Identification and characterization of constrained non-exonic bases lacking predictive epigenomic and transcription factor binding annotations

Olivera Grujic[1,2], Tanya N. Phung[3], Soo Bin Kwon[2,3], Adriana Arneson[2,3], Yuju Lee[1], Kirk E. Lohmueller[3,4,5], Jason Ernst[1,2,3,6,7,8*]

[1] Computer Science Department, University of California, Los Angeles, Los Angeles, California, USA.

[2] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California, USA.

[3] Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, California, USA.

[4] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, California, USA.

[5] Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, USA.

[6] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, Los Angeles, California, USA.

[7] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, USA.

[8] Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, USA.

*Correspondence: J.E. (jason.ernst@ucla.edu)

Abstract

Genome-wide maps of epigenomic marks and transcription factor binding provide cell type and condition specific information for annotating genomes and interpreting genetic variation. Predictions of evolutionarily constrained bases provide an orthogonal genomic annotation of potentially important bases in the genome. Evolutionary constrained non-exonic bases that are not effectively predicted from large-scale epigenomic and transcription factor binding data could suggest noteworthy gaps in the coverage of such data. To investigate this, we developed the Constrained Non-Exonic Predictor (CNEP), and applied it to the human genome using over ten thousand features defined from large-scale epigenomic and transcription factor binding data to score the evidence of each base being in a constrained non-exonic element from such data. We find that a large subset of constrained non-exonic bases is well predicted by CNEP, but another large subset is not and the predictive power for bases varies substantially with their ConsHMM conservation state annotations. Human genetic variation provided evidence to support that a set of called constrained non-exonic bases with low CNEP scores are under selection, but to a lesser extent than those with high scores. We analyzed the potential biological role of constrained non-exonic bases with low CNEP scores using regulatory sequence motifs, mouse epigenomic data, and additional prospectively considered human data. These analyses highlight how a subset of these bases may have specialized regulatory roles related to embryonic development, the brain, or response to stimuli not well annotated by commonly used compendia of epigenomic and transcription factor binding data.

Introduction

A large majority of genetic variation associated with common disease falls into non-exonic regions of the human genome(Hindorff et al., 2009). Genome-wide maps of histone modifications and variants, transcription factor (TF) binding, open chromatin, and chromatin state annotations have become important resources for interpreting and prioritizing potential phenotype associated genetic variants in the non-coding genome(Claussnitzer et al., 2015; ENCODE Project Consortium, 2012; Ernst et al., 2011; Maurano et al., 2012; Roadmap Epigenomics Consortium et al., 2015). However, these resources are specific to the condition and cell or tissue type in which the experiments underlying them were conducted. Downstream conclusions about genetic variants based on such data can thus be missed or biased due to the specific experiments that have or have not been conducted. It is therefore important to have an understanding of the extent to which large-scale compendia of epigenomic and TF binding data are capturing putatively important genome bases and the nature of putatively important bases not captured by them.

Evolutionarily constrained elements provide an orthogonal genome annotation, which does not depend on the specific cell or tissue types or the experimental mark chosen for mapping(Davydov et al., 2010; Garber et al., 2009; Lindblad-Toh et al., 2011; Siepel et al., 2005). Supporting the importance of these annotations, heritability analyses have suggested they are heavily enriched for disease associated variants(Finucane et al., 2015). Furthermore, annotations of evolutionarily constrained elements and scores have been an important feature to integrative methods for prioritizing potentially deleterious non-exonic mutations(Huang et al., 2017; Kircher et al., 2014; Zhou and Troyanskaya, 2015). Non-exonic evolutionarily constrained bases that lack informative annotations from epigenomic and TF binding data can suggest an incomplete coverage of the latter types of data. As a specific example of this, a previous study highlighted how mutations in evolutionarily constrained bases in a gene distal region associated with pancreatic agenesis did not show enhancer activity based on a large panel of datasets from the ENCODE and Roadmap Epigenomic projects(Weedon et al., 2014). However, epigenome mapping experiments specifically in the disease relevant human embryonic stem cell-derived pancreatic progenitor cells did show enhancer activity for the relevant bases.

Previous work has analyzed the overlap of evolutionary constrained bases and biochemical based genomic annotations(ENCODE Project Consortium et al., 2007; Ernst et al., 2011; Kellis et al., 2014; Lindblad-Toh et al., 2011; Margulies et al., 2007; Rands et al., 2014; Ward and Kellis, 2012). However, such work generally focused on analyzing the overlap of individual annotations or had relatively ad hoc approaches to jointly considering multiple annotations. With thousands of datasets across hundreds of cell types now available, one can expect to find at least one dataset with a peak of signal called at most genomic positions just by chance. Therefore, more systematic approaches are needed to distinguish constrained bases in the genome supported by large compendiums of epigenomic and TF binding data, versus those that lack such support. To address this, we developed the Constrained Non-Exonic Predictor (CNEP), which takes a supervised machine learning approach to produce a score for each base of the genome reflecting the probability that the base is in a constrained non-exonic element, given the information in large-scale compendia of epigenomic and TF binding data. We focus specifically on non-exonic bases since they comprise a much larger portion of the genome and are less well annotated compared to exons. Also, constraint in such bases can be expected to be largely associated with distinct patterns of epigenomic marks and TF binding relative to that found in exons.

A number of methods integrate epigenomic, TF binding, comparative genomics, and other types of data and annotations to provide scores for variant prioritization (Huang et al., 2017; Ionita-Laza et al., 2016; Kircher et al., 2014; Ritchie et al., 2014). In contrast to these methods, the only features we use in our predictions are from large compendia of epigenomic and TF binding data, as we want our score to exclusively reflect the extent to which the information in such compendia provides evidence that a location is in a constrained non-exonic base. Such a score can complement scores that use comparative genomic features to prioritize bases in the genome, since when a variant is prioritized based on comparative genomic information there still remains the problem of determining the mechanism by which such a variant exerts influence. Epigenomic and TF binding data can give insights into potential mechanisms, but it is useful to first know whether the prioritization of a variant based on constraint can even be explained by information in the compendium of epigenomic and TF binding data considered, or whether additional experimental data would be needed to do so. Another related method, FitCons(Gulko and Siepel, 2018; Gulko et al., 2015), is primarily designed to use epigenomic information in conjunction with polymorphism and divergence information to provide cell type specific estimates of fitness. In contrast, the primary goal of our method is to estimate a single score, without respect to cell type, that reflects the probability that a base will be in an evolutionarily constrained non-exonic element from information within data from thousands of experiments. Other work has attempted to predict constrained elements based on sequence features, but did not consider experimental data for the prediction(Li et al., 2017).

We applied CNEP to provide a score for each base of the human genome, summarizing the probability of a base being in a constrained non-exonic element based on more than ten thousand features defined from a large compendium of epigenomic and TF binding data. A large portion of bases in constrained non-exonic elements was relatively well predicted, but another large portion was not. We analyzed the constrained non-exonic bases that were not well predicted by CNEP in the context of other annotations, including the recently developed ConsHMM conservation state annotations(Arneson and Ernst, 2019) and human population genetics data. These analyses suggest that a subset of bases in calls of constraint receiving low CNEP scores likely represent false constraint calls, but there also exists a subset of bases that truly appear to be under constraint. We conducted additional analyses using regulatory sequence motif annotations, chromatin accessibility data from mouse, and large-scale prospectively considered human epigenomic and TF binding data to provide insights into the potential role of constrained non-exonic bases not captured by commonly used compendia of epigenomic and TF binding data.

Results

Constrained Non-Exonic Predictor

We developed the Constrained Non-Exonic Predictor (CNEP) to make a probabilistic prediction based on features defined from large-scale epigenomics and TF binding data as to whether a base in the human genome will be in a constrained non-exonic element previously called based on comparative genomics sequence analysis (**Fig. 1**). We applied CNEP with 10,836 features derived from overlap of peak calls in experiments mapping TF binding, histone modifications, and chromatin accessibility, as well as TF binding footprint calls from Digital Genomic Footprinting, and chromatin state annotations from ChromHMM (**Supplementary Table S1-2; Methods**). For simplicity of presentation we refer to TF binding to also include general factor binding that is not necessarily sequence specific. The features are based on data produced by the Roadmap Epigenomics and ENCODE consortia, or data previously included as part of the portion of the ReMap database that curates TF binding data from the Gene Expression Omnibus and ArrayExpress (ENCODE Project Consortium, 2012; Griffon et al., 2015; Roadmap Epigenomics Consortium et al., 2015).

CNEP trains an ensemble of logistic regression classifiers to discriminate between bases overlapping evolutionarily constrained elements outside of a GENCODE annotated exon and those bases in the rest of the genome (**Methods**). For each chromosome, CNEP trains a separate set of classifiers based on subsamples of positions from all chromosomes except the target chromosome. CNEP then makes a probabilistic prediction between 0 and 1 for each base on the target chromosome based on the input features that the base is in a constrained non-exonic element. We applied CNEP with constrained element sets previously produced by four different methods: PhastCons (Rosenbloom et al., 2014; Siepel et al., 2005), GERP++ (Davydov et al., 2010), SiPhy-pi and SiPhy-omega (Garber et al., 2009; Lindblad-Toh et al., 2011) (**Methods**). The PhastCons constrained element set was called based on a 100-way vertebrate alignment, while the other constrained element sets were derived based on a subset of mammals.

Predictions based on training on any two different constrained element sets were all highly correlated, with correlations ranging from 0.91 to 0.96 (**Supplementary Fig. S1**). We therefore averaged the predictions to derive a single score, which we termed the CNEP score (**Fig. 2a, Supplementary Fig. S2**). We confirmed that the CNEP score, based on averaging the four predictions, was even more highly correlated with the individual predictions with correlations ranging between 0.97 and 0.99 (**Supplementary Fig. S1**).

CNEP score associates with signatures of regulatory activity

We next investigated the relationship between CNEP scores and the input features to CNEP. For each input feature, we computed the observed genome-wide average CNEP score of bases overlapped by the feature and compared it to the expected average CNEP score computed based on the proportion of the feature's bases overlapping with constrained non-exonic elements on average for the four element sets (**Methods**, **Supplementary Table S2, Fig. 2b**). Even though the CNEP score is based on more than ten thousand input features, we found that the observed and expected average CNEP score for these features are very strongly correlated (pearson correlation 0.997) (**Fig. 2b**). We also confirmed that the genome-wide observed average CNEP score of 0.041 matched the expected average score of 0.041 based on the average genome coverage of the four constrained element sets.

We next investigated whether bases that received a higher CNEP score were more likely to show signatures of regulatory activity in more experiments or cell and tissue types, which are subsets of input features to CNEP. We analyzed a set of 350 DNase I hypersensitivity experiments from the Roadmap Epigenomics project, and computed the average number of these experiments in which a base would be covered by a peak as a function of the CNEP score (**Fig. 2c**). Bases that received a higher CNEP score tended to be in a DNase I hypersensitivity peak in more experiments. For example, bases with a CNEP score of 0.050 were in a DNase I hypersensitive peak in 1.6% of the experiments while those with a score 0.500 were in a peak in 25.7% of the experiments on average. We saw a similar pattern when considering the chromatin state frequency from a chromatin state model defined across 127 cell and tissue types as a function of the CNEP score, which showed on average a greater presence of candidate enhancer or promoter chromatin states for larger CNEP scores (**Fig. 2d**).

CNEP score is partially predictive of bases in constrained non-exonic elements

We next analyzed the extent to which the CNEP score is able to predict bases in non-exonic constrained elements using Receiving Operator Characteristic (ROC) curves and precision-recall curves (**Fig. 2e,f**, **Supplementary Fig. S3**). We obtained area under the ROC curves in the range 0.75 to 0.82 depending on the constrained element set being predicted, with the area under the curve (AUC) values for PhastCons elements being lower than the other three element sets. The lower AUC for PhastCons might be related at least in part to this being the only element set of the three defined using alignment information that included non-mammalian vertebrates, which can make it more vulnerable to possible alignment errors, or due to the higher resolution at which these elements are defined (Arneson and Ernst, 2019). At false positive rates of 5%, 10% and 20%, the CNEP score predicts 30-39%, 41-53%, and 55-68% of bases in constrained non-exonic elements, respectively, where the specific value in each range depends on the constrained element set.

These results demonstrate that while a substantial portion of bases covered by constrained non-exonic elements is relatively well predicted by CNEP, another substantial portion is not. We next sought to better understand constrained non-exonic bases receiving a low CNEP score. To facilitate that, we defined six sets of bases for each constrained element set: (1) CNE (constrained non-exonic) - non-exonic bases covered by a constrained element;

(2) Low_CNE - bases in CNE that received a CNEP score of 0.05 or less; (3) High_CNE - bases in CNE that are not in Low_CNE; (4) notCNE - bases that are not in a constrained element and also not in an exon; (5) Low_notCNE - bases in notCNE that received a CNEP score of 0.05 or less; (6) High_notCNE - bases in notCNE that are not in Low_notCNE (**Fig. 2a, Supplementary Fig. S2, Supplementary Table S3**). From these definitions, 1.0-2.0% of bases in the genome were in the Low_CNE set. This range corresponds to 34-47% of CNE bases falling into the Low_CNE set, thus a substantial fraction of constrained non-exonic bases called based on sequence constraint had a low CNEP score. For comparison, 15.2-16.5% of the bases in the genome were in the High_notCNE set. These bases received a higher CNEP score than all Low_CNE bases, despite the former not overlapping a base in a constrained element. However, as we investigate below, many High_notCNE bases would be expected simply because of the lower resolution of epigenomic and TF binding data relative to constrained elements. The specific value in these ranges depended on the constrained element set being considered.

