Identification of essential regulatory elements in the human genome

Alex Wells[1], David Heckerman[2], Ali Torkamani[3], Li Yin[3], Bing Ren[4], Amalio Telenti[3,5,6*], Julia di Iulio[3,5,6*]

[1] Stanford University, Stanford, CA 94305

[2] Department of Computer Sciences, University of California Los Angeles, Los Angeles, CA 90024, USA.

[3] Scripps Research Translational Institute, La Jolla, CA 92037

[4] Ludwig Institute for Cancer Research, La Jolla, CA 92093

[5] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037

[6] Equal contribution

*Correspondence: jdiiulio@scripps.edu, atelenti@scripps.edu

The identification of essential regulatory elements is central to the understanding of the consequences of genetic variation. Here we use novel genomic data and machine learning techniques to map essential regulatory elements and to guide functional validation. We train an XGBoost model using 38 functional and structural features, including genome essentiality metrics, 3D genome organization and enhancer reporter STARR-seq data to differentiate between pathogenic and control non-coding genetic variants. We validate the accuracy of prediction by using data from tiling-deletion-based and CRISPR interference screens of activity of *cis*-regulatory elements. In neurodevelopmental disorders, the model (*ncER*, non-coding Essential Regulation) maps essential genomic segments within deletions and rearranged topologically associated domains linked to human disease. We show that the approach successfully identifies essential regulatory elements in the human genome.

There is rapid improvement in the understanding of the human genome, the organization of function, and the consequences of human genetic variation. This understanding enables multiple innovations in medical genetics. For example, exome sequencing accelerates the diagnosis and contributes to the clinical management of rare genetic disorders. However, exome sequencing only examines less than 2% of the genome sequence and the current diagnostic yield stands at around 30%[1-4]. The role of genetics in the remaining 70% of rare disorders is at present unknown, and the exact mechanisms of disease remain largely unexplored. One potential mechanism is through perturbation of important regulatory regions of the genome. Recent efforts at extending the search space from the coding to the immediate regulatory regions show that new pathogenic variants can be identified in a small fraction of cases[5,6]. In parallel, there are recent examples of diseases that implicate distal enhancers and changes in the 3D genome structure[6,7]. Thus, the next milestones in the interpretation of the human genome sequence will emerge from the analysis of functional consequences of genetic variants in the non-protein coding (here in referred to as non-coding) genome – the remaining 98% of genome sequence that includes the regulatory machinery. As previously done for coding genes[8], we define a regulatory element as essential when loss of its function may compromise viability of the individual or results in profound loss of fitness or in disease.

Interpretation of the non-coding genome requires the identification of landmarks, features and structures – analogous to the same first principles that aid the interpretation of the coding genome. This translates to the characterization of regulatory genomic elements, rules of functional essentiality and redundancy, and the interpretation of the organization of the genome. Genome-wide epigenomic maps have revealed hundreds of thousands regions showing signatures of enhancers, promoters, and other gene-regulatory elements[9]. However, the high-resolution dissection of driver nucleotides as well as the characteristics of essentiality of those regions remain limited at present[10]. There are multiple sources of data (biochemical, genetic, evolutionary) that convey functional information on the non-coding genome[11]. These data are used by different scoring algorithms [12-21] that aim at ranking variants according to their predicted deleteriousness. The accuracy of these methods typically increases through collective integration of multiple models, eg. ensemble-based classifiers[22]. In addition, there is an opportunity to increase the precision of functional and deleteriousness prediction by learning from novel data sources – in particular, resources that have not been included in previous analyses. Those novel sources of data include studies of the patterns of human-specific constraints that are revealed by population genomic analyses[23], analyses of 3D organization of the genome (eg., promoter capture Hi-C)[24,25], and from high throughput screens of enhancer function[26]. In this work we implement state-of-the art machine learning tools to rank-classify the most essential regulatory elements of the non-coding genome with an emphasis on the contribution of new data modalities. We then use tiling array deletion and CRISPR interference (CRISPRi) data to assess the accuracy and functional relevance of the predictions. Lastly, we use the new predictive tools to identify essential regulatory regions associated with human disease – in structural variants associated with the autism spectrum disorder (ASD) and in developmental disorders resulting from the disruption of topological associated domains (TADs). The study design is summarized in **Suppl. Figure S1**.

Training a model to identify essential regulatory elements. To train a supervised machine learning model, we included non-coding pathogenic variants from ClinVar[27] and Human Gene Mutation Database (HGMD)[28] (N= 1,095, **Methods**). The set of control variants was built by using all variants from gnomAD (http://gnomad.broadinstitute.org/) with allele frequencies greater than 1% across populations and sub-selecting (N=8,093) those that matched the pathogenic variant set based on distance to splice sites and genomic element distribution (**Methods**). For validation, we used non-coding pathogenic variants not included in the original dataset: an independent set of manually curated non-coding Mendelian pathogenic variants (n=425)[23], and a new release of ClinVar and HGMD (total of N=599, including N=245 mapping to the ncRNAs); **Methods**.

We trained an XGBoost model (https://xgboost.readthedocs.io), an implementation of gradient-boosted decision trees consisting of a collection of decision trees, where a node in a single decision tree splits the training data into subsets (deleterious versus benign). During testing, new variants with the same feature sets were given to each tree to make a prediction (essential or non-essential). The outputs of each tree were combined (“ensembling”) to generate a final prediction. Each variant in the dataset was annotated with 38 features from 4 major categories

(**Suppl. Table S1, Methods**). (i) Essentiality features, such as context-dependent tolerance score (CDTS)[23] and probability of loss-of-function intolerance (pLI)[33], among others. The latter was used by mapping each non-coding genetic variant to the closest gene and assigning the gene essentiality score of that gene to the corresponding variant. (ii) Chromatin structure features, such as chromosome conformation[23,25] data used either as a binary indicator to denote whether or not a given non-coding genomic position physically interacts with gene promoters, or as a continuous feature, by attributing the respective gene essentiality of the associated promoter to the distal interacting region. The loop and anchor features were also used as discrete values representing the number of cell lines where they were identified. (iii) Gene expression related features, such as readout of high-throughput enhancer functional screens[26], and (iv) existing non-coding deleteriousness metrics: CADD[13], ncEigen[14], FATHMM[17], FunSeq2[16], LINSIGHT[21], ORION[20], ReMM[18] and ncRVIS[29].

We scored the non-coding regions genome-wide. The result of this process was a score (*ncER*, non-coding Essential Regulation) for each nucleotide, ranging from 0 (non-essential) to 1 (essential). We evaluated the model performance on a test set comprising 20% of the data through 5-fold cross validation and assessed the generalization of the model on 3 independent non-overlapping sets, consisting of curated Mendelian and two hold-out HGMD and ClinVar variants, for validation of the performance of the classifier (**Methods**). The ensemble algorithm, ncER, with a Receiver Operating Characteristic (ROC) AUC of 93% and a Precision-Recall (PR) AUC of 75% on the test set outperformed previously reported deleteriousness metrics by at least 18% ROC-AUC and 31% PR-AUC (**Figures 1A** and **B**). The model generalized to other independent validation datasets, achieving a ROC-AUC ranging from 88% to 96% and a PR-AUC ranging from 63% to 86% (**Suppl. Figure S2).** The univariate importance of each input feature in the model is displayed in **Figure 1C**. Most of the features (34 out of 38) in the model contributed to the score. The top contributing features were CDTS[23] that measures human-specific genomic constrain, 3D organization features such as distal enhancers[46] of essential genes and GTEx[30] expression variance as indicator of tolerance to gene dysregulation. Because of sparsity of some of the data (for example, pcHi-C and Vista enhancers), some of the features did not contribute to the model because few variants mapped to informative positions. A model trained with only the new essentiality, 3D genome organization and gene expression features outperformed a model trained with only previously published metrics (**Figures 1D** and **E**). In summary, the increase in ROC AUC and PR-AUC observed after inclusion of new features in the models emphasizes their orthogonality to the previously reported scoring metrics.

For computing efficiency purposes and to increase the signal-to-noise ratio, we smoothed the scores over 10bp window and used the 10bp bin resolution for all subsequent analyses. We examined the nature of the most essential regions of the genome based on different ncER percentile thresholds (99.9th, 99.5th, 99th and 95th percentiles, representing 2.8, 14.1, 28.3 and 141 Mb of cumulative sequence). The distribution of essential genomic elements at each threshold is displayed in **Suppl. Figure S3A**. All types of genomic elements were represented in the most essential bins of the genome, although *cis*-regulatory and enhancer sequences were enriched in the highest percentiles. Essential regions were of small size, with the most common size range being the 10bp bin (**Suppl. Figure S3B**). To have an overview of the putative function of those essential regulatory regions, we did pathway analyses for the set of genes (N=2,441) with at least one promoter bin in the top 99.9% ncER values. The most significant enriched biological process GO terms included development and regulation of gene expression, such as cerebral cortex development, positive/negative regulation of transcription and gene expression, among others (**Suppl. Figure S4**).

In summary, a model that trains on novel genomic features (essentiality, 3D organization, expression) adds precision to previous models that trained on partially orthogonal features (biochemistry, conservation). The model performs well in testing and generalization using new sets of human non-coding disease variants.

Functional correlates of essential regulatory elements. To confirm that ncER signals effectively captured essential regulatory functions, we analyzed two sets of functional data. The first analysis used high-throughput CRISPRi data from 1.29 Mb of sequence in the vicinity of two essential transcription factor genes, *GATA1* and *MYC*[31]. The library deployed more than 80,000 single guide RNA (sgRNAs) pairs tiled across the genomic loci. The readout of the study was cellular proliferation of K562 erythroleukemia cells. *GATA1* or *MYC* regions are characterized by a high median ncER score, in the 87th percentile (**Figure 2A** and **2B** and **Suppl. Figure S5**) – which is consistent with the biological importance of the loci. However, we observed a shift to a median 99th ncER percentile for the regions targeted by the most functional pairs of sgRNA probes (p=2.1e-15, compared to non-functional probes). There was a dosage effect: the more essential the region, the stronger the functional readout of CRISPRi (**Figure 2B**). The parameters of predictive performance and accuracy of ncER are shown in **Suppl. Table S2**. In summary, the sequences that control cellular proliferation, cell viability and gene expression of *GATA1* and *MYC* reside in the most essential regions of the locus.

The second dataset was generated using high-throughput scanning for *cis*-regulatory elements by tiling-deletion and sequencing (CREST-seq)[32]. The area investigated encompassed 2-Mb of the *POU5F1* locus in human embryonic stem cells. *POU5F1* encodes a transcription factor that plays a key role in embryonic development and stem cell pluripotency. Knockout of *POU5F1* is associated with embryonic mortality in the mouse and scores as an essential gene in humans (pLI score of 0.89)[8,33]. Thus, *POU5F1* is expected to use regulatory elements with features of essentiality[23]. CREST-seq identified 45 *cis*-regulatory elements, including 17 previously annotated as promoters of unrelated genes that, like typical enhancers, form extensive spatial contacts with the *POU5F1* promoter. The 45 enhancers of *POUSF1* were significantly more likely to be essential compared to random genomic loci of matched size (**Figure 2C**). For example, 56% of *POU5F1* enhancers reside in regions with ncER >99th percentile, compared to 14-17% for random genomic regions matched to enhancer size, p≤6.3e-09. Similarly, the enhancers contained the most essential regions within the locus in permutation analysis (**Suppl. Figure S6**). The parameters of predictive performance and accuracy of ncER are shown in **Suppl. Table S3** and **Suppl. Figure S7**.

