution of noncoding pathogenic variants is depicted for CDTS (pink) and GERP (green). GERP as expected best captured the larger set of variants (*n* = 15,741) that mostly consisted of splice-site variants. **c**, Outside of the exon boundaries (>10 bp; *n* = 1,369) both methods are enriched for pathogenic noncoding variants at their lowest percentiles; however, the enrichment is more striking with the CDTS metric. GERP misclassifies variants at the least conserved regions. CDTS, context-dependent tolerance score; GERP, Genomic Evolutionary Rate Profiling; FC, fold change.

Supplementary materials

The human non-coding genome defined by genetic diversity

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

Tables

Supplementary Table 1. Size-normalized distribution of histone and transcription factor binding sites. The normalization is done with regards to "multiple histone marks" territory for histones and to "Others" territory for transcription factor binding sites. The relative enrichment/depletion, as well as the sample size of each element, is given in the table. A loess regression, a linear regression and a linear regression performed on the log data was computed for each combination. (Provided as a separate File, hg38 coordinates)

Supplementary Table 2. Non-coding pathogenic variants from ClinVar and HGMD. The variants tagged as "stringent" are located >10bp from a splice site position. (Provided as a separate File, hg38 coordinates)

Supplementary Table 3. Description of non-coding variants associated with Mendelian traits. (Provided as a separate File, hg38 coordinates)

Supplementary Table 4. Distribution of non-coding variants associated with Mendelian traits per genomic elements.

Supplementary Table 5. Distal interacting regions and associated genes identified by pcHi-C. (Provided as a separate File, hg19 coordinates)

Supplementary Table 4. Distribution of non-coding variants associated with Mendelian traits

	Total number of variants per Element	Number of variants within CDTS 1st percentile (% total)	Number of variants within CDTS 10th percentile (% total)
Intron (>10bp from SS)	124	5 (4.0)	23 (18.5)
UTR	170	88 (51.8)	114 (67.1)
Promoter	69	28 (40.1)	41 (59.4)
Promoter Flanking	9	5 (55.6)	8 (88.9)
Enhancer	20	3 (15.0)	6 (30.0)
Unannotated	0	0 (NA)	0 (NA)
H3K27me3	1	0 (0)	0 (0)
H3K9me3	0	0 (NA)	0 (NA)
Multiple Histone marks	19	0 (0)	16 (84.2)
Others	15	2 (13.3)	6 (40.0)
Total	427	131 (30.7)	214 (50.1)