CNEP score's ability to predict bases in constrained non-exonic elements varies with ConsHMM conservation state

The previous analysis treated all bases in a constrained element set and those not in a constrained element set as two homogenous sets. However, there is additional information in the multiple species sequence alignment, which might associate with how well the CNEP score can predict CNE bases. To investigate this, we leveraged an annotation of the human genome into 100 conservation states based on the combinatorial and spatial patterns of which species match and align the human genome that was recently produced using the ConsHMM method applied to a 100-way vertebrate alignment(Arneson and Ernst, 2019). We previously showed that the enrichment of bases in constrained elements for specific individual epigenomic datasets can vary substantially depending on the conservation state(Arneson and Ernst, 2019).

Consistent with our previous analyses of enrichments for individual datasets, the CNEP scores' ability to recover CNE bases can vary drastically depending on the conservation state they overlap (**Fig. 3, Supplementary Fig. S4-5**). For example, for PhastCons, while the overall AUC for predicting bases in CNE was 0.75, the AUC was as high as 0.88 when considering CNE bases in ConsHMM state 2, which is a state associated with high frequency of all vertebrates except fish aligning to and matching the human reference genome. This state contains 11.9% of all CNE bases. In contrast, the AUC was less than 0.65 for 64 states, generally having low align frequencies for many mammals, and comprising 20.4% of all CNE bases. We note that state 1, the only state associated with high aligning and matching frequencies for all vertebrates, had an AUC of 0.80, which was less than the AUC value for some other states. We hypothesized that bases being proximal to, but outside exons, could explain the lower AUC for this state. Such bases could be expected to still be constrained, for example because of a role in splicing, but at the resolution the epigenomic and TF binding data was available such bases would not be expected to show a distinct pattern from neighboring bases in exons. We thus repeated the analysis, but when computing the ROC curves and AUC values extended exons to include 200bp of flanking regions on each side. Consistent with our expectations, we saw the AUC for state 1 increase from 0.80 to 0.91, becoming the largest AUC value for any state (**Fig. 3b**, **Supplementary Fig. S4, S6**).

We also directly contrasted the enrichment of Low_CNE to CNE bases and contrasted the enrichment of High_notCNE to notCNE bases in each state. We observed similar results as with the AUC analysis. For example, state 2 had the greatest depletion for Low_CNE bases relative to CNE bases while having the greatest enrichment for High_notCNE compared to notCNE bases (**Fig. 3a,b, Supplementary Fig. S4**). These results highlight that the CNEP score is more predictive for a subset of constrained element calls that are likely to have fewer false positives. However, even for CNE bases in ConsHMM states for which CNEP was most predictive, there was still a substantial subset of bases receiving low scores. For example, for PhastCons, states 2, 4, and 5 all had greater than six fold enrichment for Low_CNE bases and contained 21% of Low_CNE bases.

CNE bases with low CNEP scores and notCNE bases with High CNEP scores are partly explained by proximity to exons and CNE bases

To further understand the extent to which CNE bases with low CNEP scores could be explained by proximity to exons, we computed the prevalence and enrichment of Low_CNE bases as a function of distance to nearest exon (**Fig. 4a,b**, **Supplementary Fig. S7**). For PhastCons CNE bases, we found a 1.6-fold enrichment for bases immediately next to an exon, with the enrichment decreasing further away from the exon. The cumulative enrichment at 200bp away from the nearest exon was 1.2 fold and contained 9.5% of Low_CNE bases. We saw similar results for the other constrained element sets. Therefore, while enriched proximal to exons, the large majority of Low_CNE bases are not explained by proximity to exons.

Limitations in the resolution of the epigenomic and TF binding data relative to the resolution at which constrained elements are defined can also cause notCNE bases that are near a CNE base receiving a high CNEP score to also receive a high score. This would cause other sites with CNE bases to have relatively lower CNEP scores than they otherwise should. To investigate the extent to which High_notCNE bases can be explained by proximity to CNE bases, we computed the prevalence and enrichment of High_notCNE bases as a function of distance to the nearest CNE (**Fig. 4c,d, Supplementary Fig. S8**). For PhastCons High_notCNE bases, we found a 2.7 fold enrichment immediately next to a CNE and decreasing as the distance increases (**Fig. 4c,d**). Among the set of High_notCNE bases, 55% were within 200 base pairs of a CNE base at a 1.8 fold enrichment. We saw similar results for the other constrained element sets. These results indicate that a substantial fraction High_notCNE bases are likely a result of the coarser resolution of the epigenomic and TF binding data. However, there are still many High_notCNE bases that are not proximal to a CNE base, but receive higher CNEP scores than Low_CNE bases.

Low_CNE bases show evidence for purifying selection within humans

To test whether the set of Low_CNE bases are still enriched for bases under purifying selection in humans despite the limited support from the epigenomics and TF binding data considered, we turned to human population genetics data. Specifically, we considered a set of 105 unrelated individuals of the Yoruba in Ibadan (YRI) population from the 1000 Genomes Project and first examined the proportional site frequency spectrum (SFS) (**Methods**). Comparing Low_CNE bases to High_notCNE bases, we observed that there is a significant difference in the distribution ($p<10^{-15}$), with a greater proportion of low-frequency variants for Low_CNE bases, especially singletons and doubletons, and a lower proportion of common variants (**Fig**. **5a**, **Supplementary Fig. S9a,c,e**, comparing orange and purple bars). The skew towards low-frequency variants and the deficit in high-frequency variants suggest stronger purifying selection in the Low_CNE bases relative to High_notCNE bases. As an additional evaluation of whether purifying selection has been stronger in the Low_CNE bases as compared to High_notCNE bases, we examined the absolute SFS normalized by the number of base pairs and an estimated average mutation rate (Carlson et al., 2018) (**Methods**). We found that there are fewer SNPs at the Low_CNE bases relative to High_notCNE bases across all bins of allele frequencies (**Fig. 5b**, **Supplementary Fig. 9b,d,f**, comparing orange and purple bars). These results further suggest that the Low_CNE bases have experienced stronger purifying selection than the High_notCNE bases. We verified that similar results were obtained when controlling for differences in estimated background selection in different sets (McVicker et al., 2009) (**Methods**, **Supplementary Fig. S10**).

We also observed that High_CNE relative to Low_CNE bases and High_notCNE relative to Low_notCNE bases had reduced common variation (**Fig. 5a,b**, **Supplementary Fig. S9**). However, these differences were generally smaller than the differences of CNE to notCNE bases. Additionally, we compared SFS of Low_CNE bases to subsets of High_notCNE bases that satisfied more stringent thresholds on the CNEP score than 0.05 (**Supplementary Fig. S11**), which suggested that the Low_CNE are under stronger purifying selection in humans than notCNE bases that receive substantially higher CNEP scores.

Low_CNE bases show enrichments for TF binding motifs

Since human population genetics data still supported the importance of Low_CNE bases despite the low CNEP score, we investigated whether regulatory sequence motif analysis, which is cell type and condition invariant, provides evidence of a regulatory role for Low_CNE bases. For each constrained element set, we computed individual motif enrichments of Low_CNE bases relative to control motif instances from a compendium of 1,646 regulatory motifs, and did the same for the other five sets above (**Fig. 5c, Supplementary Fig. S12, Methods**). We then analyzed the distributions of these motif enrichments relative to those obtained from randomizing the motif instances. Both the Low_CNE and High_notCNE sets have motif enrichments that are above background, though less than High_CNE bases. The motif enrichments for the Low_CNE bases were substantially stronger than compared to the High_notCNE and notCNE bases. To place the motif enrichments of Low_CNE bases in additional context, we compared them to High_notCNE defined at more stringent thresholds (**Supplementary Fig. S13**). Similar to what we saw with the SFS analysis, the Low_CNE bases had greater enrichment for motifs than High_notCNE at more stringent thresholds of the CNEP score, with the specific thresholds depending on the constrained element set being considered.

To understand the specific motifs driving the overall motif distribution enrichments of the Low_CNE and High_CNE bases, we investigated directly the enrichments of individual motifs (**Fig. 5d, Supplementary Fig. S14, Supplementary Table S4**). Motifs that had the greatest increase in $\log_2$ fold enrichments for High_CNE compared to Low_CNE bases included motifs for the promoter associated *NRF1* and *E2F* TFs (Ernst and Kellis, 2013), *RFX* family of TFs, and the extensively mapped *CTCF*. While globally the High_CNE bases had stronger motif enrichments than Low_CNE bases, we did observe some motifs that showed enrichment for Low_CNE bases and for which the enrichment was greater than for High_CNE bases. Among the motifs which had a $\log_2$ enrichment for Low_CNE greater than 0.5, the greatest difference in

$\log_2$ enrichments relative to High_notCNE included motifs corresponding to *CUX1*, *ESRR* family of TFs, and the *NR5A* family of TFs. We conducted a Gene Ontology (GO) enrichment for TFs corresponding to the set of motifs that had at least a $\log_2$ fold enrichment of 0.5 in High_CNE or Low_CNE bases broken down into three sets: (i) 'High_CNE strongly preferred' - those motifs for which difference for High_CNE and Low_CNE bases was greater than 0.5; (ii) 'High_CNE moderately preferred' - the difference was between 0.5 and 0, which by definition all still showed enrichment for Low_CNE bases; (iii) 'Low_CNE preferred' - which had a greater enrichment for Low_CNE bases than High_CNE bases (**Supplementary Table S5**). TFs associated with 'High_CNE strongly preferred' motifs showed significant enrichment for protein dimerization activity and core-promoter GO terms. TFs associated with 'High_CNE moderately preferred' motifs showed enrichment for development related GO terms. Finally, TFs associated with 'Low_CNE preferred' motifs showed enrichment for lipid binding and response to stimulus related GO-terms (corrected p-values <0.05). These results suggest that some of the CNE bases that are more difficult to predict from the epigenomic and TF binding data considered are associated with sites that might only be active in specific developmental stages or under specific stimuli.

Low_CNE bases show some enrichments for mapped mouse DNase I hypersensitive sites

As mouse experiments can potentially have coverage of cell or tissue types and developmental stages not represented in human data, we investigated the extent to which DNase I hypersensitive sites (DHS) in mouse mapped to the human genome enriched for Low_CNE bases. Specifically, we analyzed a set of 156 DNase I hypersensitivity experiments from the mouse ENCODE project(Vierstra et al., 2014; Yue et al., 2014). For each experiment, we mapped the set of mouse DHS to the human genome and did the same for datasets where we first randomized the location of the peaks in mouse (**Methods**). For each constrained element set, we then computed the enrichment of Low_CNE bases for the actual dataset relative to the randomized dataset, and also did the same for CNE, High_CNE, High_notCNE, Low_notCNE, and notCNE bases.

We found that for all element sets, the Low_CNE bases showed enrichment for the mapped DHS at least for a subset of the experiments (**Fig. 5e, Supplementary Fig. S15**). These enrichments were modest, not exceeding 2-fold for any DHS experiment or constrained element set, and were lower than what was seen for CNE and High_CNE bases. However, the enrichments were greater than for Low_notCNE and notCNE bases and comparable to High_notCNE bases for at least the most enriched experiments. We observed that the DHS experiments that tended to have the greatest enrichment for Low_CNE bases were either for whole brain or cerebrum or conducted in mouse embryos at day 11.5 (**Fig. 5f**, **Supplementary Fig. S16, Supplementary Table S6**). For example, for PhastCons, the 32 DHS experiments with the greatest enrichment included all 25 experiments done on whole brain and cerebrum and six of eight experiments from mouse embryos at day 11.5. The only additional experiment in the top 32 was conducted in retina at the newborn-1 day stage. Of the eight experiments in mouse embryos conducted at day 11.5, the six experiments in mesoderm, forelimb bud, and hindlimb bud were in the top 32, while the two experiments for a headless embryo were ranked much lower. These results provide evidence to suggest that a limited portion of Low_CNE bases are regulatory active in corresponding positions in available mouse samples, particularly related to the brain or embryonic development.

Additional information to predict CNE bases in specialized human datasets

The CNEP predictions were based on only datasets that were available by 2015. Since then, many additional datasets have become available or more accessible. We leveraged large collections of additional human datasets to conduct a prospective analysis to identify datasets that provide additional marginal additive information predictive of CNE bases, beyond the information already summarized in the CNEP score. Identifying such datasets will highlight types of datasets that are underrepresented in commonly used compendiums of data relative to the information they provide about CNE bases in the genome.

Specifically, we analyzed 31,901 datasets of peak calls from ChIP-atlas, which provides a uniform processing of Short Read Archive data(Leinonen et al., 2011; Oki et al., 2018), 1,755 datasets of peak calls of TF binding from ReMap 2018(Chèneby et al., 2018), and 16,711 peak calls from the ENCODE portal(Davis et al., 2018; ENCODE Project Consortium, 2012). CNEP's use of the ReMap database was limited to the 395 datasets of peak calls from the 2015 version of curated public data, and CNEP's use of ENCODE data was limited to data from the second phase of the project(ENCODE Project Consortium, 2012). In contrast, here we used all human ChIP-seq and DNase-seq peak calls available on the ENCODE portal as of May 2018. We note that differences in data processing procedures between databases lead to differences in the sets of peak calls between databases in addition to differences in the underlying data they use.