In summary, the excellent performance of ncER scores for the identification of deleterious variants and for the prediction of functional read-outs in the non-coding genome may help map critical regulatory and structural elements of the non-coding genome in human disease.

Mapping essential regulatory elements in genetic diseases. We hypothesized that severe genetic diseases that do not have causal variants in the coding region could result from damage to essential functional elements. To investigate this concept, we chose two different models that represent challenges for the accurate mapping of the regulatory sites in relation to function and disease: (i) the identification of essential functional areas within non-coding structural variants/deletions associated with autism and ASD, and (ii) the impact of reorganization of TADs in the setting of a human developmental disorder that has been modeled in the mouse.

For the first disease model and as follow-up to our previous collaboration with Sebat et al.[34], we assessed a dataset of 136 transmitted *cis*-regulatory structural variants (SV, deletions) in the setting of autism and ASD. The deletions spanned a wide range of sizes: <1kb, N=32 probands and N=10 healthy sibling controls; 1-25kb, N=65 cases and N=4 controls; 25-100kb, N=17 cases and N=2 controls. The last group, >100kb deletions, found in N=6 probands did not have matched controls (**Suppl. Figure S8**). We observed that probands carry structural variants with only slightly higher ncER scores (median ncER percentile: 86 in probands versus 70 in healthy siblings for the <1kb deletions, p=0.21, **Suppl. Figure S9**). However, autism and ASD probands were more likely to carry structural variants with localized essential functional domains compared to healthy siblings (**Figure 3A**). For example, 41% of the <1kb deletions in probands contain regions with ncER >99th percentile, compared to 4% for random genomic regions matched to the size of the deletions, p=1.7e-10 and also compared to the few deletions present in non-probands (**Figure 3A**). Similar findings were observed for larger deletions (**Suppl. Figure S10**). As example, 20 unique <1kb deletions with at least one essential domain are displayed in **Figure 3B**. These results indicate that the identification of essential regulatory regions could map the most plausible causal region within a structural variant.

Next, we chose a human disease that involves the rearrangement of the regulatory landscape of *IHH* (encoding Indian hedgehog), which cause developmental defects including craniosynostosis and synpolydactyly[35,36]. Will et al.[37] identified nine enhancers with individual tissue specificities in the digit anlagen, growth plates, skull sutures and fingertips. The *IHH* region in humans shares a common structure with the mouse locus that was used for the model by Will et al.[37] In their study, consecutive deletions resulted in growth defects of the skull and long bones that confirmed that the enhancers function in an additive manner. Deletions and duplications caused dose-dependent upregulation and misexpression of *Ihh*, leading to abnormal phalanges, fusion of sutures and syndactyly. We identified the critical enhancers to reside in an extensive region of essentiality as scored by ncER, in particular for the regions shared across human duplications associated with disease (**Figure 3C**, Blue box). Within the locus, the critical enhancers were also endowed with features of essentiality (**Figure 3D**). For example, 67% of the enhancers contain regions with ncER >99th percentile, compared to 16% for random genomic regions matched to size, p=9.3e-04. Similarly, the nine enhancers carried the most essential regions within the locus in permutation analysis (**Supp**. **Figure S11**). Thus, highly essential enhancers that associate with human disease are correctly prioritized by computational approaches.

The present work contributes to the debate about reconciling redundancy and conservation of regulatory elements in the genome. Work on developmentally expressed genes supports the widespread existence of functionally redundant enhancers in mammalian genomes[38]. Redundancy reduces the likelihood of severe consequences resulting from genetic or environmental challenge. However, Osterwalder et al.[38] also suggest that contributions of enhancers to overall gene expression levels are relevant for organismal fitness under specific pressures, thus subjecting enhancers to purifying selection over evolutionary time. We have in the past indicated that essential genes will use proximal and distant regulatory elements that are conserved and constrained – thus representing essentiality in the non-coding genome[8,23]. We now show hallmarks of proximal or distal regions that regulate the expression of medically important genes. The current model supports the prioritization of variants and regions across the non-protein coding human genome for diagnostics and for functional analysis.

Materials and Methods

Training Features. To train our model, we leveraged a total of 38 features from 4 major categories: (i) gene essentiality, (ii) 3-dimentional chromatin structure, (iii) gene expression and other regulatory/functional data and (iv) existing variant pathogenicity/deleteriousness scores. A complete list of features along with their descriptions and accession links can be found in **Suppl. Table S1**.

We have previously identified a coordination of constrains between genes and their respective *cis* and distal regulatory elements[23]. We implement this concept in the present study by including the following essentiality features: (i) CDTS (our recently developed approach to score the non-coding genome essentiality, based on human genetic diversity[23], (ii) probability of loss-of-function intolerance (pLI)[33], (iii) haploinsufficiency score[39], (iv) gene dosage sensitivity score from ClinGen[40], (v) autosomal dominant or recessive categorization[41,42] and (vi) OMIM association [43]. For the metrics that solely provide scores for the genic portion of the genome, the respective essentiality features were calculated by mapping each non-coding genomic position to the nearest gene and assigning the corresponding essentiality metric score to the genomic position.

Chromatin 3D structure features included (i) nucleosome positioning extracted from MNase data (https://www.encodeproject.org/), (ii) multiple cell type anchor, loop and domain regions extracted from Hi-C data[44], (iii) frequently interacting regions (FIRE) and topologically associated domains (TADs) extracted from Schmitt et al.[45] and (iv) distal enhancer-TSS associations extracted from CAGE pairwise expression correlation (FANTOM)[46]. The 3D organization features were either used as binary indicators to denote whether or not a given non-coding genomic position physically interacted with gene promoters, or as discrete values representing the number of cell lines were the structures were identified. Finally, to combine both essentiality and chromatin structure features, we created distal essentiality features, by attributing the respective coding gene essentiality score (pLI) to distal regulatory elements identified through pcHi-C or CAGE pairwise expression correlations.

The model used gene expression, long non-coding RNA (lncRNA) annotations and functional regulatory data that have not been used by other existing metrics. Those included (i) median gene expression and variance across tissues (GTEx)[30], (ii) functionally tested genomic regions with enhancer activity identified through (ChIP-)STARR-seq experiments[26,47] or validated with transgenic mice[48] and (iii) lncRNAs identified through CAGE and transcriptome analysis[49].

Lastly, variant pathogenicity/deleteriousness scores used in the model included CADD[13], ncEigen[14], FATHMM[17], FunSeq2[16], LINSIGHT[21], ncRVIS[29], Orion[20] and ReMM[18]. We downloaded pre-computed genome-wide scores for each of these metrics (hg19 reference build). In the minority of cases where a per alternative variant score was provided, we used the most "deleterious" value at each position. For the metrics that solely provide scores for the genic portion of the genome, the respective features were calculated by mapping each non-coding genomic position to the nearest gene and assigning the corresponding metric score to the genomic position.

Training Set. We used a total of 9,188 single nucleotide variants (SNVs) to train the model. The pathogenic dataset comprised 1,095 non-coding SNVs, located at least 10bp from the nearest spice site, obtained from HGMD (2016_R1)[28] and ClinVar (July 2016)[27]. The selection criteria for HGMD SNVs were "DM and high" tags, while for ClinVar, SNVs had to be labeled as "Pathogenic" or "Likely Pathogenic", with star 1 or more and no conflicting assertion. HGMD were further filtered out for variants overlapping SNVs annotated as "benign" or "likely benign" in ClinVar (with star 1 or more and no conflicting assertion). The control genomic training set consisted of 8,093 variants chosen from a larger set of variants that were present in the gnomAD dataset at an allelic frequency > 1% and matched for the distance to the nearest splice sites and genomic elements. The matching was performed as follows: all pathogenic variants and gnomAD variants with allelic frequency > 1% were annotated with their respective distance to the closest splice site and the genomic element they mapped to (see Reference build, Annotation and Genomic Element Categorization section). For each pathogenic variant the subset of control variants falling within the same genomic element was extracted. Within this subset, the 10 control variants with the most similar distance to splice site as compared to the pathogenic variant were kept. Finally, duplicated control variants (if any) were removed from the final set. 80/20 percent of the variants were respectively used as training/test sets.

Machine Learning Model. We trained an XGBoost model in order to differentiate between pathogenic and control genomic positions in our training set. Hyperparameters were tuned using 5-fold cross validation and a randomized search method. 1,000 sets of randomly selected hyperparameters were evaluated using 5-fold cross validation, and the model that achieved the highest ROC-AUC score was selected. These hyperparameters were then used to train the final model on the entirety of the training set. After hyperparameter tuning, we found that using 233 estimators, a maximum depth of 31, a learning rate of 6.1e-2, and a minimum child weight of 6.17 maximized model performance. We evaluated our model with Receiver Operating Characteristic (ROC) AUC and Precision-Recall (PR) AUC on the test set (representing 20% of the data).

We annotated each position in the genome with our set of features and used the tuned XGBoost model to make a functionality prediction at each genomic position to score the entire genome. Of note, the model is trained to assess the functionality and essentiality of regulatory regions and should therefore be interpreted as such even in protein coding regions.

Validation Sets. The generalization of the model was assessed on three independent sets of variants. The non-coding pathogenic sets included 425 manually curated variants associated with Mendelian traits[23], 354 and 245 new HGMD and ClinVar variants[28] mapping outside/inside non-coding RNA genes (HGMD 2017_R2 and ClinVar January 2018). The control genomic training sets consisted respectively of 1,863, 2,604 and 1,876 variants in the gnomAD dataset at an allelic frequency > 1% and matched to the pathogenic sets for the distance to the nearest splice sites and genomic elements as explained above.

Reference build, Annotation and Genomic Element Categorization. All input features and the model were mapped to the human reference build hg19. To investigate the element distribution, we built an annotation track that combined annotations from GenCode (v.27 mapped to GRCh37) and ENCODE (annotated features and multicell regulatory elements, Ensembl v91 Regulatory Build) and used a prioritization scheme to assign each genomic position a single annotation category (described in [23]). For **Figure 3** and **Suppl. Figure S3**, *Intron* refers to intronic regions (from protein coding or non-coding genes), *ncRNA* refers to exonic regions of non-coding RNAs, *Cis Regulatory* encompasses promoters and untranslated regions (UTRs), *Enhancers and Others* encompasses promoter flanking regions, enhancers, open chromatin, CTCF and other transcription factor binding sites, *Intergenic* refers to unannotated regions and *Histone marks* encompasses H3K9me3 and/or H3K27me3 as well as other histone marks combinations.

ncER score. For computing power purposes and to minimize the signal-to-noise ratio, the ncER score were averaged over 10bp bins and then expressed as percentiles genome-wide. The binned/percentile scores are provided at https://github.com/TelentiLab and can be accessed directly at OMNI (https://omni-variants.herokuapp.com/). Intersection of ncER score with other datasets was performed using bedops utility (v2.4.30)[50].

