A novel functional genomics atlas coupled with convolutional neural networks facilitates clinical interpretation of disease relevant variants in non-coding regulatory elements

Ruizhi Deng[1,*], Elena Perenthaler[1,*†], Anita Nikoncuk[1,^], Soheil Yousefi[1,^], Kristina Lanko[1], Rachel Schot[1], Michela Maresca[1], Michael J. Parker[2], Wilfred F.J. van Ijcken[3,4], Joohyun Park[5], Marc Sturm[5], Tobias B. Haack[5,6], Genomics England Research Consortium, Gennady V Roshchupkin[7,8], Eskeatnaf Mulugeta[4] and Tahsin Stefan Barakat[1, #]

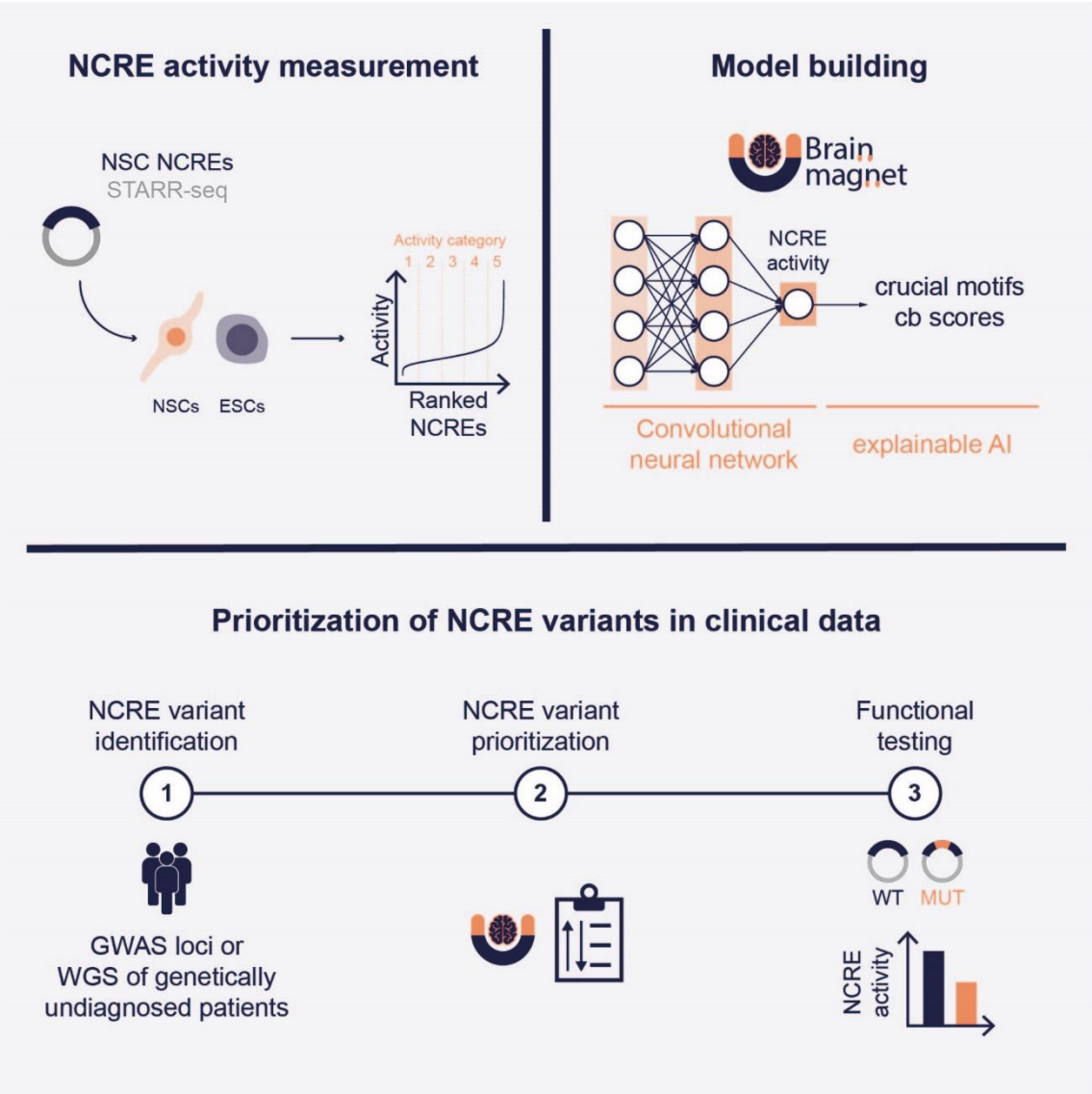
[1] Department of Clinical Genetics, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[2] Sheffield Children's NHS Foundation Trust, Sheffield, UK.
[3] Center for Biomics, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[4] Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[5] Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany
[6] Centre for Rare Diseases, University of Tübingen, Tübingen, Germany
[7] Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands
[8] Department of Epidemiology, Erasmus MC University Medical Center, Rotterdam, The Netherlands
* These authors contributed equally as first authors.
^These authors contributed equally as second authors.
† current affiliation: Institute of Biophysics, CNR, Trento, Italy
#author for correspondence: t.barakat@erasmusmc.nl