For each dataset, we computed the observed average CNEP score for bases in each peak, and the expected average CNEP score based on its overlap with constrained element calls, as we did in **Fig. 2b** (**Supplementary Fig S17, Supplementary Table S7**). We observed overall high correlations, with correlations in the range of 0.97-0.99 for sets of peaks covering at least 200kb. We then compared the distribution of the difference between the expected and observed CNEP score to the distribution obtained if we shuffled the location of the peaks within each set of peak calls (**Fig. 6a-c**). We also compared these distributions to the distribution that would be obtained if instead of the observed CNEP score we used the genome-wide expected average CNEP score when computing the difference. There was a relatively large gap in distribution based on using the observed CNEP score instead of genome-wide expected average CNEP score when computing the difference, highlighting that the CNEP score captures a relatively large amount of information about CNE bases. There is a difference in distributions between using the CNEP score on the actual peaks and a shuffled version of the peaks, though it was much smaller. These results suggest that CNEP captures most of the marginal information contained in any peak call dataset about the expectation on the frequency of CNE bases. However, there are some datasets that capture some additional marginal information on CNE bases than given by the CNEP score.

To better understand datasets that overlap CNE bases more than is expected by the CNEP score, we defined the CNEP underestimation value for a dataset as the difference between the expected and observed average value of the CNEP score for bases overlapping a peak in the dataset. We then plotted this quantity as a function of the number of bases the peaks overlap, for those datasets with a positive CNEP underestimation value (**Fig. 6d**). Particularly informative datasets for explaining CNE bases beyond what is in the CNEP score would have a large value for the CNEP underestimation and also cover many bases of the genome. As controls, we did the same for features directly provided to CNEP and shuffled versions of each dataset (**Supplementary Fig. S18**). We observed datasets with greater underestimation values for the same genomic coverage than in both controls, demonstrating additional marginal additive information about CNE bases in these datasets. However, in most cases the CNEP underestimate for a dataset was relatively small (<~0.01) or only applied to relatively few bases, indicating that the additional marginal information in those datasets is limited.

Among the exceptions was a dataset of narrow peak calls from a DNase I hypersensitivity experiment in spinal cord of a 59-day embryo that covered 71.6 million bases and had an underestimation value of 0.066. No experiment that covered more than 4.5 million bases had a greater underestimation value than this dataset. A broader set of peak calls for this same experiment covered 179.5 million bases and had an underestimation value of 0.054 bases, which was the largest underestimation value of any dataset covering more than 71.6 million bases. Excluding these two peak calls sets and additional version of the peak calls from this same experiment, a peak call set for a DNase I hypersensitivity experiment in the embryonic brain covering 57.3 million bases had a higher underestimate value than any dataset covering more than 11.6 million bases. Other peak call sets or experiments of DNase I hypersensitivity of embryonic spinal cord, brain, eye and retina also had notably high combinations of underestimate values and genome coverage (**Fig. 6d**).

In **Fig. 6d**, we highlight selected experiments for TF binding that had relatively high underestimate values for their genome coverage. Some of these had even higher CNEP underestimate values than observed in the embryonic DNase I hypersensitivity experiments, though covered a smaller fraction of the genome. These include *PDX1* and *ONECUT1* in pancreatic progenitor cells from the study which previously showed the specific relevance of an enhancer highly specific to this cell type to pancreatic agenesis (Weedon et al., 2014), as well as other experiments mapping *PDX1* (Teo et al., 2015; Wang et al., 2015). Experiments for TFs including *PHOX2B*, *HOXC9*, and *TEAD4* in neuroblastoma cell lines, BEC2 or CLB-GA, also had high underestimate values(Boeva et al., 2017; Rajbhandari et al., 2018; Wang et al., 2013). Experiments mapping *AR* and HOXB13 in LHSAR cells, a prostate epithelium cell line, with overexpression of *HOXB13*, were other experiments with relatively high underestimate values for their genome coverage (Pomerantz et al., 2015; Yin et al., 2017). *CDX2* in colon carcinoma (Colo320) (Salari et al., 2012) and *HOXA9* in Human Embryonic Kidney cells (HEK293) (Rio-Machin et al., 2017) were two additional experiments that both covered at least 300kb and had an underestimate value greater than 0.10. We also leveraged the ChIP-atlas cell type class metadata annotations to determine if there were specific cell type classes which had datasets significantly enriched with underestimate values greater than 0.02, restricted to those datasets covering at least 200kb (**Supplementary Table S8**). We found enrichment of datasets of pluripotent stem cell, pancreas, and neural to be the most significantly enriched (corrected p-value <0.01).

We also compared the underestimate values and coverage of these datasets to new exons added to GENCODE between v19 and v28. New exons in GENCODE covered 12.3 million bases with an underestimate of 0.038. These values, while greater than those seen for many experiments, were still less than a single set of peak calls for both TEAD4 and ONECUT1 in terms of both genomic coverage and underestimate value and substantially less than some of the embryonic DNase I hypersensitivity experiments.

Discussion

In this work we developed and applied the Constrained Non-Exonic Predictor (CNEP) to provide a score for each base of the human genome that reflects the probability that the base overlaps a CNE from information in large-scale collections of epigenomic and TF binding data. We used information from more than ten thousand features derived from epigenomic and TF binding data spanning a wide range of cell and tissue types. We showed that a substantial portion of constrained non-exonic bases is relatively well predicted by CNEP. However, despite the large number of features used, a substantial portion of CNE bases was not well predicted. For example, for PhastCons, at a CNEP score threshold of 0.05, 47% of CNE bases would not be predicted, while 16.5% of the genome in notCNE bases would be.

To better understand the nature of CNE bases that were not well predicted by CNEP, Low_CNE bases, we conducted analyses with a number of other types of data. We showed that there was an enrichment for Low_CNE bases near exons receiving low scores, but this provides only a limited explanation for some of the disagreement. Using the recently developed ConsHMM conservation state annotations we observed that Low_CNE bases were less likely to be found in conservation states with robust alignment patterns compared to CNE bases in general. However, even in states with robust alignment patterns through distal vertebrates, we observed a substantial percentage of Low_CNE bases. Using human population genetic variation data, we provided evidence to suggest that the Low_CNE bases are under constraint in humans, though to a lesser extent than the High_CNE bases. Consistent with the trends of the human population variation data, regulatory sequence motifs also showed enrichment in Low_CNE bases, though to a lesser extent than in the High_CNE bases. We found development related TFs to be associated with motifs that enriched in Low_CNE bases and showing only moderately stronger enrichment in High_CNE bases. Stimulus response TFs were associated with motifs that had stronger enrichments in the Low_CNE set of bases than the High_CNE. Mouse DHS mapped to human also showed enrichment for Low_CNE bases, particularly for brain and embryonic development related samples.

We also identified a substantial fraction of the genome that received high CNEP scores, but were not in constrained elements, High_notCNE bases. A simple explanation for many

High_notCNE bases was their proximity to constrained elements, and the finer resolution at which constrained elements are defined compared to the resolution of the epigenomic and TF binding data. However, there was also a substantial fraction of High_notCNE bases not proximal to constrained elements. High_notCNE bases showed greater enrichment for regulatory motifs and less genetic variation compared to Low_notCNE bases, though this was not the case when comparing to CNE or even Low_CNE bases. A subset of High_notCNE bases might correspond to bases that are under evolutionary constraint in humans, but not actually in a constrained element call, which was supported by the conservation state enrichments for High_notCNE bases. Another subset of High_notCNE bases may correspond to recently evolved bases with a potentially important regulatory role that share epigenomic marks and TF binding patterns associated with CNE bases.

One route for improving CNEP predictions of CNE bases would be to incorporate additional epigenomic and TF binding data, particularly data distinct from that well covered by the compendia of data already used. To provide a perspective of the marginal information available in additional experimental datasets we evaluated the overlap of peaks in additional datasets with CNE bases and compared it to the CNEP score at those bases. We found most datasets offered at most very small additional marginal additive information than the information already in the CNEP score to predict CNE bases, though we did identify some additional datasets that would provide more substantial additional information. These datasets included DNase I hypersensitivity datasets for embryonic spinal cord, brain, and eye as well as some specialized TF binding experiments. We saw an enrichment of datasets with relatively informative additional marginal information categorized as being from pluripotent stem cell, pancreas, or neural cell type classes. While specific, these additional datasets could provide important information in the context of certain diseases. For example, these datasets included some that were previously shown to provide unique information for studying pancreatic agenesis(Weedon et al., 2014). Also, chromatin accessibility data from fetal brain has been implicated to give specific information relevant to neuropsychiatric diseases(de la Torre-Ubieta et al., 2018). In addition to information from the assays considered here in additional cell types and conditions, it is also possible different types of assays would enable improved prediction of CNE bases.

Another route for improving CNEP predictions of CNE bases is to have more accurate calls of CNE bases. Of the four constrained element sets considered the only one that considered non-mammalian vertebrates and used the largest set of mammal sequences actually had a lower AUC. This suggests potential room for improvement in defining CNE bases, and consistent with that we showed that the predictability of CNE bases depends heavily on the ConsHMM conservation state to which it is assigned. We note, however, that the quality of a constrained set of predictions should not be evaluated solely on how well they can be predicted from epigenomic and TF binding data, since the resolution for defining constraint is finer than that at which predictions based on epigenomic and TF binding data can make meaningful distinctions.

The objective of the CNEP score is different and complementary from those scores that integrate epigenomic, comparative genomic, and other annotations to prioritize variants. Since the set of features CNEP uses is more restricted and notably does not include comparative genomics information it would not be expected to be competitive at variant prioritization tasks. However, by only using features from epigenomic and TF binding data in a compendium, the CNEP score provides a way to assess whether the datasets in a compendium provides evidence to support a variant potentially prioritized because of non-exonic constraint at a genomic position. Additionally, the CNEP score is learned based on many more epigenomic and TF binding features than used to derive other scores, and thus might become a useful feature for integrative methods that produce scores based on a more limited, but more diverse set of annotations and other scores.

This work highlights that available compendia of epigenomic and TF binding data contain information to explain a substantial fraction of CNE bases. However, there are also subsets of CNE bases that are difficult to predict from such compendiums. The importance of a subset of such bases is supported by orthogonal evidence, thus also highlighting the remaining challenge to a comprehensive understanding of the non-exonic genome.

Methods

Availability of CNEP scores and CNEP software

The CNEP scores and software are available from https://github.com/ernstlab/CNEP.

Genome assembly and gene annotations

All predictions and analysis were done on human genome assembly hg19 and were restricted to chr1-22 and chrX. For gene annotations we used the GENCODE v19 annotations obtained from ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz. Exon annotations include exon bases that are non-coding.

Constrained element sets

We used four different constrained element sets based on the PhastCons(Siepel et al., 2005), GERP++(Davydov et al., 2010), SiPhy-omega, and SiPhy-pi (Garber et al., 2009; Lindblad-Toh et al., 2011) methods. The PhastCons constrained elements were based on the human hg19 100-way vertebrate alignment and obtained from the UCSC genome browser(Rosenbloom et al., 2014). The SiPhy-omega and SiPhy-pi elements were called based on a 29-way mammalian alignment and were the hg19 version obtained from https://www.broadinstitute.org/mammals-models/29-mammals-project-supplementary-info. The GERP++ elements were called based on the mammalian subset of the UCSC genome browser hg19 46-way vertebrate alignment obtained from http://mendel.stanford.edu/SidowLab/downloads/gerp/.

Epigenomics and TF binding features

We used 10,836 binary features defined from functional genomics data. The sources of the features are found in **Supplementary Table S1** and a list of features can be found in **Supplementary Table S2**. The features were derived from ChIP-seq data of histone modifications, TFs including general factors, DNase I hypersensitivity data, and FAIRE data. The data was produced by the ENCODE consortium during its second phase(ENCODE Project Consortium, 2012), Roadmap Epigenomics consortium(Roadmap Epigenomics Consortium et al., 2015), and part of the ReMap public dataset(Griffon et al., 2015), which is a reprocessing of non-ENCODE ChIP-seq data of TFs from the Gene Expression Omnibus and ArrayExpress. The peak calls for the Roadmap Epigenomics data was based on the unconsolidated datasets. In total 5,579 features were based on peak calls. For these features the data was encoded as a ‘1’ if the corresponding base overlapped a peak and ‘0’ otherwise. Additionally, we had 42 features defined based on the position of Digital Genomics Footprints(Neph et al., 2012; Roadmap Epigenomics Consortium et al., 2015). For these features the data was encoded as a ‘1’ for those bases overlapping a footprint and a ‘0’ otherwise. We also had 5,215 features defined based on chromatin state calls from three different ChromHMM models(Ernst and Kellis, 2012). The three models were: (1) a 15-state model defined across 9-ENCODE cell types based on eight histone modifications and CTCF; (2) the 15-state ‘core’ model based on 5-histone modifications defined across 127-reference epigenomes based on consolidated data processed by the Roadmap Epigenomics consortium (111 reference epigenomes were derived from data produced by the Roadmap Epigenomics project and 16 from the ENCODE project); (3) a 25-state model based on imputed data for 12-chromatin marks (10 histone modifications, H2A.Z, and DNase I hypersensitivity) defined across the same 127 reference epigenomes(Ernst and Kellis, 2015). For each model we had a separate feature for each chromatin state and cell type or reference epigenome combination. A feature value was encoded as a '1' if a base overlapped the chromatin state in the cell type or reference epigenome and a '0' otherwise.