External Datasets. The CRISPRi coordinates and scores used for analyses displayed in **Figures 2A** and **2B** and **Suppl. Figure S5** were obtained from Fulco et al.[31] (http://science.sciencemag.org/highwire/filestream/686019/field_highwire_adjunct_files/2/aag2445_Table_S2.xlsx). As recommended in their paper, the CRISPRi scores were smoothed over 20 subsequent pairs. . Only regions that were assessed by at least 20 different pairs and <1kb long (to prevent size biases) were retained for analysis, resulting in a total of N=77,368 remaining probed pairs.

CREST-seq peaks coordinates used for analyses displayed in **Figure 2C** and **Suppl. Figure S6** were obtained from Diao et al.[32] (https://media.nature.com/original/nature-assets/nmeth/journal/v14/n6/extref/nmeth.4264-S7.xlsx). Statistical enrichment of the 11,570 tested sgRNA pairs used for analyses in **Suppl. Figure S7** were also obtained from [32] (https://media.nature.com/original/nature-assets/nmeth/journal/v14/n6/extref/nmeth.4264-S5.xlsx). The locus used for random extraction of same size regions was chr6:30132133-32138339. The matched size random extraction (both in the same locus and genome-wide) was performed 100 times.

Cis regulatory transmitted deletions used for analyses displayed in **Figures 3A** and **B** and **Suppl. Figures S8-10** were obtained from Brandler et al.[34] (http://science.sciencemag.org/highwire/filestream/708877/field_highwire_adjunct_files/9/aan2261_TableS7.xlsx, Replication CR Trans sheet). The matched size genome-wide random extraction was performed 100 times.

The nine mouse enhancer data used for analyses displayed in **Figure 3C** and **Suppl. Figure S11** were obtained from Will et al.[37] (https://media.nature.com/original/nature-assets/ng/journal/v49/n10/extref/ng.3939-S1.pdf, Supplementary Table S4). The mouse coordinates were mapped to human using CrossMap (v.0.2.5. http://crossmap.sourceforge.net/) using the mm9ToHg19.over.chain.gz chain. When the mouse enhancers were mapped discontinuously to the human genome, the left most and right most coordinates in the human genome were used as start and end. The locus used for random extraction of same size regions was chr2: 219940039-220025587. The matched size random extraction (both in the same locus and genome-wide) was performed 100 times.

Statistics. Statistical analyses and plotting were performed with R v3.4.3 (https://www.R-project.org/), notably using the package ggplot2 (http:// ggplot2.org/). Data mining was performed using Python (v.2.7.11). The performance predictors in **Figures 1** and **2, Suppl. Figures 2,5** and **7**, **Suppl. Tables S2** and **S3** were assessed as follows: sensitivity or true positive rate or recall is (TP/(TP+FN)) * 100, specificity is (TN/(TN+FP)) * 100, false positive rate is (FP/(FP + TN)) * 100, accuracy is ((TP+TN)/(TP+TN+FP+FN))*100, positive predictive value or precision is (TP/(TP+FP))*100 and negative predictive value is (TN/(TN+FN))*100. Where TP is a true positive, TN is a true negative, FP is a false positive and FN is a false negative.

Acknowledgments. We thank J. Fellay for comments on the manuscript. Work of A.Te. is supported by the Qualcomm Foundation and the NIH Center for Translational Science Award (CTSA, grant number UL1TR002550).

Author contributions. Conception and design of the study: J.d.I, A.Te. Performed the analyses: A.W., J.d.I. Built the browser and code repositories: L.Y. Contributed methods and analytical strategies: D.H., A.To., B.R. Wrote the manuscript: J.d.I., A.Te.

Competing financial interests. None.

References

1. Lee, H. *et al.* Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**, 1880-7 (2014).
2. Chong, J.X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**, 199-215 (2015).
3. Thevenon, J. *et al.* Diagnostic odyssey in severe neurodevelopmental disorders: toward clinical whole-exome sequencing as a first-line diagnostic test. *Clin Genet* **89**, 700-7 (2016).
4. Iglesias, A. *et al.* The usefulness of whole-exome sequencing in routine clinical practice. *Genet Med* **16**, 922-31 (2014).
5. Pena, L.D.M. *et al.* Looking beyond the exome: a phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases. *Genet Med* **20**, 464-469 (2018).
6. Short, P.J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611-616 (2018).
7. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-10 (2015).
8. Bartha, I., di Iulio, J., Venter, J.C. & Telenti, A. Human gene essentiality. *Nat Rev Genet* **19**, 51-62 (2018).
9. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
10. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. *https://www.biorxiv.org/content/early/2017/09/27/193136* (2017).

11. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-8 (2014).

12. Torkamani, A. & Schork, N.J. Predicting functional regulatory polymorphisms. *Bioinformatics* **24**, 1787-92 (2008).

13. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

14. 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**, 214-20 (2016).

15. 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).

16. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).

17. Shihab, H.A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536-43 (2015).

18. 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).

19. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-4 (2015).

20. Gussow, A.B. *et al.* Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. *PLoS One* **12**, e0181604 (2017).

21. 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).

22. Rokach, L. Ensemble-based classifiers. *Artif Intell Rev* **33**, 1-39 (2010).

23. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat Genet* **50**, 333-337 (2018).

24. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-4 (2013).

25. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).

26. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074-7 (2013).

27. Landrum, M.J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-8 (2016).

28. Stenson, P.D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665-677 (2017).

29. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet* **11**, e1005492 (2015).

30. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

31. Fulco, C.P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).

32. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).

33. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).

34. Brandler, W.M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327-331 (2018).

35. Klopocki, E. *et al.* Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am J Hum Genet* **88**, 70-5 (2011).

36. 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**, 902-6 (2015).

37. Will, A.J. *et al.* Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat Genet* **49**, 1539-1545 (2017).

38. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239-243 (2018).

39. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).

40. Rehm, H.L. *et al.* ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**, 2235-42 (2015).

41. Berg, J.S. *et al.* An informatics approach to analyzing the incidentalome. *Genet Med* **15**, 36-44 (2013).

42. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-9 (2008).

43. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-7 (2005).

44. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).

45. Schmitt, A.D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059 (2016).

46. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

47. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**, 141-149 (2018).

48. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L.A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92 (2007).

49. Hon, C.C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199-204 (2017).

50. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-20 (2012).

Figure 1. Ensemble learning for the prediction of essential regulatory domains.

Performance ROC-AUC (**panel A**) and PR-AUC (**panel B**) on the test set (N=231 non-coding pathogenic and N=1,607 control variants) of ncER model (in grey) compared to previously published scores. The color codes for each metric is shown in the legend. The various input features to ncER (feature importance, the features that have the most effect in the model) are shown in **panel C**. Green, new essentiality features; Blue, published scores; Orange, regulatory/functional screen features; Red, 3D chromatin structure features. Performance ROC-AUC (**panel D**) and PR-AUC (**panel E**) of model trained only with published deleteriousness metrics (blue), only with new features (orange) and with both new features and published metrics (green, ncER).

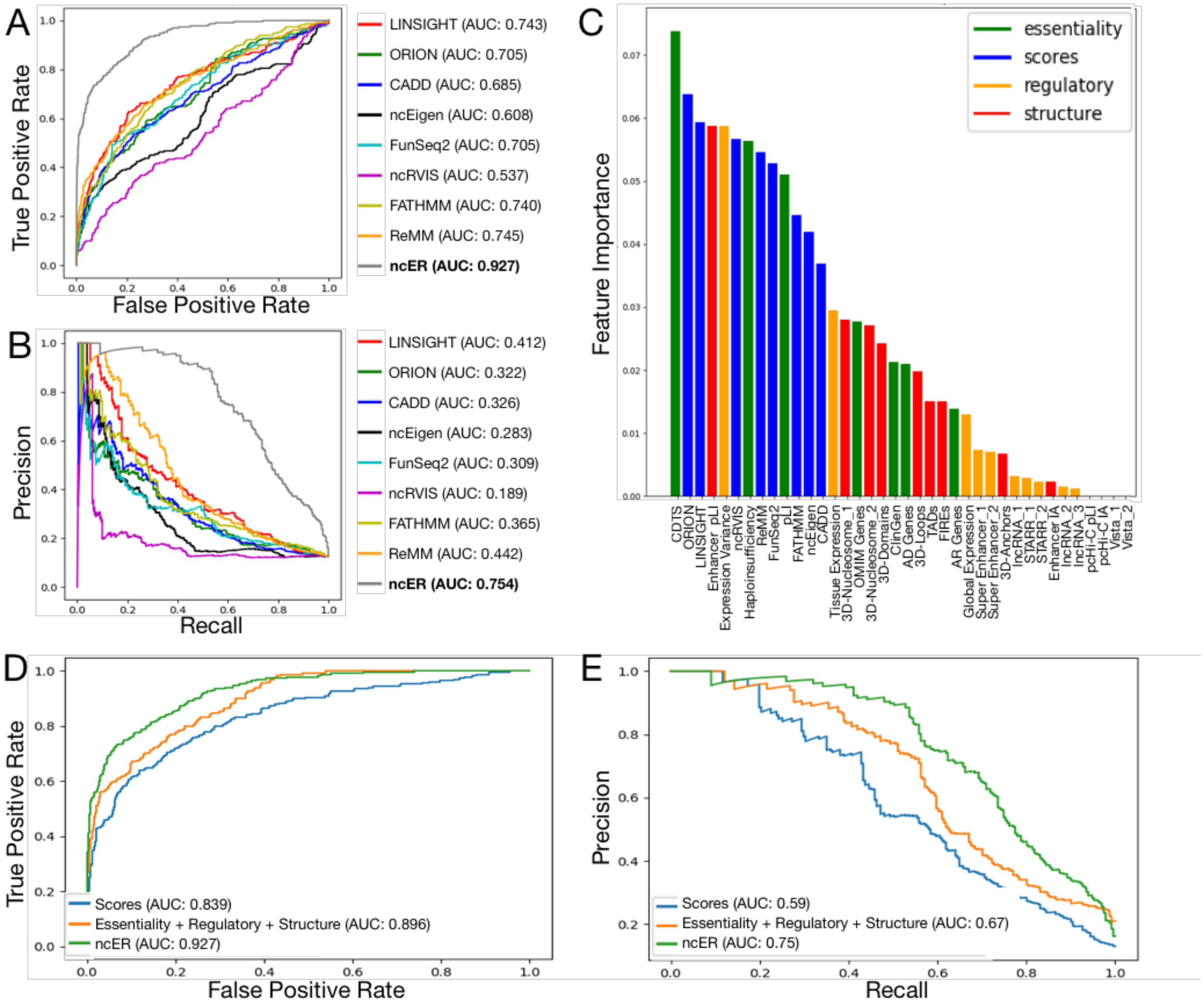


Figure 2. Comparison of experimental functional assays with *in silico* predictions of essentiality.

Panel A. CRISPRi effect on cell viability (77,368 sgRNA probes pairs) from Fulco et al.[31] and the corresponding maximum ncER score accross the *GATA1* and *MYC* loci. Accuracy at four ncER thresholds is shown in yellow, orange, red and dark-red respectively for the 95th, 99th, 99.5th and 99.9th percentiles. **Panel B.** Distribution of maximum ncER at different bins of cell viability (0 to less than -3 log2 fold change). P values were computed with independent 2-group Man-Whitney Unpaired Test. **Panel C.** Upper display represents the experimental locus and the identified CRESTseq peaks (n=45, green), from Diao et al.[32]. Lower display presents the fraction of regions with at least one essential bin. Essential bins are defined by four different ncER percentile thresholds. CRESTseq peaks are shown in red (N=45), random matched sized *in silico* regions extracted from the same locus (“random same locus, N=4,500) in dark grey and random matched sized *in silico* regions extracted genome-wide (“random GW”, N=4,500), in light grey. P values were computed with Fisher Exact Test. Panel C pictogram is adapted from[32]. FC, fold change. CRESTseq, *cis*-regulatory elements by tiling-deletion and sequencing. GW, genome-wide.

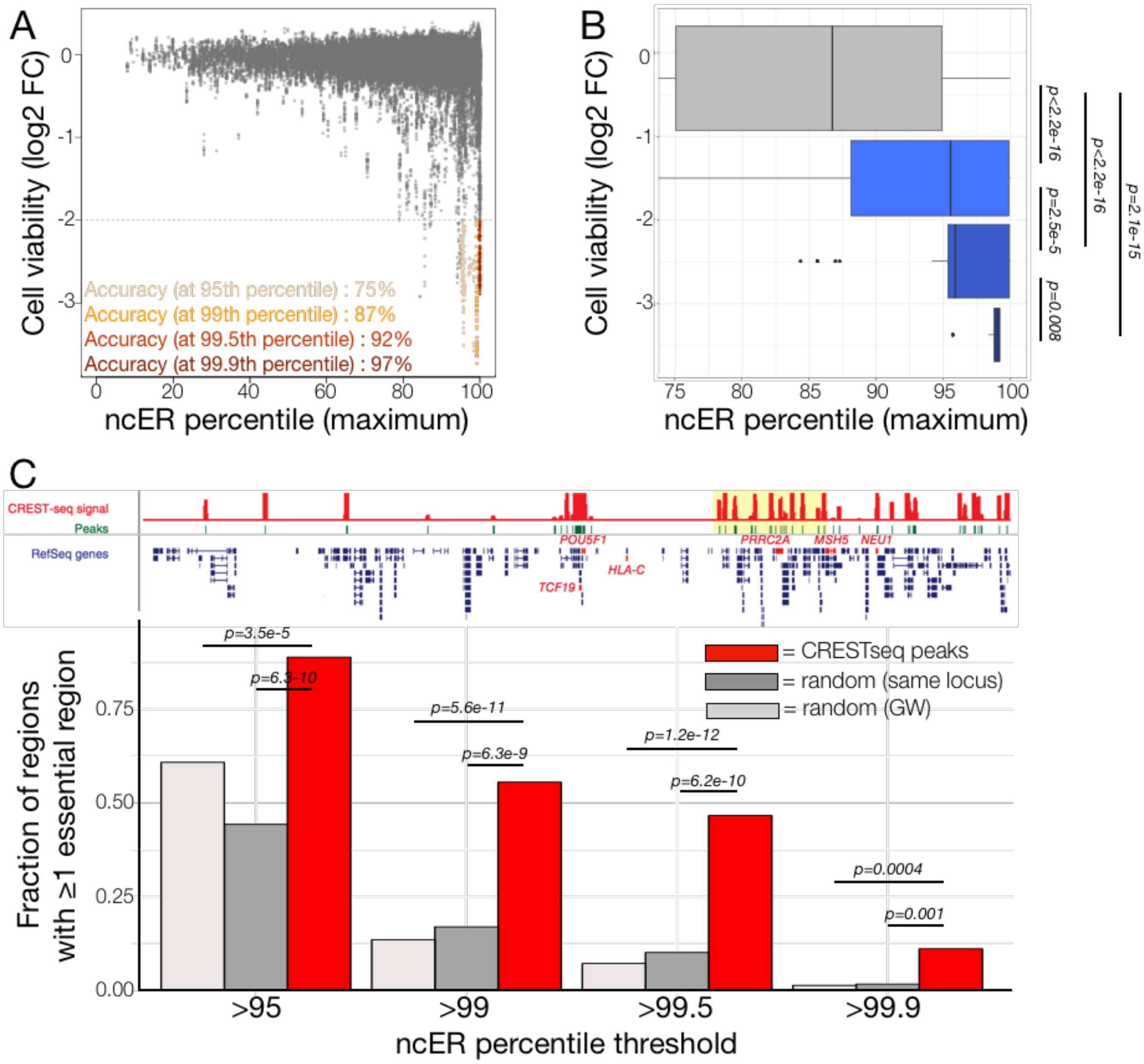
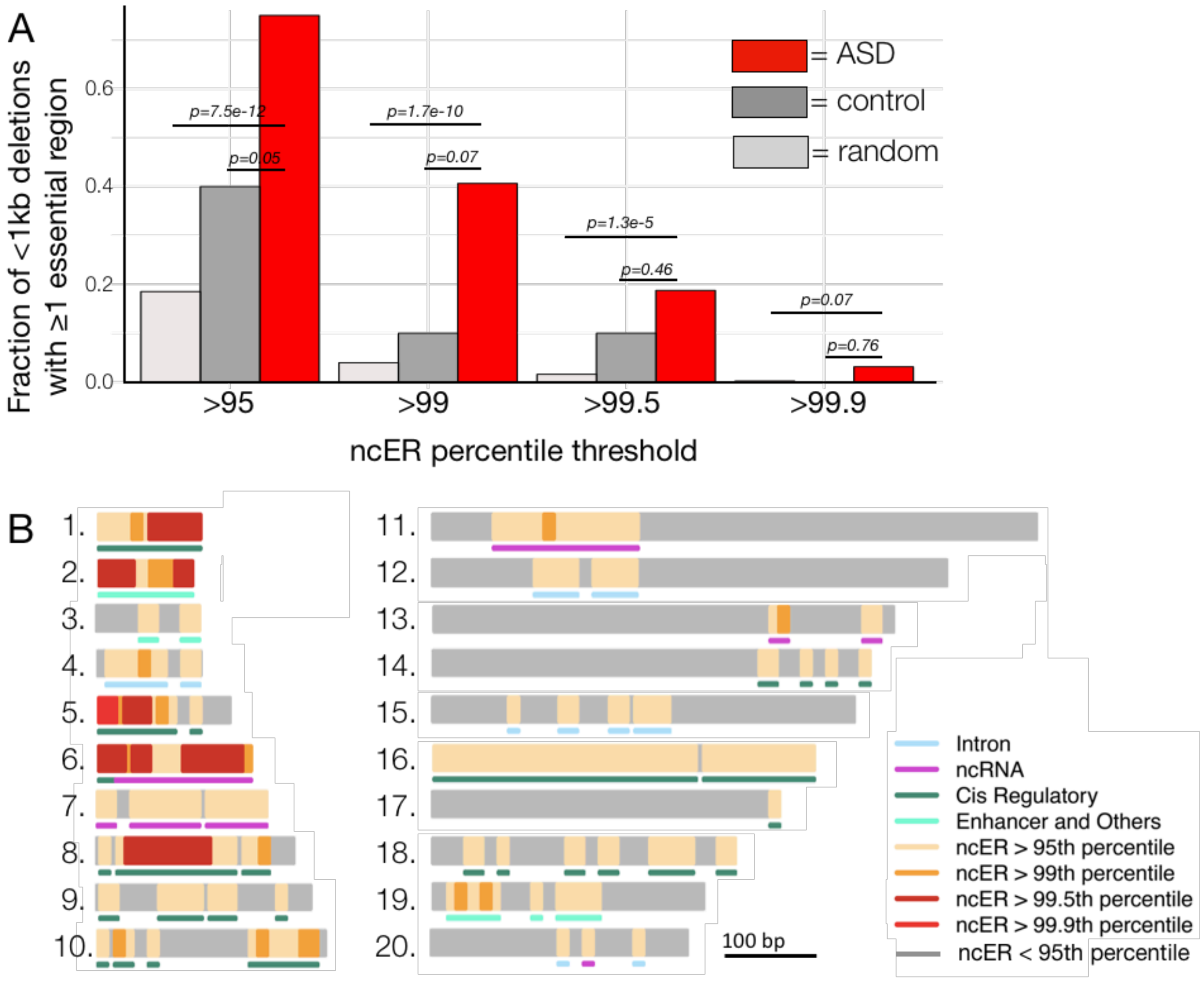
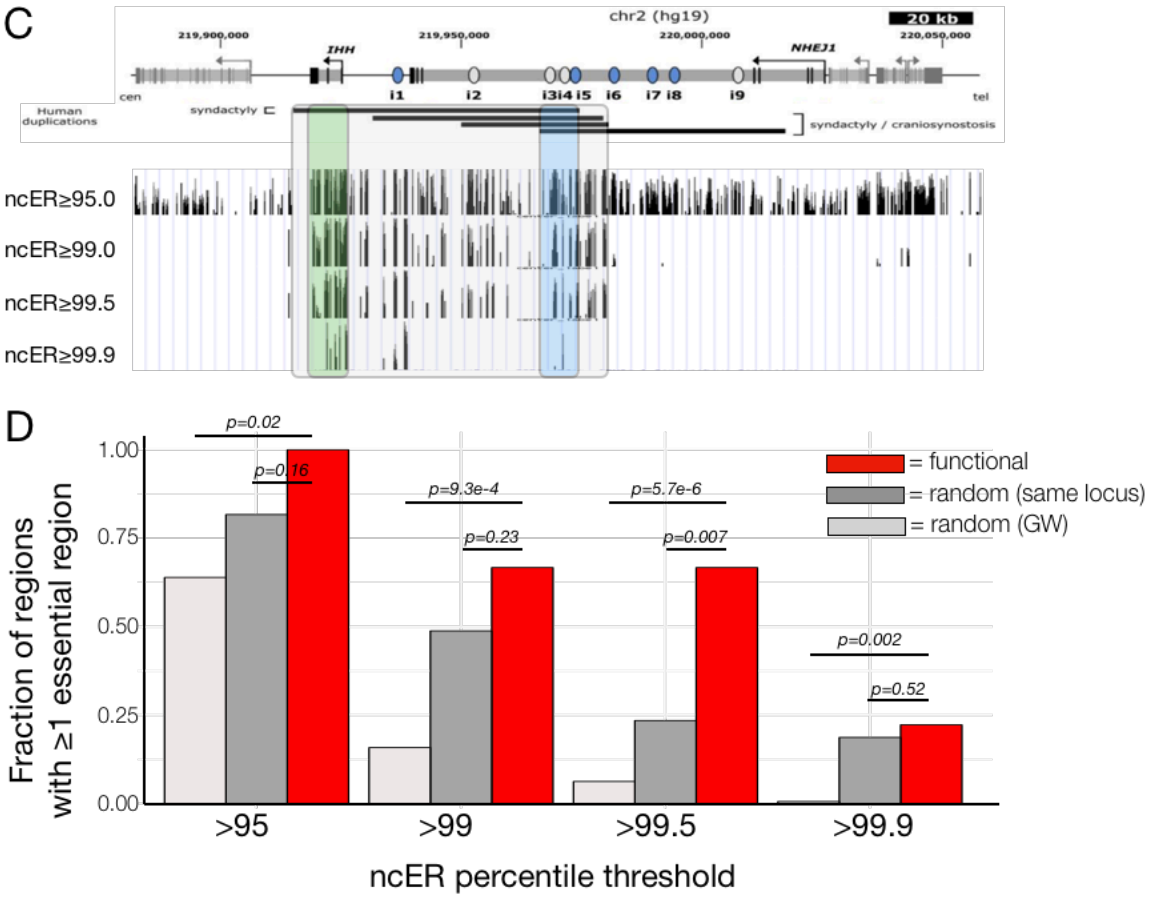


Figure 3. Mapping of essential regulatory domains in disease models.