CNEP method

The CNEP scores are generated by first training an ensemble of logistic regression classifiers. For a given constrained element set and a set of binary functional genomics features, CNEP trains logistic regression classifiers to discriminate between bases in a constrained element that are outside of all exons as a positive set from all other bases as a negative set. For generating CNEP scores on one chromosome based on one constrained element set, CNEP trained ten logistic regression classifiers using different 1,000,000 randomly sampled positions from the other 22 chromosomes. We repeated this for each of the constrained element sets and 23 chromosomes thus training in total 920 logistic regression classifiers. CNEP used the Liblinear v.2.1(Fan et al., 2008) software to train the logistic regression classifiers using L_1 regularization (-s 6) with a bias term (-B 1), with the default regularization parameter value of 1 (-c 1). The exclusive use of binary features allowed us to make efficient use of the sparse representation of the data in Liblinear. For generating genome-wide predictions based on one constrained element for each chromosome, CNEP computed and averaged the probabilistic predictions from its ten corresponding logistic regression classifiers and then outputted the predictions to the nearest 0.001 value. To generate the CNEP score we then averaged the outputted predictions based on each of the four constrained element sets.

Computing observed and expected average CNEP scores for features and genome-wide

For computing the observed average CNEP score for a feature, we computed the average CNEP score in all bases in the genome where the feature was defined as being present. For computing the expected average CNEP score for a feature, we computed the average over the four constrained element sets of the number of bases for which the feature was present and overlapped a constrained non-exonic element divided by the total number of bases in which the feature was present. We computed the genome-wide observed and expected CNEP scores the same way except all bases in the genome were included.

Analysis of CNEP score's relationship to Roadmap Epigenomics DNase I hypersensitive peak coverage

For computing the relationship between CNEP score and average fraction of Roadmap Epigenomics DNase I hypersensitivity experiments in a peak (**Fig. 2c**), we used 350 narrowPeak call files with 'ChromatinAccessibility' in the file name available from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/unconsolidated/narrowPeak/. For each value of the CNEP score computed to the nearest 0.001 value covering at least 1000 bases, we took all bases in the genome having that score and determined the average fraction of the 350 experiments in which the bases are overlapped by a peak call.

Analysis of CNEP score's relationship to chromatin states

For the analysis of the relationship between CNEP score and chromatin state annotation (**Fig. 2d**), we used the 25-state ChromHMM chromatin state annotations defined across 127 epigenomes based on imputed data for 12-chromatin marks (Ernst and Kellis, 2015). For each CNEP score, which were computed to the nearest 0.001 value, we took all bases in the genome having that score and determined the average fraction of the 127 epigenomes in which each of the 25-states overlapped the base. We then stacked bar graphs with fractions starting from the state with the greatest state number (25_Quies) to the lowest state number (1_TssA) with the state numbers and colors from (Ernst and Kellis, 2015). In the plot we did not differentiate between different states that were previously given the same color and thus the graph provides information on the 14-state groups that were colored differently.

Defining sets of bases for analyses

For each of the four constrained element sets considered, we defined the following set of bases used in some analyses: (1) CNE – bases in a constrained element that do not overlap a GENCODE exon; (2) Low_CNE – bases in a constrained element that have a CNEP score less than or equal to 0.05; (3) High_CNE – bases in a constrained element that have a CNEP score greater than 0.05; (4) notCNE – bases not in a constrained element and do not overlap a GENCODE exon; (5) Low_notCNE – bases not in a constrained element and do not overlap a GENCODE exon and have a CNEP score less than or equal to 0.05; (6) High_notCNE – bases not in a constrained element and do not overlap a GENCODE exon and have a CNEP score greater than 0.05 (**Supplementary Table S3**).

Conservation state analysis

The ConsHMM conservation state annotations were the 100-conservation state annotations for hg19 from (Arneson and Ernst, 2019). These conservation state annotations were defined on the same 100-way vertebrate alignment for which the PhastCons bases we used were defined. Fold enrichments for CNE, Low_CNE, notCNE, and High_notCNE bases in the conservation states were computed using the OverlapEnrichment command of ChromHMM v1.17 with the options '-b 1 -lowmem' specified. Conservation state assignments to chrY were excluded from the background in this analysis, as the CNEP scores were not defined on this chromosome. The per state ROC and AUC values for the CNEP score were computed by considering a positive base a CNE in a specific conservation state and a negative base any base in the genome that was not in a CNE. Bases in the CNE set that were in a different conservation state were excluded when generating the ROC and computing the AUC. The ROC and AUC based on extending exons 200bp in each direction was computed in the same way, except first adjusting the exon start and end positions.

Positional enrichment analysis relative to exons and CNE bases

For computing the enrichment of Low_CNE bases in proximity to exons we computed for each base in the genome the distance to the nearest base of any exon. The enrichment of Low_CNE at a specific distance to the nearest exon was defined as the ratio of the fraction of Low_CNE bases whose nearest exon was at that distance to the fraction of all bases in the genome whose nearest exon was at that distance. A similar set of enrichments was computed for High_notCNE bases in relation to their distance to the nearest CNE base.

Human variation analysis

The human variation analysis was conducted on a set of 105 unrelated individuals from the Yoruba in Ibadan (YRI) population part of the 1000 Genomes Project. We focused on this population for analyzing the effects of selection since it is associated with greater genetic diversity and has a simpler demographic history than non-African populations (Gutenkunst et al., 2009). We selected high quality sites by applying a mask from 1000G where a site was defined as high quality if its depth (DP) is within 1.5x the mean DP across all sites (1000 Genomes Project Consortium et al., 2015). For this analysis, we restricted it to the autosomes and variant calls that were bi-allelic. For each set of coordinates analyzed, we computed a count c_n of how many of the variants occurred in exactly *n* individuals for each value of *n*=1,…,10 (low and intermediate frequency variants) and also a count $c_{>10}$ of how many occurred in greater than 10 individuals (common variants). We then computed the proportional site frequency spectrum (SFS) as each of these individual counts divided by the sum of all of the counts. We assessed the statistical significance between pairs of coordinate sets by applying a chi-square test to the 11 count values.

The absolute SFS contains the numbers (rather than proportions) of SNPs at particular minor allele counts. Because the count of SNPs is affected by the number of base pairs analyzed (more base pairs would lead to more SNPs) as well the mutation rate (higher mutation rates lead to more SNPs), we normalized for both of these factors. To do this, first we obtained mutation rate estimates from http://mutation.sph.umich.edu/hg19/ (Carlson et al., 2018). We associated each base with a single mutation rate by averaging its three mutation rates, each corresponding to a mutation from the reference nucleotide to an alternative nucleotide. For a coordinate set, we computed the sum of the mutation rates at all bases that were high quality sites as defined above and had a mutation rate available. This sum is equivalent to the number of base pairs analyzed in a coordinate set times their average mutation rate. We computed the unnormalized count values as described above for the proportional SFS except excluded positions that did not have mutation rates available. We then divided these counts by the sum of the mutation rates.

To compute SFS controlled for background selection we used the version of *B*-values in hg19 as part of the CADD annotation set, which are based on the *B*-values from (McVicker et al., 2009). For the proportional SFS, we reweighted variant calls in each coordinate set so that the *B*-value distribution was effectively the same as the distribution of *B*-values at all non-exonic bases with a variant call. This analysis was restricted to non-exonic variants that had an estimated *B*-value available. The weighting for a variant with a *B*-value, *x*, was $p_a(x)/p_s(x)$ where $p_a(x)$ and $p_s(x)$ are the proportions of variants with the *B*-value *x* among all variants considered and the subset in the coordinate set, respectively. For the absolute SFS density normalized by its average mutation rate, we reweighted bases in each coordinate set so the *B*-value distribution was effectively the same as the distribution of *B*-values at all non-exonic bases. This analysis was restricted to non-exonic bases that were in a high quality site and had both estimated mutation rates and *B*-values available. The weighting for a base with a *B*-value, *x*, was $p_c(x)/p_s(x)$ where $p_c(x)$ and $p_s(\mathrm{x})$ are the proportions of variants with the *B*-value *x* among all bases considered and the subset in the coordinate set respectively. The weighting was used in both counting variants and the sum of the mutation rates.

Motif enrichment analysis

For the motif enrichment analysis (**Fig. 5c,d, Supplementary Fig. S12-S14, Supplementary Table S4**), we used motif instances from http://compbio.mit.edu/encode-motifs/matches-with-controls.txt.gz (Kheradpour and Kellis, 2014). We used motif instances for a set of 1,646 motifs that excluded motifs that were in the compendium based on being discovered from ENCODE ChIP-seq data, so that the set of motifs we analyzed were independent of the features provided to CNEP. The motif instances were called outside of coding, 3' UTR, and repetitive regions and called independent of conservation. For each motif, there were also a set of corresponding control motif instances called(Kheradpour and Kellis, 2014), which control for biases from sequence composition or background. To compute the enrichment of a specific motif in a target set of bases, we computed the ratio of the fraction of motif instance bases that also overlapped the target set to the fraction of corresponding control motif instance bases that also overlapped the target set. For each of the four constrained element sets, these enrichments for individual motifs were reported for High_CNE and Low_CNE bases (**Fig. 5d**, **Supplementary Fig. S14**, **Supplementary Table S4**). For the analyses on the distribution of motif enrichments for a target set, we generated three randomized versions of the motif instance calls with controls. To generate a randomized version for each chromosome we performed column-wise random permutations where one column is the motif identifier and the other column contains the motif coordinates. For each target set considered, we computed the distribution of motif enrichments on each of the randomized motif instances, using the same procedure as the actual motif instances. We then ordered the enrichments separately for the actual and three randomized datasets. At each ranked position in the ordering we took the difference between the $\log_2$ value of the actual enrichment and the $\log_2$ value of the median enrichment from the three randomized datasets.

Gene Ontology analysis

Gene Ontology enrichment analysis for the TFs corresponding to sets of motifs was conducted using the STEM software v.1.3.11 with default settings (**Supplementary Table S5**) (Ernst and Bar-Joseph, 2006). The Gene Ontology and human gene annotations were downloaded using the STEM software on September 17, 2017. We used as a base set all TFs corresponding to a motif in the compendium. The corresponding TF for a motif was taken to be the portion before the '_' in the motif ID.

Mouse DNase I hypersensitive site enrichment analysis

For the mouse DNase I hypersensitivity site (DHS) analysis (**Fig. 5e,f, Supplementary Fig. 15-16, Supplementary Table S6**), we used the 156 narrowPeak files from the University of Washington mouseENCODE group available from

http://hgdownload.soe.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/ and

http://hgdownload.soe.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDgf/ (Vierstra et al., 2014; Yue et al., 2014). We also generated a randomized version of each set of DHS by randomly selecting a different position for each DHS in the original file on the same chromosome. We lifted over both the real and randomized versions of the DHS files from mm9 to hg19 using the liftOver tool from the UCSC genome browser with the options '-bedPlus=3 -minMatch=0.00000001'. The lower value for the minMatch parameter enables a more permissive mapping of peaks from mouse to human and thus an enrichment estimate that is more reflective of a background that includes all mouse DHS. For both the real and randomized version of each set of DHS, for each of the four constrained element sets considered, we computed enrichments for the six target sets: CNE, High_CNE, Low_CNE, notCNE, High_notCNE, and Low_notCNE. Enrichments for each set of DHS were computed by taking the ratio between the fraction of bases in human covered by a DHS that are in the target set to the fraction of bases in the genome that are in the target set. The reported enrichment for an experiment is the ratio of this enrichment for the real DHS compared to the corresponding enrichment for the randomized DHS.

Analysis on additional human datasets

For the analysis of additional human datasets we used data from ChIP-atlas(Oki et al., 2018), ReMap 2018(Chèneby et al., 2018), the ENCODE portal(Davis et al., 2018; ENCODE Project Consortium, 2012), and updated exon annotations from GENCODE. For the ChIP-atlas we used all peaks called at the 10^{-5} threshold for hg19 available from http://dbarchive.biosciencedbc.jp/kyushu-u/hg19/eachData/bed05/ on May 11, 2018. We excluded files that did not have any peaks called on the chromosomes we considered. For the ENCODE portal data we downloaded all files for the ChIP-seq or DNase-seq assay available in narrowPeak or broadPeak format for hg19 on May 11, 2018 produced by the ENCODE project from https://www.encodeproject.org/. For the ReMap 2018 database we used the peaks in the hg19 files restricted to the 'Public' data (non-ENCODE). We note that some of the ChIP-atlas datasets were generated by the ENCODE project, but processed differently. We did not exclude any dataset for being based on the same experiment as used to generate the CNEP predictions. For the updated GENCODE exon annotations we used release 28 mapped to hg19/GRCh37 available from

ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_28/GRCh37_mapping/gencode.v28lift37.annotation.gtf

We excluded any base in an exon from release 19. For generating the shuffled data we used the shuffleBed command of BEDTools(Quinlan and Hall, 2010).

For computing the cell type class enrichments, we used STEM software v.1.3.11 with user provided annotations treating each dataset as if it was a gene and the cell type class as the annotation category (Ernst and Bar-Joseph, 2006). The foreground for the enrichment were those ChIP-atlas datasets with peaks covering at least 200kb and having an underestimate value greater than 0.02. The background set for the enrichment analysis was all ChIP-atlas datasets with peaks that covered at least 200kb. We used default settings except changed the minimum number of genes parameter to 1 and multiple hypothesis testing correction to 'Bonferroni'.