Panel A. Fraction of <1kb deletions with at least one essential bin. Essential bins are defined at four different ncER percentile thresholds. Autism and ASD deletions are shown in red (ASD, N=32), control deletions in dark grey (control, N=10) and random size-matched *in silico* deletions extracted genome-wide in light grey (random, N=4,200). See **Suppl. Figure 8** for distribution of other sizes of deletions. **Panel B**. Schematic illustration of 20 unique deletions (grey bars) identified in Autism and ASD probands that harbor essential regions (highlighted in yellow, orange, red and dark red, based on the ncER thresholds). The corresponding genomic elements present at the essential regions are displayed under the deletions. Introns are shown in blue, ncRNA in magenta, *cis* regulatory in dark green and enhancers and others regulatory elements in light green (See **Methods** for categorization of genomic element classes). **Panel C**. The upper panel illustrates the human *IHH* genomic locus associated with developmental defects including craniosynostosis and synpolydactyly[35,36]. It harbors the 9 enhancers identified in mice (from Will et al.[37]). Lower panel, UCSC genome browser view of the essential region in the locus. Essential bins are shown at four ncER percentile thresholds. The grey box inset highlights the region of essentiality across human pathogenic duplications, the blue box the essentiality at the maximal overlap of genomic lesions in humans, and the green box that includes the *IHH* region present in duplications causing syndactyly Leuken type engineered in Will et al.[37]. Panel C pictogram is adapted from[37]. **Panel D.** Fraction of regions with at least one essential bin. Essential bins are defined by four different ncER percentile thresholds. Mouse to human mapped enhancers are shown in red ("functional", N=9), random size-matched *in silico* deletions extracted from the same locus ("random same locus", N=900) in dark grey and random size-matched *in silico* deletions extracted genome-wide ("random GW", N=900) in light grey. GW, genome-wide.


A
Fraction of <1kb deletions with ≥1 essential region
0.0
0.2
0.4
0.6
p=7.5e-12
p=0.05
p=1.7e-10
p=0.07
p=1.3e-5
p=0.46
p=0.07
p=0.76
🟥 = ASD
▨ = control
□ = random
>95
>99
>99.5
>99.9
ncER percentile threshold
B
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.
100 bp
Intron
ncRNA
Cis Regulatory
Enhancer and Others
ncER > 95th percentile
ncER > 99th percentile
ncER > 99.5th percentile
ncER > 99.9th percentile
ncER < 95th percentile


C
chr2 (hg19)
20 kb
219,900,000
219,950,000
220,000,000
220,050,000
IHH
NHEJ1
cen
tel
i1
i2
i3i4 i5
i6
i7 i8
i9
Human duplications
syndactyly
syndactyly / craniosynostosis
ncER≥95.0
ncER≥99.0
ncER≥99.5
ncER≥99.9
D
Fraction of regions with ≥1 essential region
1.00
0.75
0.50
0.25
0.00
p=0.02
p=0.16
p=9.3e-4
p=0.23
p=5.7e-6
p=0.007
p=0.002
p=0.52
🟥 = functional
▨ = random (same locus)
□ = random (GW)
>95
>99
>99.5
>99.9
ncER percentile threshold

Supplementary materials

Identification of essential regulatory elements in the human genome

Alex Wells[1], David Heckerman[2], Ali Torkamani[3], Li Yin[3], Bing Ren[4], Amalio Telenti[3,5*], Julia di Iulio[3,5*]

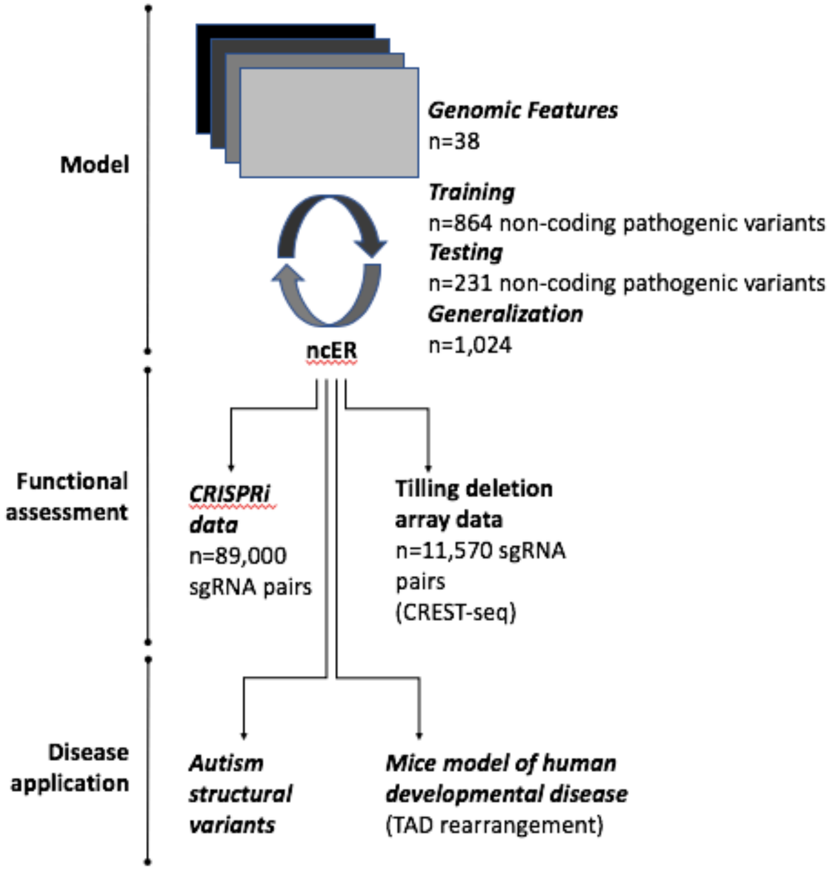
Figures

Tables

Suppl. Figure S1. Study design.

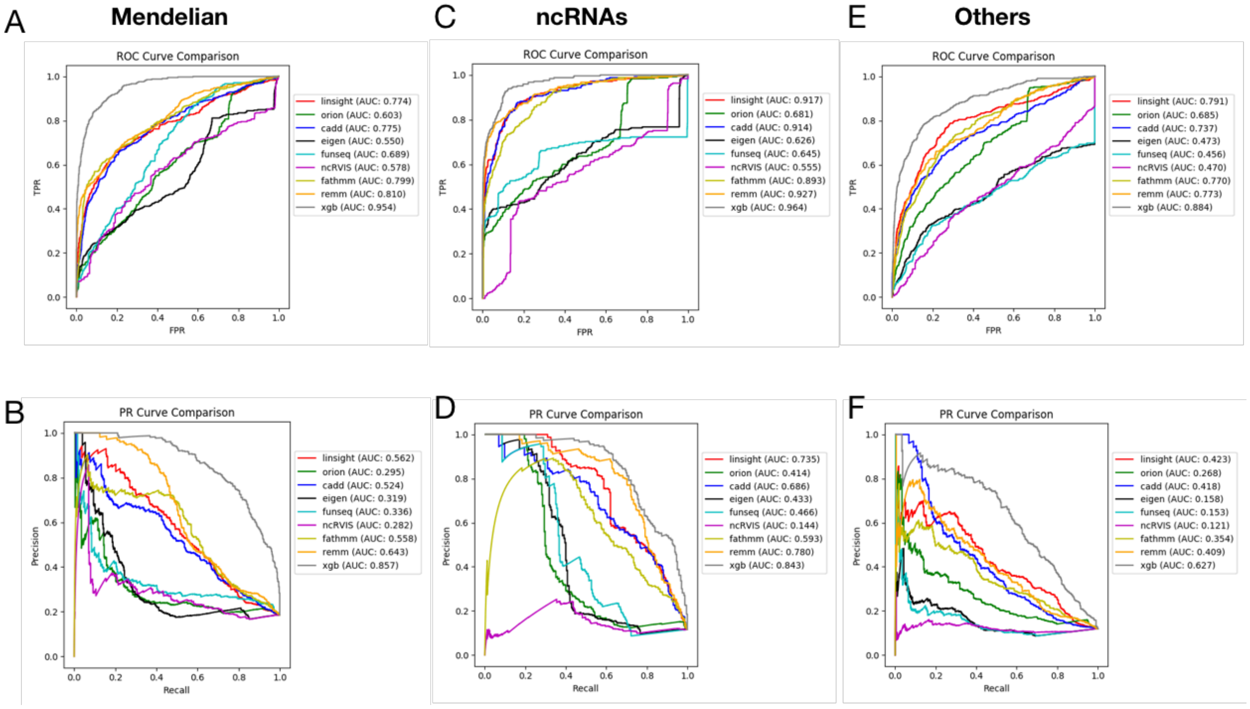
CREST-seq= *cis*-regulatory element scan by tiling-deletion and sequencing.

CRISPRi= clustered regularly interspaced short palindromic repeats (CRISPR) interference.



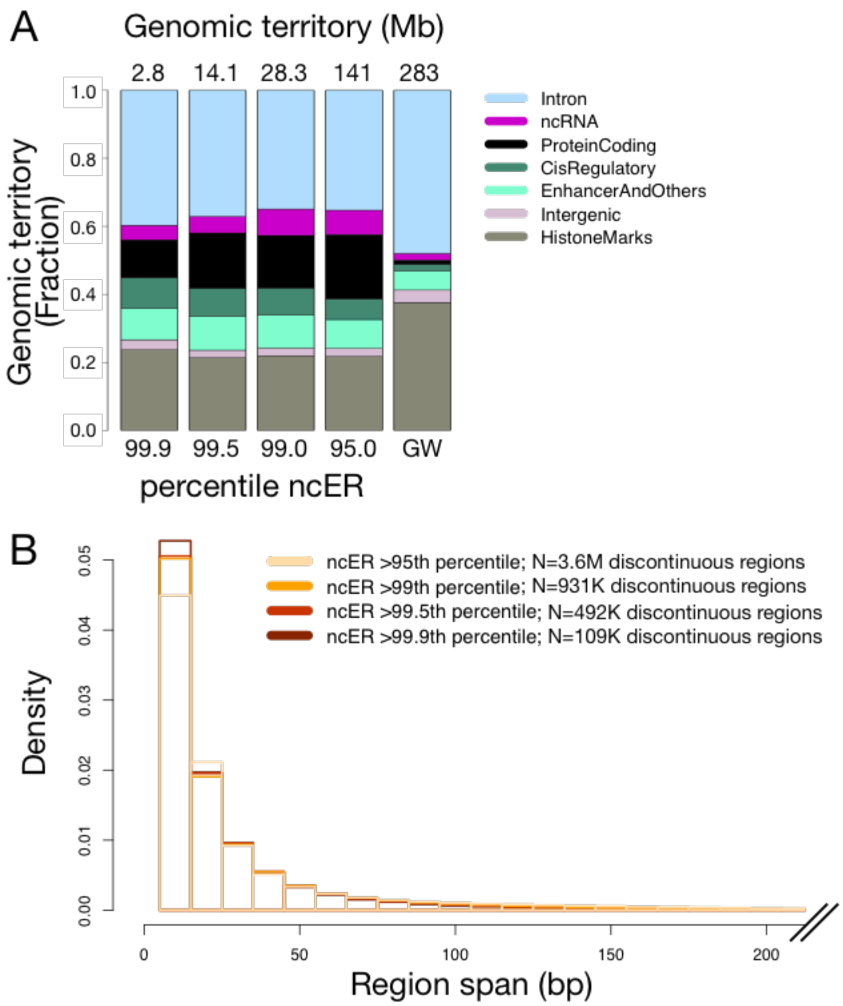
Suppl. Figure S2. ncER generalization to independent variant sets.

Performance ROC-AUC (**panel A**) and PR-AUC (**panel B**) for the independent set of manually curated non-coding Mendelian variants (N=425 pathogenic and N=1,863 control)[23]. Performance ROC-AUC (**panel C**) and PR-AUC (**panel D**) for the independent set of left out non-coding HGMD and ClinVar variants mapping to ncRNAs (N=245 pathogenic and N=1,876 control). Performance ROC-AUC (**panel E**) and PR-AUC (**panel F**) for the independent set of left out non-coding HGMD and ClinVar variants mapping outside of ncRNAs (Others; N=354 pathogenic and N=2,604 control). ncRNA, non-coding RNA. TPR, true positive rate. FPR, false positive rate. Xgb, XGBoost model or ncER.



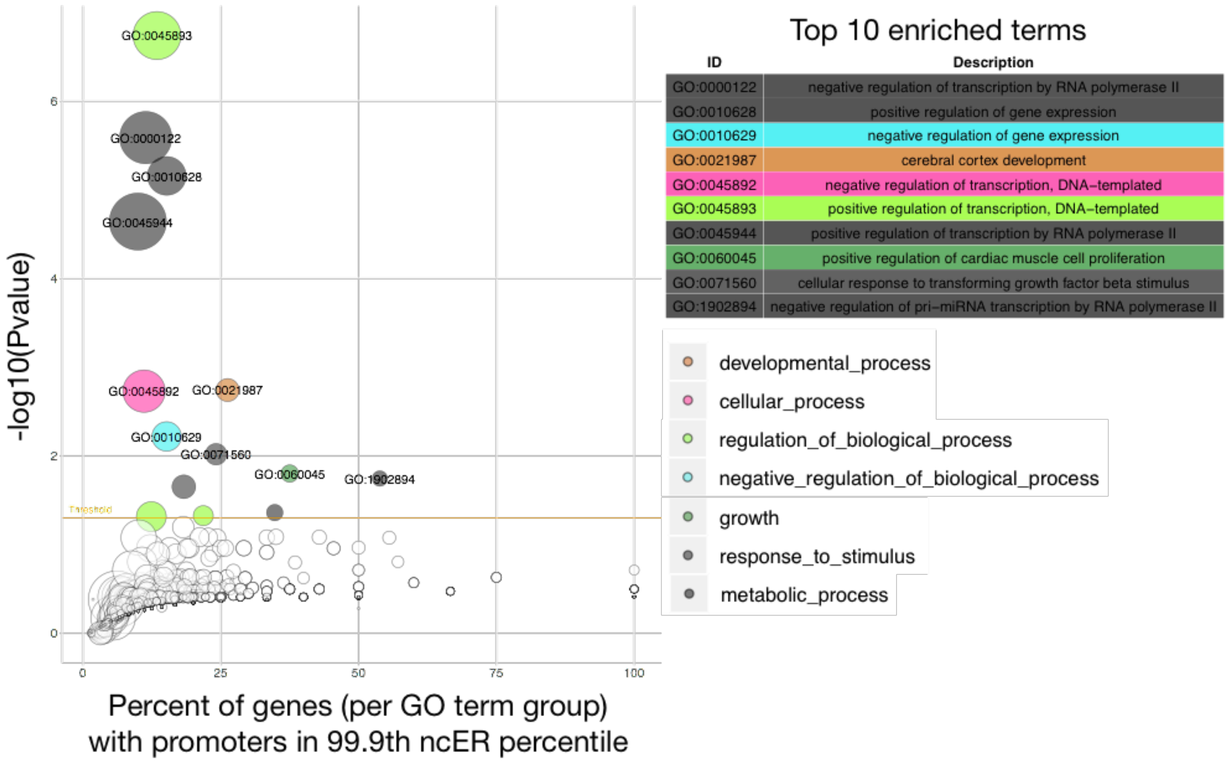
Suppl. Figure S3. Essential regions distribution in the genome.

Panel A. Bar plot displaying the cumulative territory covered by each element family at different percentile threshold (indicated at the bottom of the bars). The total genomic territory is displayed at the top of the bars. The percentiles are based on the rank of ncER values. The elements appear in the bar plot in the same order as in the legend. Genomic elements grouping is detailed in **Methods**. **Panel B**. Size distribution of essential regions defined at different ncER percentile thresholds (indicated by the color code in the top-right of the figure). GW, genome-wide.



Suppl. Figure S4. The GO terms enrichment for biological processes in essential regions.

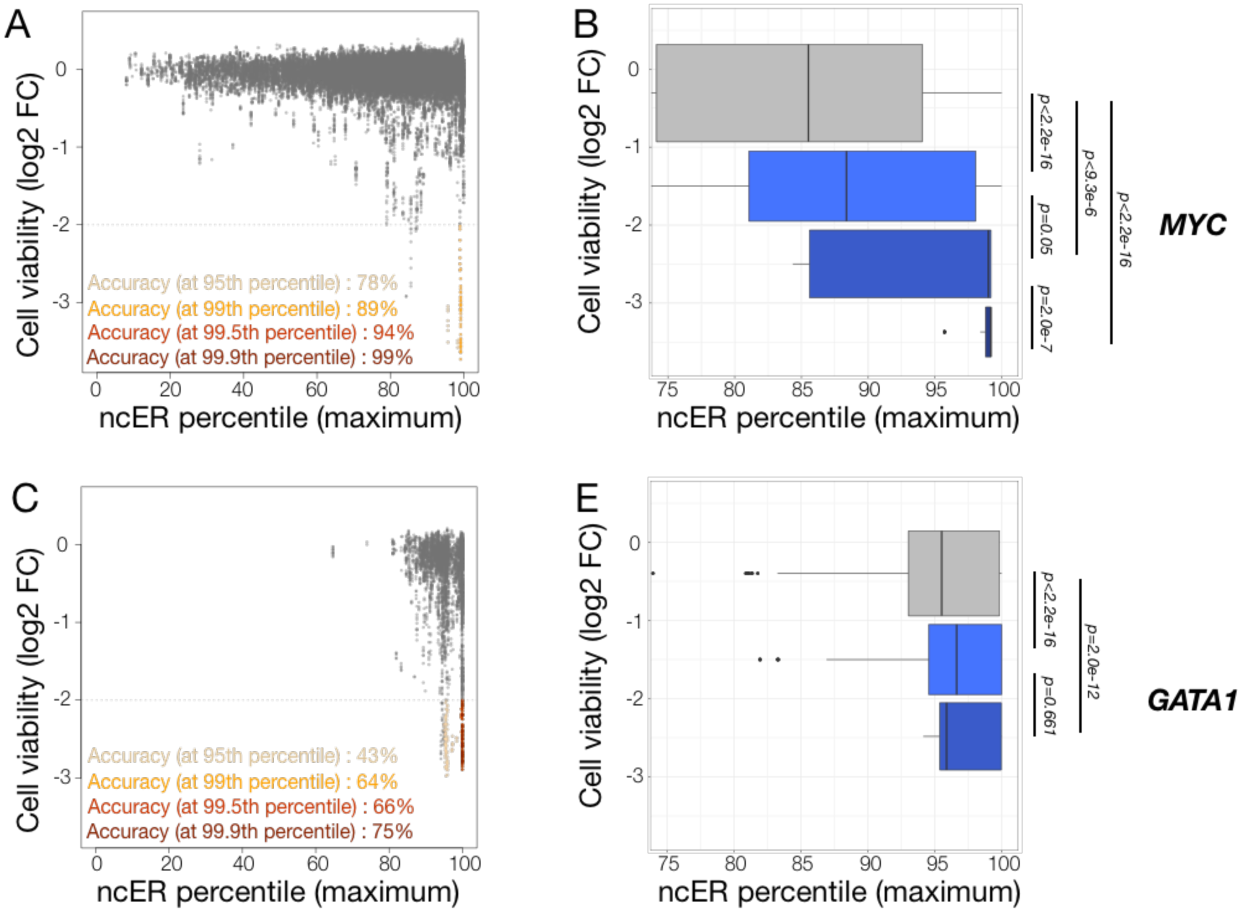
The bubble plot represents the significance of the enrichment for a given term (y axis, minus log10(pvalue)) versus the percentage of genes associated with a given biological process that were present in the set of genes with at least one essential promoter bin (x axis). Each circle in the plot represents one biological process. The size of the circles is proportional to the number of genes in the specific GO term class. Only the significant terms are colored. The circles and the rows in the associated table are colored according to the highest ranking hierarchical term (to facilitate pathway and redundant information detection). When multiple terms could be the highest hierarchical ancestor, the coloring was randomly selected among the multiple possibilities. The associated table provides the name of the top ranked terms. Promoter regions were defined as the 600bp upstream the transcription start site. A modified version of the GOBubble function from the GOplot R package (http://wencke.github.io/) was used to generate this figure.



ID	Description
GO:0000122	negative regulation of transcription by RNA polymerase II
GO:0010628	positive regulation of gene expression
GO:0010629	negative regulation of gene expression
GO:0021987	cerebral cortex development
GO:0045892	negative regulation of transcription, DNA–templated
GO:0045893	positive regulation of transcription, DNA–templated
GO:0045944	positive regulation of transcription by RNA polymerase II
GO:0060045	positive regulation of cardiac muscle cell proliferation
GO:0071560	cellular response to transforming growth factor beta stimulus
GO:1902894	negative regulation of pri–miRNA transcription by RNA polymerase II