Acknowledgements

We thank Xiaorui Fan and Petko Fiziev for useful discussions and assistance. We acknowledge funding from US National Institutes of Health grants DP1DA044371, R01ES024995, U01HG007912 and U01MH105578 (J.E.), T32CA201160 (A.S.), and R35GM119856 (K.E.L.), US National Science Foundation CAREER Award #1254200 (J.E.), a Kure-IT award and an Alfred P. Sloan Fellowship (J.E.). We acknowledge the ENCODE and Roadmap Epigenomic consortia for generating some of the data used and making it available pre-publication.

Figures Legends

Figure 1: Example of Constrained Non-exonic Predictor (CNEP) scores. An example genomic locus illustrating CNEP scores. The top line is the GENCODE gene annotation track followed by the CNEP score track. In general, the CNEP score ranges between 0 and 1, but in this image the y-scale is capped at 0.5. Below the CNEP score track is the PhastCons element

track, the PhastCons score track, the UCSC Genome Browser ENCODE TF binding summary track (Txn Fac ChIP V2), and the ENCODE DNase I summary track (DNase Clusters V3). These tracks are then followed by chromatin state annotation across 127 samples based on a previously defined 25-state ChromHMM annotation based on imputed data (Ernst and Kellis, 2015). A color legend for the chromatin state annotations is also available in **Fig. 2d**.

Figure 2: Properties of the CNEP score. (a) The graph shows the cumulative distribution of the CNEP score genome-wide (green), in PhastCons constrained non-exonic (CNE) bases (red), and bases that are not in PhastCons constrained elements and also not in exons (notCNE) (blue). **(b)** A scatter plot with each point corresponding to one feature that CNEP uses. The x-axis shows the average CNEP score in bases that have the feature present, while the y-axis shows the expected CNEP score based on the feature's overlap with constrained non-exonic bases. Only features that cover at least 200kb are shown, which is 10,741 of the 10,836 features. The full table corresponding to these values can be found in **Supplementary Table S2**. The diagonal line is the y=x line. The vertical line corresponds to the genome-wide average CNEP score. The horizontal line corresponds to the genome-wide expected average CNEP score. **(c)** A plot showing the average fraction of the 350 Roadmap DNase I experiments in which the base overlaps a called peak for each CNEP score value, rounded to the nearest 0.001, covering at least 1000 bases. In total, there was 996 such values. **(d)** A plot showing the average fraction of bases assigned across the 127 epigenomes to each of 14-groups based on 25 ChromHMM chromatin states previously assigned the same color for each CNEP score value, rounded to the nearest 0.001 (Ernst and Kellis, 2015). A color with the state abbreviations is displayed at the bottom of the panel. **(e)** A plot of the ROC curve for the CNEP score predicting PhastCons non-exonic bases. The area under this curve is 0.75. **(f)** A plot of the precision-recall curve for the CNEP score identifying PhastCons non-exonic bases. ROC and precision-recall curves for other constrained element sets can be found in **Supplementary Fig. S3**.

Figure 3: CNE prediction depends on conservation state. (a) Heatmap representation of conservation state parameters of the ConsHMM conservation state model defined in (Arneson and Ernst, 2019). Rows correspond to different conservation states. The states were previously clustered into eight groups based on these parameters and colored accordingly. The left half indicates for each state the probability of each species having a nucleotide aligning the human reference genome, regardless of whether it matches the human reference. The right half indicates for each state the probability of each species having a nucleotide matching the human

reference genome. Individual columns correspond to species, the names of which are available in (Arneson and Ernst, 2019). The major groups of species are colored and labeled. Color scale for the heatmap is shown at the bottom. **(b)** The first column reports the genome % of each state excluding chrY. The second column contains the AUC of the CNEP score for predicting CNE bases in each state, where for this and the remaining columns the constrained elements are from PhastCons. CNE bases that are not in the target conservation state are excluded when computing the AUC. The next column reports the AUC when exons are first extended by 200bp. The next three columns contain the fold enrichment for CNE bases, Low_CNE bases, and the ratio of the enrichment of Low_CNE bases to CNE bases. The next three columns contain the fold enrichment for notCNE bases, High_notCNE bases, and the ratio of the enrichment of High_notCNE bases to notCNE bases. Adjacent pairs of columns on a red-white color scale are on the same color scale. The other columns are on a column specific color scale. The bottom row gives the base % of the genome for the four sets. Results based on all the constrained element sets can be found in **Supplementary Fig. S4**. **(c)** ROC curves for the CNEP score identifying PhastCons CNE bases in specific ConsHMM conservation states. Curves are colored based on their corresponding state, as shown in panel (a). **(d)** Plot showing the AUC values for each ROC curve in (c). The AUC values are displayed from left to right in decreasing value and positioned along the x-axis based on the cumulative fraction of PhastCons CNE bases they cover. The points are color-coded based on the conservation state coloring shown in (a). States with the highest AUC values are labeled. Similar plots to (c) and (d), but for additional constrained element sets can be found in **Supplementary Fig. S5** and based on excluding bases within 200bp of exons can be found in **Supplementary Fig. S6**.

Figure 4: Low_CNE bases enrichment near exons and High_notCNE bases enrichment near CNE bases. (a) The plot shows the cumulative fraction of PhastCons Low_CNE bases at each distance to the nearest exon, up to 3,000bp. **(b)** The plot shows the fold enrichment for the cumulative number of PhastCons Low_CNE bases being within each distance to the nearest exon, up to 3,000bp. Similar plots to (a) and (b) for other constrained element sets can be found in **Supplementary Fig. S7**. **(c)** The plot shows the cumulative fraction of PhastCons High_notCNE bases at each distance to the nearest CNE base up to 3,000 bp. **(d)** The plot shows the fold enrichment for the cumulative number of PhastCons High_notCNE bases being within each distance to the nearest CNE base, up to 3,000bp. Similar plots to (a) and (b) for other constrained element sets can be found in **Supplementary Fig. S8**.

Figure 5: CNEP score's relationship to human variation, TF sequence motifs, and DNase I Hypersensitive Sites in mouse. (a) The plot shows for PhastCons High_CNE, CNE, Low_CNE, High_notCNE, notCNE, and Low_CNE bases the proportional site frequency spectrum based on a set of 105 unrelated individuals in the YRI population in terms of # SNPs per base pair eligible for a SNP to be called, normalized for the number of sites with a variant in each set (**Methods**). The last column includes all SNPs with minor allele count greater than 10. **(b)** Similar plot to (a), except showing the absolute site frequency spectrum per base pair eligible for a SNP to be called normalized by estimated mutation rates. Corresponding plots for additional constrained elements can be found in **Supplementary Fig. S9** and plots controlling for difference in background selection can be found in **Supplementary Fig. S10**. Plots at higher thresholds of the CNEP score for notCNE bases can be found in **Supplementary Fig. S11**. **(c)** The plot shows the difference of the distribution of motif enrichments relative to the distribution for a randomized set of the motifs for the PhastCons High_CNE, CNE, Low_CNE, High_notCNE, notCNE, and Low_CNE bases. The x-axis is the rank position of the motif among the 1,646 motifs. The y-axis is the difference between the $\log_2$ fold enrichment based on the actual motif calls and the median $\log_2$ fold enrichment from three randomized versions at the same rank position (**Methods**). Similar plots for other constrained elements can be found in **Supplementary Fig. S12** and at other thresholds for defining notCNE high bases in **Supplementary Fig. S13**. **(d)** Scatter plot of individual motif enrichments. The x and y axes corresponds to the $\log_2$ fold enrichments in PhastCons Low_CNE and High_CNE bases respectively. The blue lines separate the three regions used for the GO enrichment analysis, High_CNE strongly preferred, High_CNE moderately preferred, and Low_CNE preferred, where at least one of the Low_CNE or the High_CNE $\log_2$ enrichment is greater than or equal to 0.5 (**Supplementary Fig. S4**). The gray line is the y=x line where both Low_CNE and High_CNE $\log_2$ enrichments are less than 0.5. Similar plots based on other thresholds of the CNEP score can be found in **Supplementary Fig. S14**. Selected motifs are labeled. **(e)** The distribution of enrichments for DNase I Hypersensitive Sites (DHS) from 156 experiments in mouse, where the sites are mapped to human and enrichments are computed relative to enrichments for a randomized DHS, for PhastCons High_CNE, CNE, Low_CNE, High_notCNE, notCNE, and Low_CNE bases (**Methods**). Similar plots for other constrained elements can be found in **Supplementary Fig. S15**. **(f)** A bar graph corresponding to the enrichments shown in (e) for Low_CNE bases. Bars are colored to indicate if the experiment is of whole brain or cerebrum (red), embryonic day 11.5 (dark blue), or neither (gray). Similar plots for other constrained

elements can be found in **Supplementary Fig. S16**. A table of the enrichment values can be found in **Supplementary Table S6**.

Figure 6: Analysis on additional human datasets. Plots for **(a)** ChIP-atlas **(b)** ENCODE portal, and **(c)** ReMap 2018 showing the distribution of prediction underestimate values for datasets with peaks covering at least 200kb. The prediction underestimate value for a dataset is the average difference between the expected CNEP score (Methods) and the prediction value for each base covered by a peak. Results are shown for prediction values based on the genome-wide average expected CNEP score (blue) and the CNEP score (red). Also shown is the distribution of using the CNEP score for the prediction values, but applied to a shuffled version of each dataset (green). **(d)** Scatter plot where each point corresponds to a dataset, the x-axis is the number of bases it covers, and the y-axis is the prediction underestimate value when using the CNEP score for prediction values. Selected datasets with a high combination of base coverage and underestimate values are labeled or placed in a box if they correspond to a DNase I hypersensitive experiment of embryonic brain, spinal cord, or eye. The color of the box corresponds to brain, spinal cord, or eye as indicated in the legend. The color and shape of the points are based on whether the point corresponds to a ChIP-atlas, ENCODE portal, or ReMap 2018 dataset or the set of new GENCODE exons between v19 and v28. Only datasets with a positive underestimate value are shown. Datasets covering more than 200 million base pairs are not shown, but all had an underestimate value of less than 0.01. Three datasets that had an underestimate value greater than 0.13 are not shown, but all covered less than 9,000 base pairs. Versions of these plots based on the input features to CNEP and based on shuffled versions of the additional datasets can be found in **Supplementary Fig. S18**.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Arneson, A., and Ernst, J. (2019). Systematic discovery of conservation states for single-nucleotide annotation of the human genome. Commun. Biol. *2*, 248.

Boeva, V., Louis-Brennetot, C., Peltier, A., Durand, S., Pierre-Eugène, C., Raynal, V., Etchevers, H.C., Thomas, S., Lermine, A., Daudigeos-Dubus, E., et al. (2017). Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries. Nat. Genet. *49*, 1408–1413.

Carlson, J., Locke, A.E., Flickinger, M., Zawistowski, M., Levy, S., Myers, R.M., Boehnke, M., Kang, H.M., Scott, L.J., Li, J.Z., et al. (2018). Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nat. Commun. *9*, 3753.

Cebola, I., Rodríguez-Seguí, S.A., Cho, C.H.-H., Bessa, J., Rovira, M., Luengo, M., Chhatriwala, M., Berry, A., Ponsa-Cobas, J., Maestro, M.A., et al. (2015). TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. Nat. Cell Biol. *17*, 615–626.

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. *46*, D267–D275.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol *6*, e1001025.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics *7*, 191.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Ernst, J., and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. *23*, 1142–1154.

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. *33*, 364–376.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. J. Mach. Learn. Res. *9*, 1871–1874.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54–i62.

Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., and Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Nucleic Acids Res. *43*, e27–e27.

Gulko, B., and Siepel, A. (2018). An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences. Nat. Genet.

Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genet. *5*, e1000695.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. *111*, 6131–6138.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Res. *39*, D19–D21.

Li, Y., Quang, D., and Xie, X. (2017). Understanding Sequence Conservation With Deep Learning. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, (New York, NY, USA: ACM), pp. 400–406.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. *17*, 760–774.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread Genomic Signatures of Natural Selection in Hominid Evolution. PLOS Genet. *5*, e1000471.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., Kawaji, H., Nakaki, R., Sese, J., and Meno, C. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. EMBO Rep. e46255.

Pomerantz, M.M., Li, F., Takeda, D.Y., Lenci, R., Chonkar, A., Chabot, M., Cejas, P., Vazquez, F., Cook, J., Shivdasani, R.A., et al. (2015). The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. Nat. Genet. *47*, 1346–1351.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. *26*, 841–842.

Rajbhandari, P., Lopez, G., Capdevila, C., Salvatori, B., Yu, J., Rodriguez-Barrueco, R., Martinez, D., Yarmarkovich, M., Weichert-Leahey, N., Abraham, B.J., et al. (2018). Cross-Cohort Analysis Identifies a TEAD4–MYCN Positive Feedback Loop as the Core Regulatory Element of High-Risk Neuroblastoma. Cancer Discov. *8*, 582–599.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLOS Genet. *10*, e1004525.

Ritchie, G.R.S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. Nat. Methods *11*, 294–296.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. gku1177.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Teo, A.K.K., Tsuneyoshi, N., Hoon, S., Tan, E.K., Stanton, L.W., Wright, C.V.E., and Dunn, N.R. (2015). PDX1 Binds and Represses Hepatic Genes to Ensure Robust Pancreatic Commitment in Differentiating Human Embryonic Stem Cells. Stem Cell Rep. *4*, 578–590.

de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., and Geschwind, D.H. (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell *172*, 289-304.e18.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science *346*, 1007–1012.

Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N.A., Muth, K., Palmer, J., Qiu, Y., Wang, J., et al. (2015). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell Stem Cell *16*, 386–399.