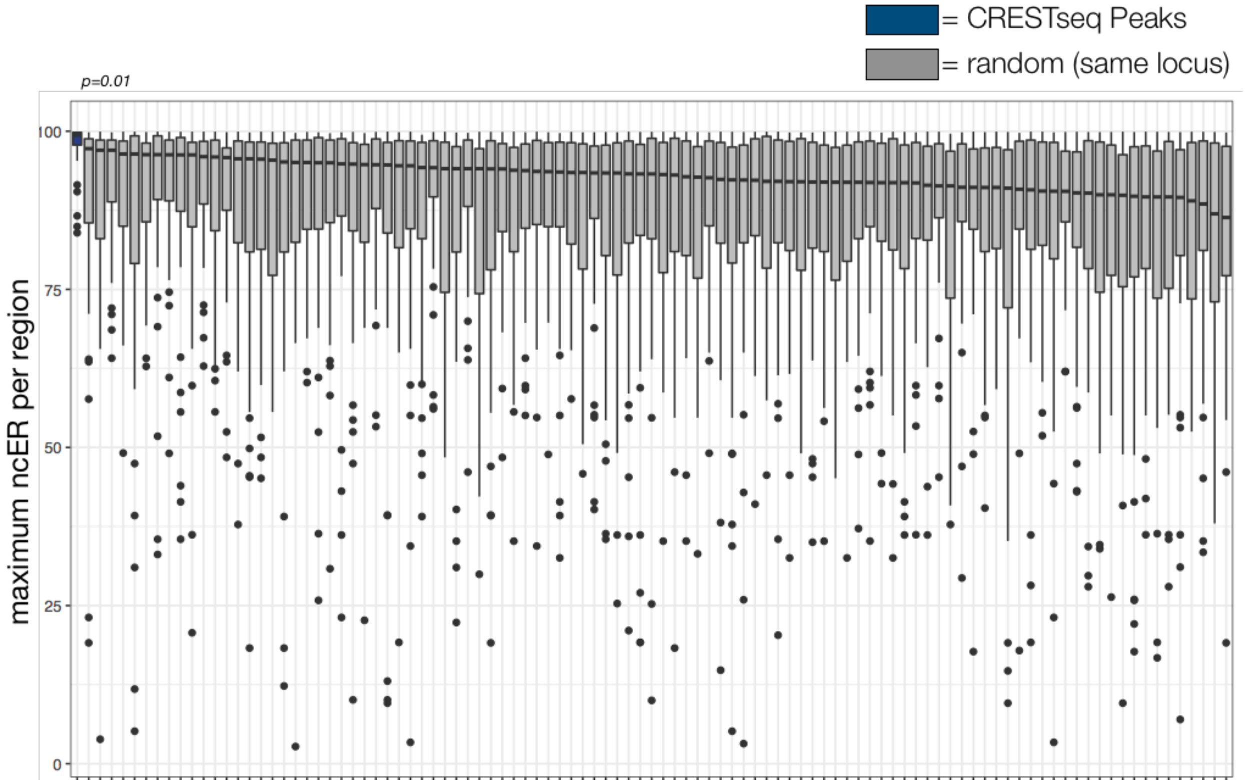
Suppl. Figure S5. Comparison of experimental CRISPRi functional assays with *in silico* predictions of essentiality.

Panel A. CRISPRi effect on cell viability (71,404 sgRNA probe pairs targeting the *MYC* locus) and the corresponding maximum ncER score within the tested region. Accuracy at four ncER thresholds is shown in yellow, orange, red and dark-red respectively for the 95th, 99th, 99.5th and 99.9th ncER percentiles. **Panel B.** Distribution of maximum ncER at different bins of cell viability (0 to lower than -3 log2 fold change). P values were computed with independent 2-group Man-Whitney Unpaired Test. **Panel A.** CRISPRi effect on cell viability (5,856 sgRNA probe pairs targeting the *GATA1* locus) and the corresponding maximum ncER score within the tested region. Accuracy at four ncER thresholds is shown in yellow, orange, red and dark-red respectively for the 95th, 99th, 99.5th and 99.9th ncER percentiles. **Panel B.** Respective distribution of maximum ncER at different bins of cell viability (0 to lower than -3 log2 fold change). P values were computed with independent 2-group Man-Whitney Unpaired Test.



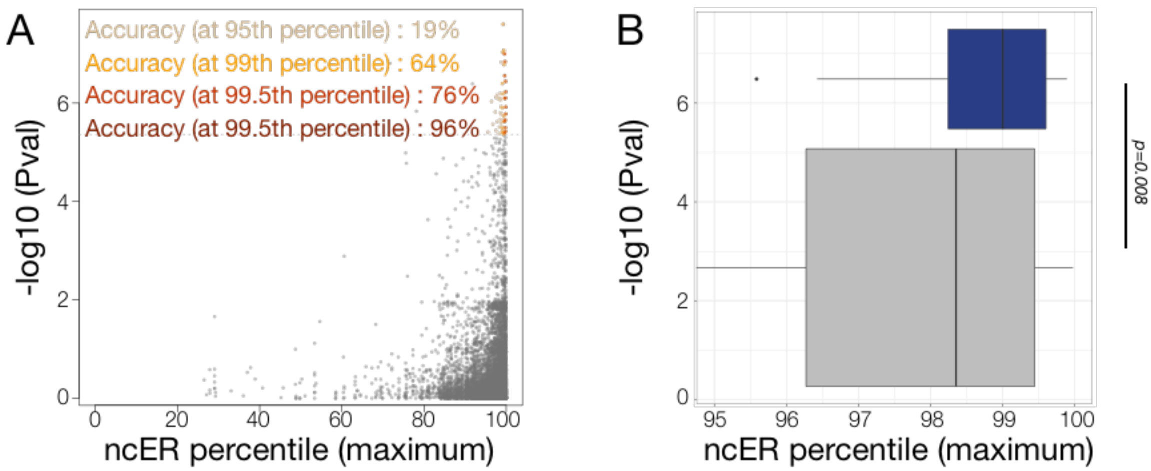
Suppl. Figure S6. CREST-seq peaks enrichment in essential regions.

CREST-seq peaks (N=45, dark blue) display the highest ncER percentile distribution, compared to 100 permutations (grey), each containing 45 regions matched by size to the CREST-seq peaks and from the same genomic locus.



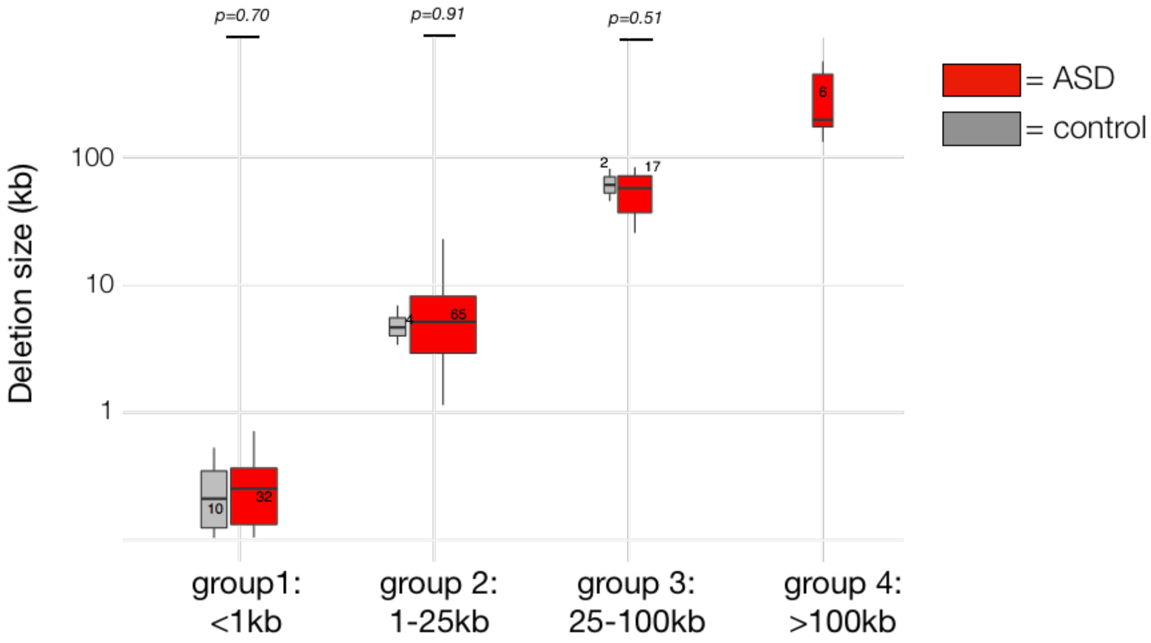
Suppl. Figure S7. Comparison of experimental CREST-seq functional assays with *in silico* predictions of essentiality.

Panel A. P values determined by Diao et al. [32] comparing *POU5F1* expression in targeted cells versus controls (11,570 sgRNA probe pairs) and the corresponding maximum ncER score within the tested region. Accuracy at four ncER thresholds is shown in yellow, orange, red and darkred respectively for the 95th, 99th, 99.5th and 99.9th ncER percentiles. **Panel B.** Distribution of maximum ncER at different bins of -log10(p value) (up to 5.36 (non-significant), above 5.36 which corresponds to -log10(0.05/11,570)). P values were computed with independent 2-group Man-Whitney Unpaired Test.



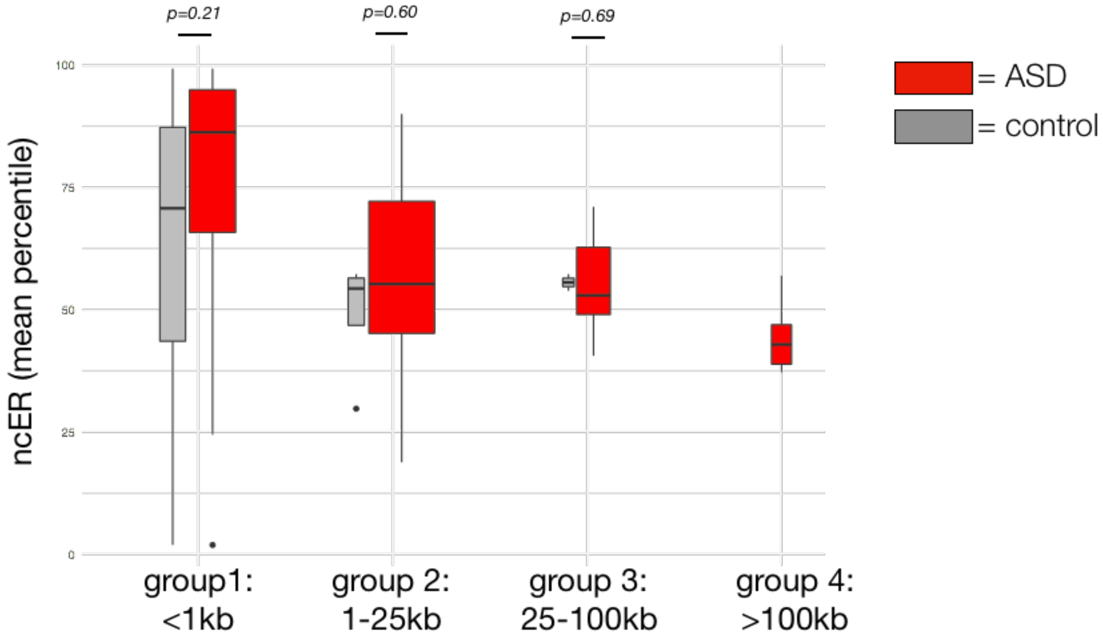
Suppl. Figure S8. Size distribution of Cis-regulatory transmitted deletions.

The transmitted cis-regulatory deletions were split into 4 groups of matched size (with no significant difference in the distribution of control and ASD deletion within the same group size). Deletions present in cases are shown in red, deletions present in controls are shown in grey. The y axis is on logarithmic scale. P values were computed with independent 2-group Man-Whitney Unpaired Test.