Wang, X., Choi, J.-H., Ding, J., Yang, L., Ngoka, L.C., Lee, E.J., Zha, Y., Mao, L., Jin, B., Ren, M., et al. (2013). HOXC9 directly regulates distinct sets of genes to coordinate diverse cellular processes during neuronal differentiation. BMC Genomics *14*, 830.

Ward, L.D., and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science *337*, 1675–1678.

Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat. Genet. *46*, 61–64.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

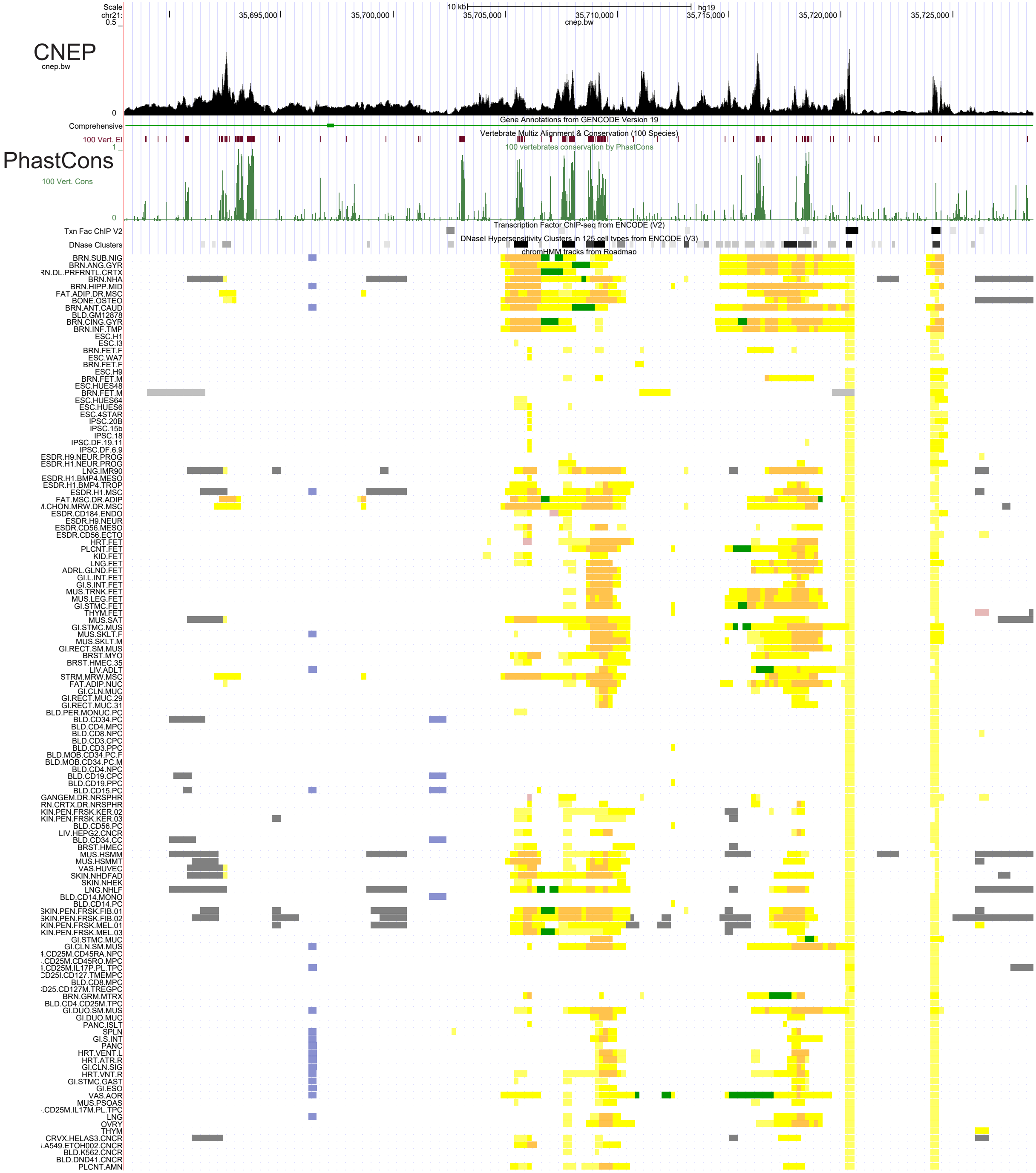

Scale
chr21:
0.5
10 kb
hg19
35,695,000
35,700,000
35,705,000
35,710,000
35,715,000
35,720,000
35,725,000
cnep.bw
CNEP
cnep.bw
0
Comprehensive
Gene Annotations from GENCODE Version 19
100 Vert. El
Vertebrate Multiz Alignment & Conservation (100 Species)
1
PhastCons
100 vertebrates conservation by PhastCons
100 Vert. Cons
0
Txn Fac ChIP V2
Transcription Factor ChIP-seq from ENCODE (V2)
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
chromHMM tracks from Roadmap
BRN.SUB.NIG
BRN.ANG.GYR
RN.DL.PRFRNTL.CRTX
BRN.NHA
BRN.HIPP.MID
FAT.ADIP.DR.MSC
BONE.OSTEO
BRN.ANT.CAUD
BLD.GM12878
BRN.CING.GYR
BRN.INF.TMP
ESC.H1
ESC.I3
BRN.FET.F
ESC.WA7
BRN.FET.F
ESC.H9
BRN.FET.M
ESC.HUES48
BRN.FET.M
ESC.HUES64
ESC.HUES6
ESC.4STAR
IPSC.20B
IPSC.15b
IPSC.18
IPSC.DF.19.11
IPSC.DF.6.9
ESDR.H9.NEUR.PROG
ESDR.H1.NEUR.PROG
LNG.IMR90
ESDR.H1.BMP4.MESO
ESDR.H1.BMP4.TROP
ESDR.H1.MSC
FAT.MSC.DR.ADIP
I.CHON.MRW.DR.MSC
ESDR.CD184.ENDO
ESDR.H9.NEUR
ESDR.CD56.MESO
ESDR.CD56.ECTO
HRT.FET
PLCNT.FET
KID.FET
LNG.FET
ADRL.GLND.FET
GI.L.INT.FET
GI.S.INT.FET
MUS.TRNK.FET
MUS.LEG.FET
GI.STMC.FET
THYM.FET
MUS.SAT
GI.STMC.MUS
MUS.SKLT.F
MUS.SKLT.M
GI.RECT.SM.MUS
BRST.MYO
BRST.HMEC.35
LIV.ADLT
STRM.MRW.MSC
FAT.ADIP.NUC
GI.CLN.MUC
GI.RECT.MUC.29
GI.RECT.MUC.31
BLD.PER.MONUC.PC
BLD.CD34.PC
BLD.CD4.MPC
BLD.CD8.NPC
BLD.CD3.CPC
BLD.CD3.PPC
BLD.MOB.CD34.PC.F
BLD.MOB.CD34.PC.M
BLD.CD4.NPC
BLD.CD19.CPC
BLD.CD19.PPC
BLD.CD15.PC
GANGEM.DR.NRSPHR
RN.CRTX.DR.NRSPHR
KIN.PEN.FRSK.KER.02
KIN.PEN.FRSK.KER.03
BLD.CD56.PC
LIV.HEPG2.CNCR
BLD.CD34.CC
BRST.HMEC
MUS.HSMM
MUS.HSMMT
VAS.HUVEC
SKIN.NHDFAD
SKIN.NHEK
LNG.NHLF
BLD.CD14.MONO
BLD.CD14.PC
SKIN.PEN.FRSK.FIB.01
SKIN.PEN.FRSK.FIB.02
KIN.PEN.FRSK.MEL.01
KIN.PEN.FRSK.MEL.03
GI.STMC.MUC
GI.CLN.SM.MUS
4.CD25M.CD45RA.NPC
.CD25M.CD45RO.MPC
4.CD25M.IL17P.PL.TPC
CD25I.CD127.TMEMPC
BLD.CD8.MPC
D25.CD127M.TREGPC
BRN.GRM.MTRX
BLD.CD4.CD25M.TPC
GI.DUO.SM.MUS
GI.DUO.MUC
PANC.ISLT
SPLN
GI.S.INT
PANC
HRT.VENT.L
HRT.ATR.R
GI.CLN.SIG
HRT.VNT.R
GI.STMC.GAST
GI.ESO
VAS.AOR
MUS.PSOAS
.CD25M.IL17M.PL.TPC
LNG
OVRY
THYM
CRVX.HELAS3.CNCR
.A549.ETOH002.CNCR
BLD.K562.CNCR
BLD.DND41.CNCR
PLCNT.AMN

Figure 1

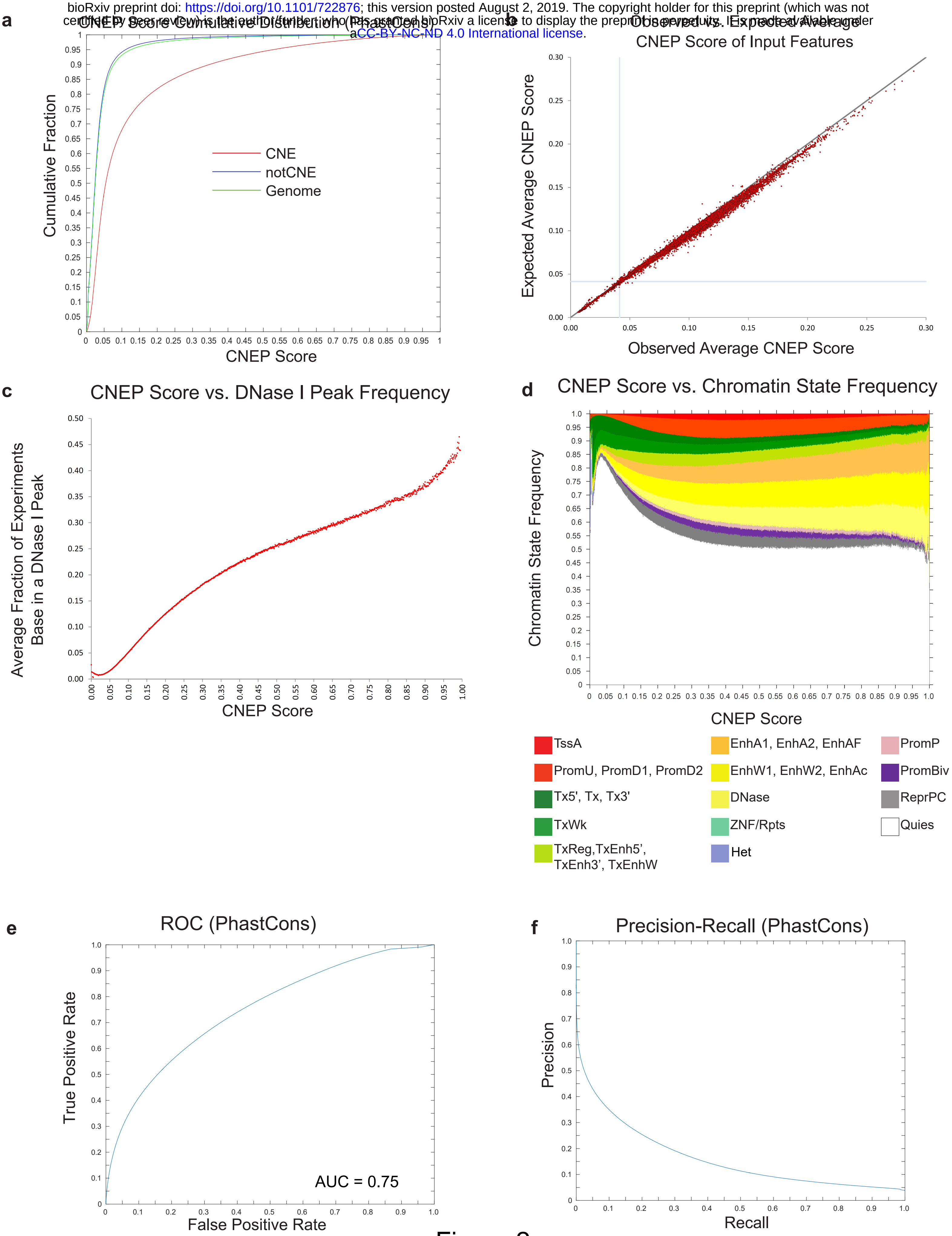
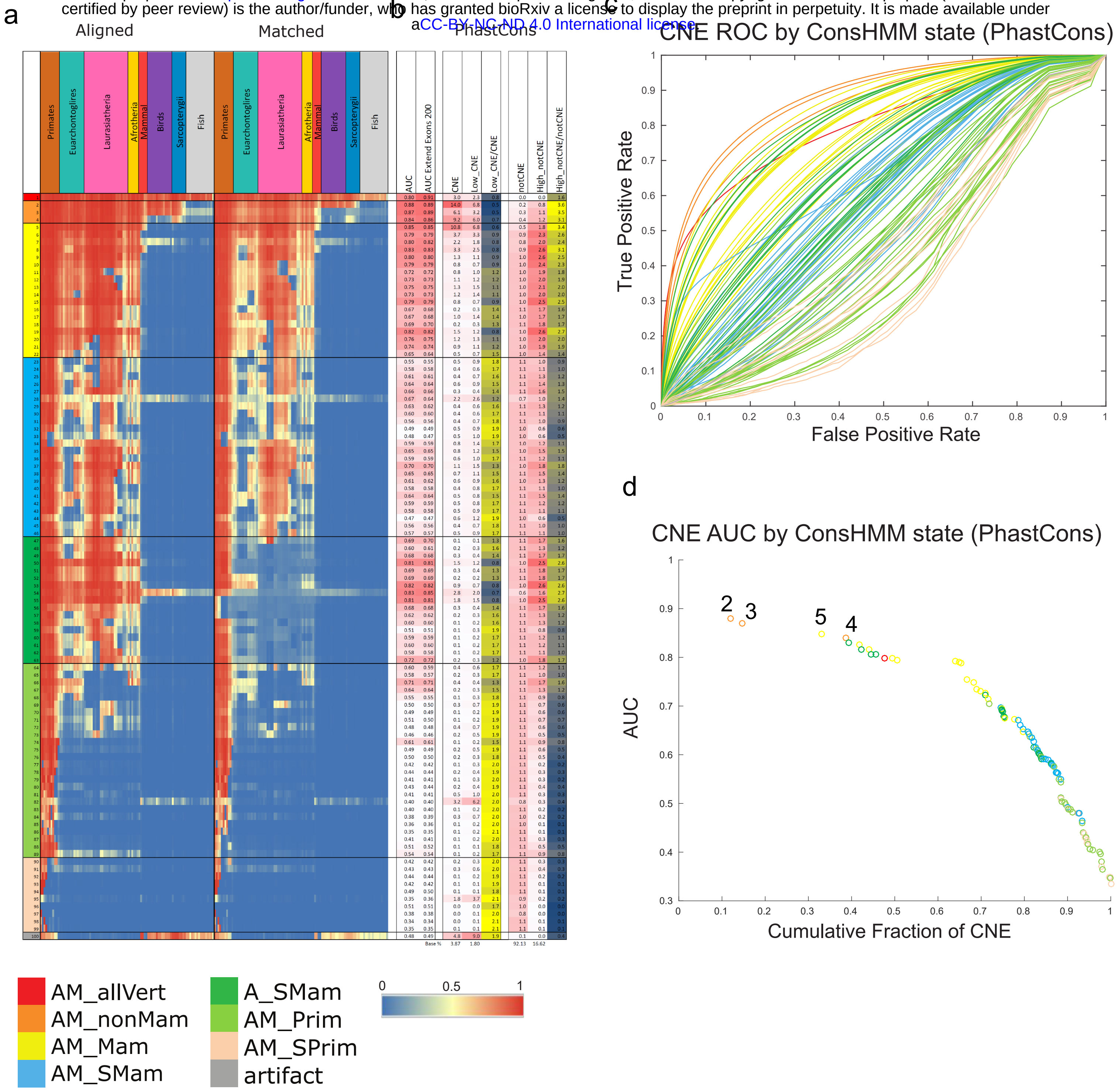

a
CNEP Score Cumulative Distribution (PhastCons)
Cumulative Fraction
CNEP Score
CNE
notCNE
Genome
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
b
Observed vs. Expected Average CNEP Score of Input Features
Expected Average CNEP Score
Observed Average CNEP Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
c
CNEP Score vs. DNase I Peak Frequency
Average Fraction of Experiments Base in a DNase I Peak
CNEP Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
d
CNEP Score vs. Chromatin State Frequency
Chromatin State Frequency
CNEP Score
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1.0
TssA
PromU, PromD1, PromD2
Tx5', Tx, Tx3'
TxWk
TxReg,TxEnh5', TxEnh3', TxEnhW
EnhA1, EnhA2, EnhAF
EnhW1, EnhW2, EnhAc
DNase
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
e
ROC (PhastCons)
True Positive Rate
False Positive Rate
AUC = 0.75
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
f
Precision-Recall (PhastCons)
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 2



	AUC	AUC Extend Exons 200	CNE	Low_CNE	Low_CNE/CNE	notCNE	High_notCNE	High_notCNE/notCNE
1	0.80	0.91	3.0	2.3	0.8	0.0	0.0	1.6
2	0.88	0.89	14.0	6.8	0.5	0.2	0.8	3.6
3	0.87	0.89	6.1	3.2	0.5	0.3	1.1	3.5
4	0.84	0.86	9.2	6.0	0.7	0.4	1.2	3.1
5	0.85	0.85	10.8	6.8	0.6	0.5	1.8	3.4
6	0.79	0.79	3.7	3.3	0.9	0.9	2.3	2.6
7	0.80	0.82	2.2	1.8	0.8	0.8	2.0	2.4
8	0.83	0.83	3.3	2.5	0.8	0.9	2.6	3.1
9	0.80	0.80	1.3	1.1	0.9	1.0	2.6	2.5
10	0.79	0.79	0.8	0.7	0.9	1.0	2.4	2.3
11	0.72	0.72	0.8	1.0	1.2	1.0	1.9	1.8
12	0.73	0.73	1.1	1.2	1.2	1.0	2.0	1.9
13	0.75	0.75	1.3	1.5	1.1	1.0	2.1	2.0
14	0.73	0.73	1.2	1.4	1.1	1.0	2.0	2.0
15	0.79	0.79	0.8	0.7	0.9	1.0	2.5	2.5
16	0.67	0.68	0.2	0.3	1.4	1.1	1.7	1.6
17	0.67	0.68	1.0	1.4	1.4	1.0	1.7	1.7
18	0.69	0.70	0.2	0.3	1.3	1.1	1.8	1.7
19	0.82	0.82	1.5	1.2	0.8	1.0	2.6	2.7
20	0.76	0.75	1.2	1.3	1.1	1.0	2.0	2.0
21	0.74	0.74	0.9	1.1	1.2	1.0	1.9	1.9
22	0.65	0.64	0.5	0.7	1.5	1.0	1.4	1.4
23	0.55	0.55	0.5	0.9	1.8	1.1	1.0	0.9
24	0.58	0.58	0.4	0.6	1.7	1.1	1.1	1.0
25	0.61	0.61	0.4	0.7	1.6	1.1	1.3	1.2
26	0.64	0.64	0.6	0.9	1.5	1.1	1.4	1.3
27	0.66	0.66	0.3	0.4	1.4	1.1	1.6	1.5
28	0.67	0.64	2.2	2.6	1.2	0.7	1.0	1.4
29	0.63	0.62	0.4	0.6	1.6	1.1	1.3	1.2
30	0.60	0.60	0.4	0.6	1.7	1.1	1.1	1.1
31	0.56	0.56	0.4	0.7	1.8	1.1	1.0	0.9
32	0.49	0.49	0.5	0.9	1.9	1.0	0.6	0.6
33	0.48	0.47	0.5	1.0	1.9	1.0	0.6	0.5
34	0.59	0.59	0.8	1.4	1.7	1.0	1.2	1.1
35	0.65	0.65	0.8	1.2	1.5	1.0	1.5	1.5
36	0.59	0.59	0.6	1.0	1.7	1.1	1.2	1.1
37	0.70	0.70	1.1	1.5	1.3	1.0	1.8	1.8
38	0.65	0.65	0.7	1.1	1.5	1.1	1.5	1.4
39	0.61	0.62	0.6	0.9	1.6	1.0	1.3	1.2
40	0.58	0.58	0.4	0.8	1.7	1.1	1.1	1.0
41	0.64	0.64	0.5	0.8	1.5	1.1	1.5	1.4
42	0.59	0.59	0.5	0.8	1.7	1.1	1.2	1.2
43	0.58	0.58	0.5	0.9	1.7	1.1	1.1	1.1
44	0.47	0.47	0.6	1.2	1.9	1.0	0.6	0.5
45	0.56	0.56	0.4	0.7	1.8	1.1	1.0	1.0
46	0.57	0.57	0.5	0.9	1.7	1.1	1.1	1.0
47	0.69	0.70	0.1	0.1	1.3	1.1	1.7	1.6
48	0.60	0.61	0.2	0.3	1.6	1.1	1.3	1.2
49	0.68	0.68	0.3	0.4	1.4	1.1	1.7	1.7
50	0.81	0.81	1.5	1.2	0.8	1.0	2.5	2.6
51	0.69	0.69	0.3	0.4	1.3	1.1	1.8	1.7
52	0.69	0.69	0.2	0.2	1.3	1.1	1.8	1.7
53	0.82	0.82	0.9	0.7	0.8	1.0	2.6	2.6
54	0.83	0.85	2.8	2.0	0.7	0.6	1.6	2.7
55	0.81	0.81	1.8	1.5	0.8	1.0	2.5	2.6
56	0.68	0.68	0.3	0.4	1.4	1.1	1.7	1.6
57	0.62	0.62	0.2	0.3	1.6	1.1	1.3	1.2
58	0.60	0.60	0.1	0.2	1.6	1.1	1.3	1.2
59	0.51	0.51	0.1	0.3	1.9	1.1	0.8	0.8
60	0.59	0.59	0.1	0.2	1.7	1.1	1.2	1.1
61	0.60	0.60	0.1	0.2	1.7	1.1	1.2	1.1
62	0.58	0.58	0.1	0.2	1.7	1.1	1.1	1.0
63	0.72	0.72	0.2	0.3	1.2	1.0	1.8	1.7
64	0.60	0.59	0.4	0.6	1.7	1.1	1.2	1.1
65	0.58	0.57	0.2	0.3	1.7	1.1	1.0	1.0
66	0.71	0.71	0.4	0.4	1.3	1.1	1.7	1.6
67	0.64	0.64	0.2	0.3	1.5	1.1	1.3	1.2
68	0.55	0.55	0.1	0.3	1.8	1.1	0.9	0.8
69	0.50	0.50	0.3	0.7	1.9	1.1	0.7	0.6
70	0.49	0.49	0.1	0.2	1.9	1.1	0.6	0.6
71	0.51	0.50	0.1	0.2	1.9	1.1	0.7	0.7
72	0.48	0.48	0.4	0.7	1.9	1.1	0.6	0.6
73	0.46	0.46	0.2	0.5	1.9	1.1	0.5	0.5
74	0.61	0.61	0.1	0.2	1.5	1.1	0.9	0.8
75	0.49	0.49	0.2	0.5	1.9	1.1	0.6	0.5
76	0.50	0.50	0.2	0.3	1.8	1.1	0.5	0.4
77	0.42	0.42	0.1	0.3	2.0	1.1	0.2	0.2
78	0.44	0.44	0.2	0.4	1.9	1.1	0.3	0.3
79	0.41	0.41	0.1	0.3	2.0	1.1	0.3	0.2
80	0.43	0.44	0.2	0.4	1.9	1.1	0.4	0.4
81	0.41	0.41	0.5	1.0	2.0	1.1	0.3	0.3
82	0.40	0.40	3.2	6.2	2.0	0.8	0.3	0.4
83	0.40	0.40	0.1	0.2	2.0	1.1	0.2	0.2
84	0.38	0.39	0.3	0.7	2.0	1.0	0.2	0.2
85	0.36	0.36	0.1	0.2	2.0	1.0	0.1	0.1
86	0.35	0.35	0.1	0.2	2.1	1.1	0.1	0.1
87	0.41	0.41	0.1	0.2	2.0	1.1	0.3	0.3
88	0.51	0.52	0.1	0.1	1.8	1.1	0.5	0.5
89	0.54	0.54	0.1	0.2	1.7	1.1	0.9	0.8
90	0.42	0.42	0.2	0.3	2.0	1.1	0.3	0.3
91	0.43	0.43	0.3	0.6	2.0	1.1	0.4	0.3
92	0.44	0.44	0.1	0.2	1.9	1.1	0.2	0.2
93	0.42	0.42	0.1	0.1	1.9	1.1	0.1	0.1
94	0.49	0.50	0.1	0.1	1.8	1.1	0.1	0.1
95	0.35	0.36	1.8	3.7	2.1	0.9	0.2	0.2
96	0.51	0.51	0.0	0.0	1.7	1.0	0.0	0.0
97	0.38	0.38	0.0	0.1	2.0	0.8	0.0	0.0
98	0.34	0.34	0.0	0.1	2.1	1.1	0.1	0.1
99	0.35	0.35	0.1	0.1	2.1	1.1	0.1	0.1
100	0.48	0.49	4.8	9.0	1.9	0.1	0.0	0.4
		Base %	3.87	1.80		92.13	16.62	