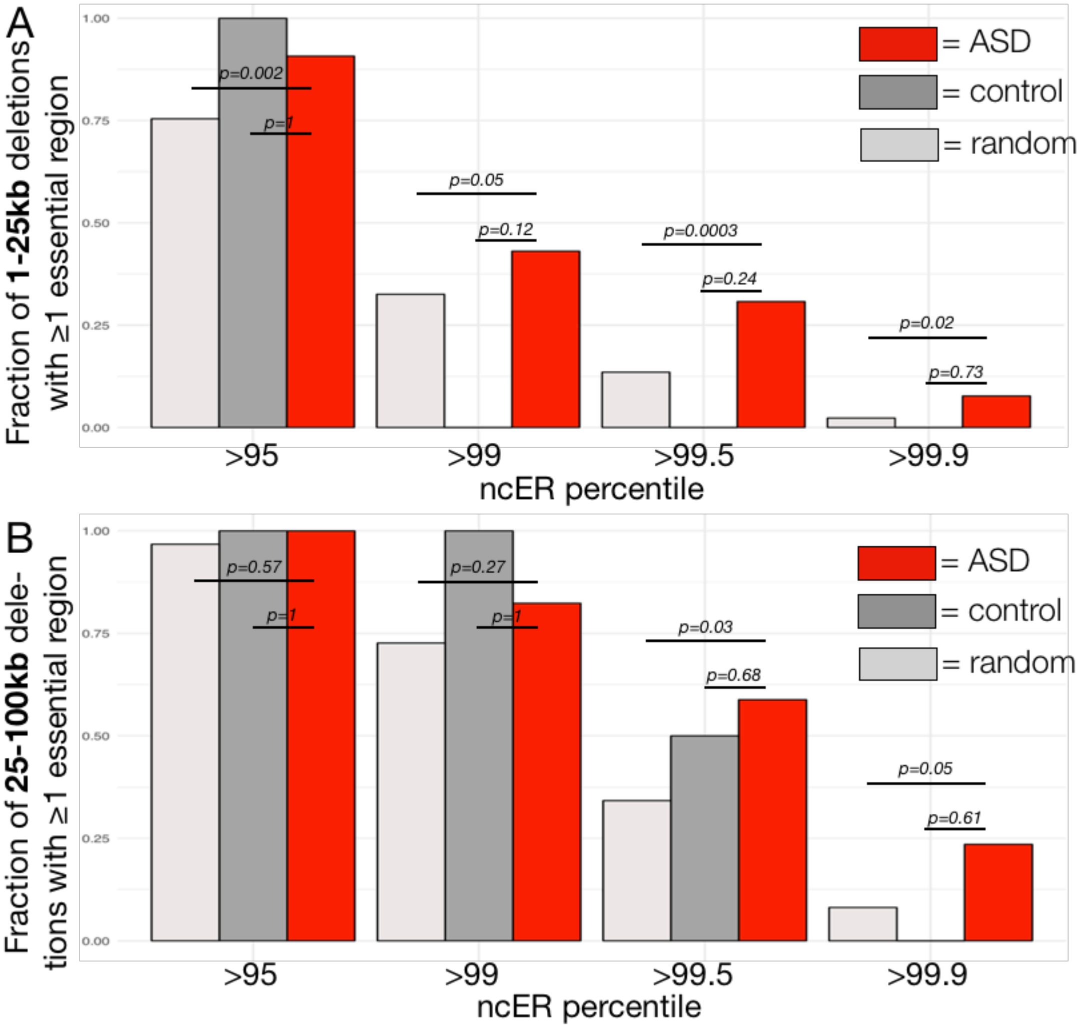
Suppl. Figure S9. ncER percentile distribution across transmitted deletion.

The <1kb cis-regulatory transmitted deletions tend to have an average ncER slightly higher in ASD than in controls, however the effect, if any, disappears for higher size deletions. The longer the deletion the more likely to approach ncER 50th percentile, the genome average. Deletions present in probands are shown in red, deletions present in controls are shown in grey. The boxplot width is proportional to the number of deletions per group (see **Suppl. Figure S8**). P values were computed with independent 2-group Man-Whitney Unpaired Test.



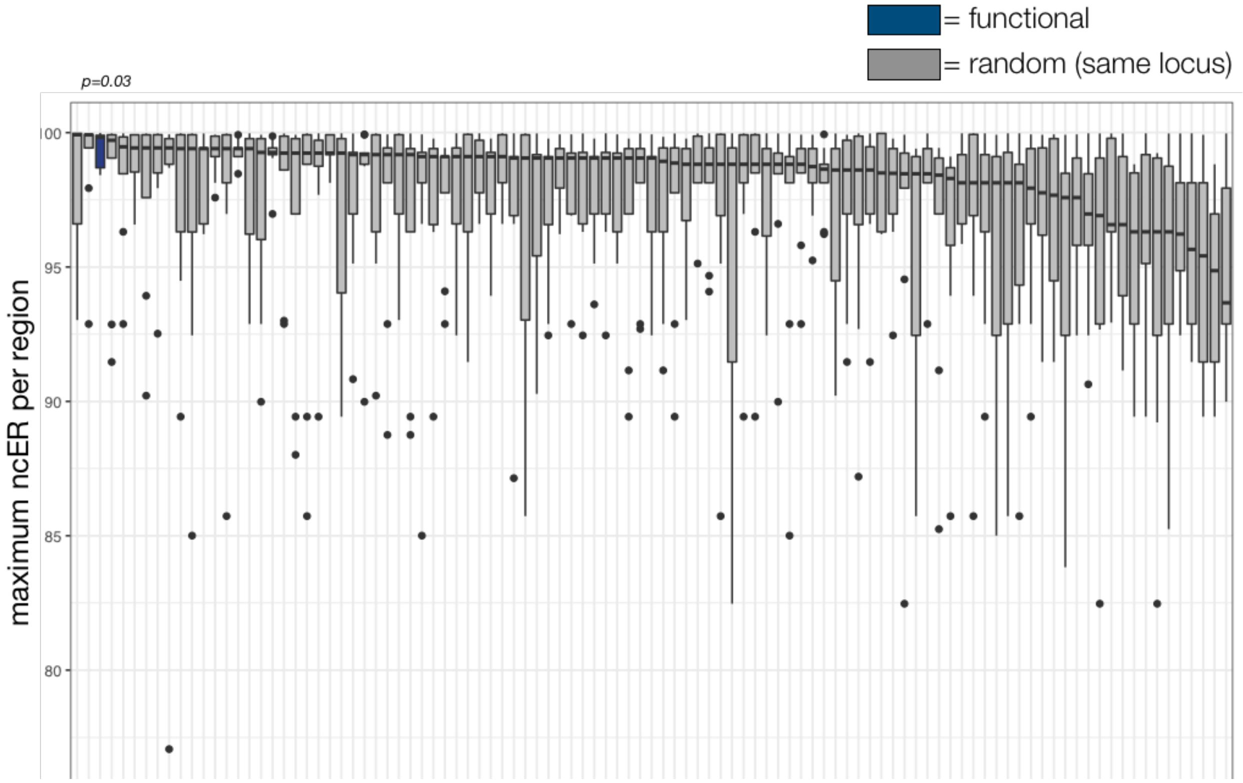
Suppl. Figure S10. Fraction of transmitted deletions with essential functional domains.

Panel A. Fraction of 1-25kb deletions with at least one essential bin. Essential bins are defined by four different ncER percentile thresholds. Autism and ASD deletions are shown in red ("ASD", N=65), control deletions in dark grey ("control", N=4) and random size-matched *in silico* deletions extracted genome-wide in light grey ("random", N=6,900). **Panel B.** Fraction of 25-100kb deletions with at least one essential bin. Essential bins are defined by four different ncER percentile thresholds. Autism and ASD deletions are shown in red ("ASD", N=17), control deletions in dark grey ("control", N=2) and random size-matched *in silico* deletions extracted genome-wide in light grey ("random", N=1,900).



Suppl. Figure S11. Enrichment in essential regions for mouse functional enhancers.

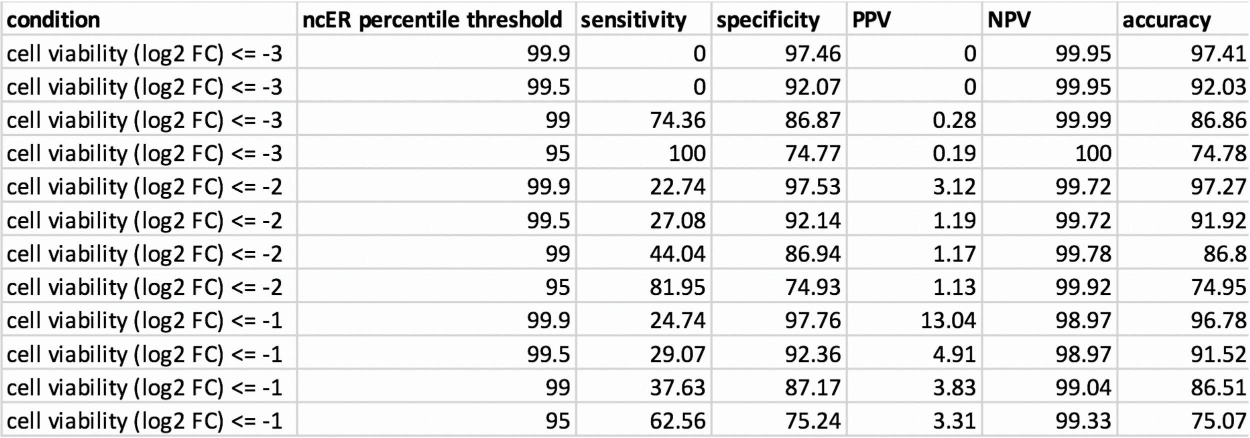
Functional enhancers (N=9, blue) are placed among the highest ncER percentile distribution compared to 100 permutations (grey), each containing 9 regions matched by size to the enhancers and issued from the same genomic locus.



Suppl. Table S1. Input feature description and accession links.

Provided as separate file.

Suppl. Table S2. Predictive performance and accuracy of ncER compared to CRISPRi functional assays.



condition	ncER percentile threshold	sensitivity	specificity	PPV	NPV	accuracy
cell viability (log2 FC) <= -3	99.9	0	97.46	0	99.95	97.41
cell viability (log2 FC) <= -3	99.5	0	92.07	0	99.95	92.03
cell viability (log2 FC) <= -3	99	74.36	86.87	0.28	99.99	86.86
cell viability (log2 FC) <= -3	95	100	74.77	0.19	100	74.78
cell viability (log2 FC) <= -2	99.9	22.74	97.53	3.12	99.72	97.27
cell viability (log2 FC) <= -2	99.5	27.08	92.14	1.19	99.72	91.92
cell viability (log2 FC) <= -2	99	44.04	86.94	1.17	99.78	86.8
cell viability (log2 FC) <= -2	95	81.95	74.93	1.13	99.92	74.95
cell viability (log2 FC) <= -1	99.9	24.74	97.76	13.04	98.97	96.78
cell viability (log2 FC) <= -1	99.5	29.07	92.36	4.91	98.97	91.52
cell viability (log2 FC) <= -1	99	37.63	87.17	3.83	99.04	86.51
cell viability (log2 FC) <= -1	95	62.56	75.24	3.31	99.33	75.07

Suppl. Table S3. Predictive performance and accuracy of ncER compared to CREST-seq functional assays.

condition	ncER percentile threshold	sensitivity	specificity	PPV	NPV	accuracy
-log10(pval)>=5.36	99.9	0.00	96.07	0.00	99.52	95.63
-log10(pval)>=5.36	99.5	26.42	76.46	0.51	99.56	76.23
-log10(pval)>=5.36	99	50.94	64.03	0.65	99.65	63.97
-log10(pval)>=5.36	95	92.45	19.02	0.52	99.82	19.36