Figure 3

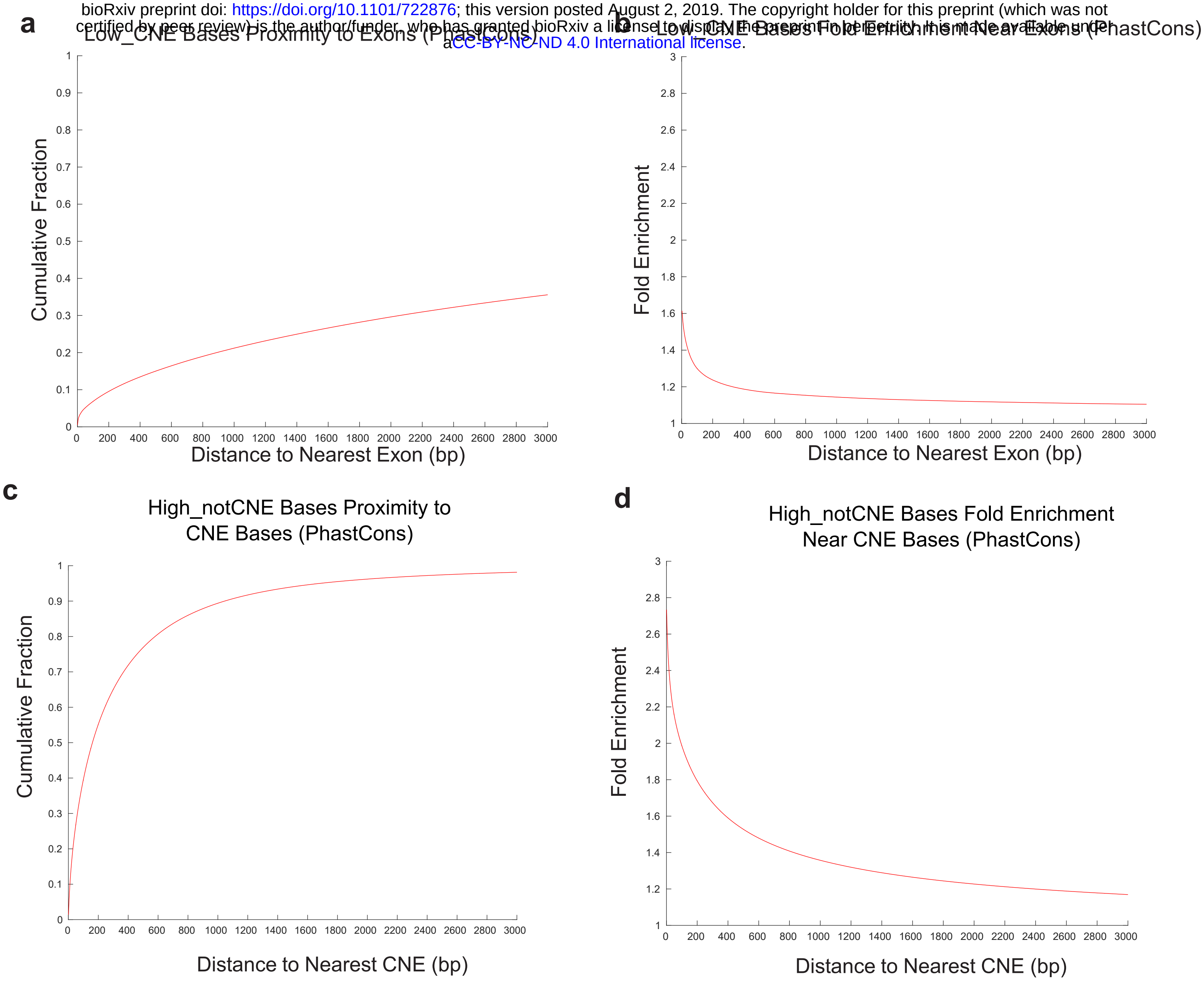

a
Low_CNE Bases Proximity to Exons (PhastCons)
Cumulative Fraction
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000
Distance to Nearest Exon (bp)
b
Low_CNE Bases Fold Enrichment Near Exons (PhastCons)
Fold Enrichment
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000
Distance to Nearest Exon (bp)
c
High_notCNE Bases Proximity to CNE Bases (PhastCons)
Cumulative Fraction
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000
Distance to Nearest CNE (bp)
d
High_notCNE Bases Fold Enrichment Near CNE Bases (PhastCons)
Fold Enrichment
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000
Distance to Nearest CNE (bp)

Figure 4

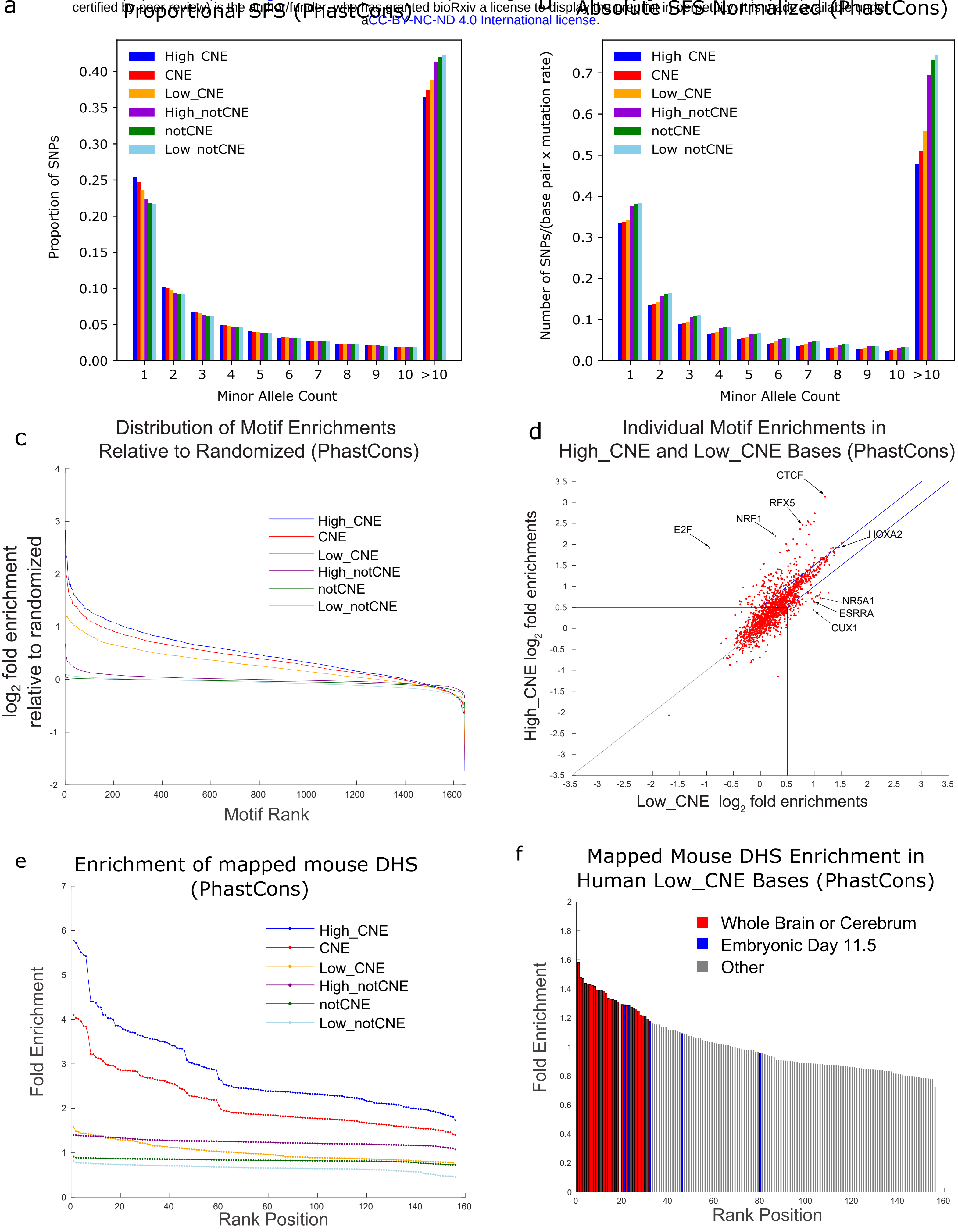

a
Proportional SFS (PhastCons)
Proportion of SNPs
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
High_CNE
CNE
Low_CNE
High_notCNE
notCNE
Low_notCNE
1
2
3
4
5
6
7
8
9
10
>10
Minor Allele Count
b
Absolute SFS Normalized (PhastCons)
Number of SNPs/(base pair x mutation rate)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
High_CNE
CNE
Low_CNE
High_notCNE
notCNE
Low_notCNE
1
2
3
4
5
6
7
8
9
10
>10
Minor Allele Count
c
Distribution of Motif Enrichments
Relative to Randomized (PhastCons)
$\log_2$ fold enrichment
relative to randomized
-2
-1
0
1
2
3
4
High_CNE
CNE
Low_CNE
High_notCNE
notCNE
Low_notCNE
0
200
400
600
800
1000
1200
1400
1600
Motif Rank
d
Individual Motif Enrichments in
High_CNE and Low_CNE Bases (PhastCons)
High_CNE $\log_2$ fold enrichments
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
CTCF
RFX5
NRF1
E2F
HOXA2
NR5A1
ESRRA
CUX1
Low_CNE $\log_2$ fold enrichments
e
Enrichment of mapped mouse DHS
(PhastCons)
Fold Enrichment
0
1
2
3
4
5
6
7
High_CNE
CNE
Low_CNE
High_notCNE
notCNE
Low_notCNE
0
20
40
60
80
100
120
140
160
Rank Position
f
Mapped Mouse DHS Enrichment in
Human Low_CNE Bases (PhastCons)
Fold Enrichment
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
Whole Brain or Cerebrum
Embryonic Day 11.5
Other
0
20
40
60
80
100
120
140
160
Rank Position

Figure 5

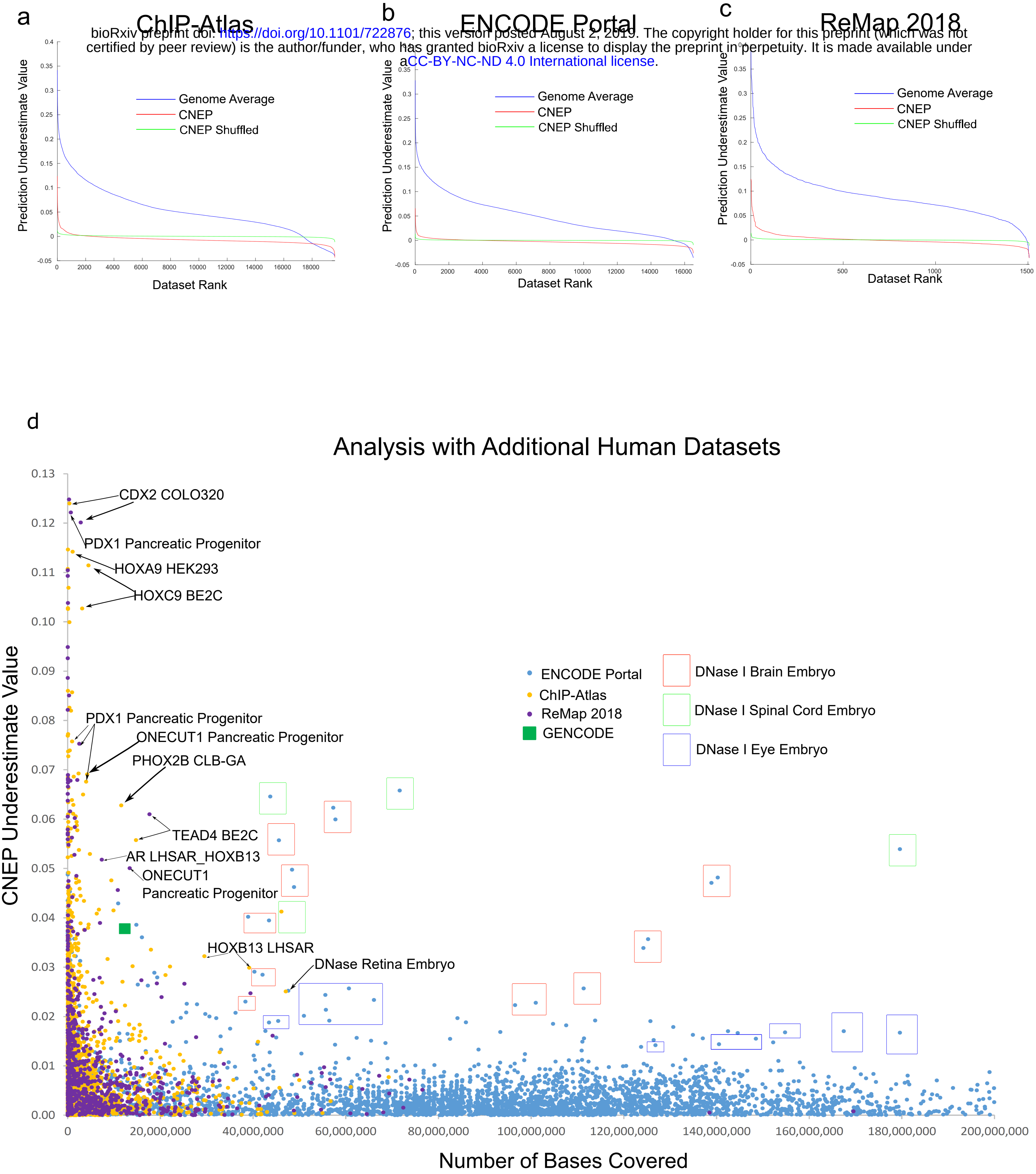

a
ChIP-Atlas
Prediction Underestimate Value
Genome Average
CNEP
CNEP Shuffled
Dataset Rank
b
ENCODE Portal
Prediction Underestimate Value
Genome Average
CNEP
CNEP Shuffled
Dataset Rank
c
ReMap 2018
Prediction Underestimate Value
Genome Average
CNEP
CNEP Shuffled
Dataset Rank
d
Analysis with Additional Human Datasets
CNEP Underestimate Value
Number of Bases Covered
ENCODE Portal
ChIP-Atlas
ReMap 2018
GENCODE
DNase I Brain Embryo
DNase I Spinal Cord Embryo
DNase I Eye Embryo
CDX2 COLO320
PDX1 Pancreatic Progenitor
HOXA9 HEK293
HOXC9 BE2C
PDX1 Pancreatic Progenitor
ONECUT1 Pancreatic Progenitor
PHOX2B CLB-GA
TEAD4 BE2C
AR LHSAR_HOXB13
ONECUT1
Pancreatic Progenitor
HOXB13 LHSAR
DNase Retina Embryo
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0
20,000,000
40,000,000
60,000,000
80,000,000
100,000,000
120,000,000
140,000,000
160,000,000
180,000,000
200,000,000

Figure 6