Abstract

Genome-wide assessment of genetic variation is becoming a routine in human genetics, but functional interpretation of non-coding variants both in common and rare diseases remains extremely challenging. Here we employed the massively parallel reporter assay ChIP-STARR-seq to functionally annotate activity of >140 thousand non-coding regulatory elements (NCREs) in human neural stem cells (NSCs) as a model for early brain development. Highly active NCREs show an increasing sequence constraint and harbour *de novo* variants in individuals affected by neurodevelopmental disorders. They are enriched for transcription factor (TF) motifs including YY1 and p53 family members and for the presence of primate-specific transposable elements, providing insights on gene regulatory mechanisms in NSCs. Examining episomal NCRE activity of the same sequences in human embryonic stem cells (ESCs) identified cell type differential activity and primed NCREs, accompanied by a rewiring of the epigenome landscape. Leveraging on the experimentally measured NCRE activity and nucleotide composition of the assessed sequences, we build BRAIN-MAGNET, a convolutional neural network that allows the prediction of NCRE activity based on DNA sequence composition, and which identifies functionally relevant nucleotides and TF motifs

within each NCRE that are required for NCRE function. The application of BRAIN-MAGNET including its functional validation allows fine-mapping of GWAS loci identified for common neurological traits, and prioritization of possible disease causing rare non-coding variants in currently genetically unexplained individuals with neurogenetic disorders, including those from the Genomics England 100,000 Genomes project. We foresee that this NCRE atlas and BRAIN-MAGNET will help reducing missing heritability in human genetics, by limiting the search space for functional relevant non-coding genetic variation.

Keywords: gene regulation, non-coding genome, enhancer, massively parallel reporter assay, Genomics England 100,000 Genomes Project, GWAS, whole genome sequencing, artificial intelligence, convolutional neural network, neurogenetics, neurodevelopmental disorders, diagnostics.



Introduction

Around 98% of the human genome does not directly encode protein coding genes and contains by far most of genetic variation (Chen et al., 2024). Although rapid advances in genomic sequencing technologies have made the identification of this genetic variation rather trivial, interpreting functionality of non-coding genomic sequences and effects of variants herein remains tremendously challenging (Chen *et al.*, 2024). This is exemplified by the plethora of common single nucleotide polymorphisms (SNPs) associated with disease traits by genome-wide association studies (GWAS) for which the underlying biological function and direct effects on gene regulation have not been resolved (Maurano et al., 2012). Another example concerns the rare disease field, where current genomic diagnostic testing modalities even when using whole genome sequencing (WGS) mostly focus on interpreting potential disease causing variants directly affecting protein coding genes (Wright et al., 2018). Such approaches at best identify a genetic cause for rare disorders in 30-50% of affected individuals (Clark et al., 2018; Smith et al., 2019). Given that in a clinical setting most of the non-coding genetic variation is not routinely assessed, and an increasing amount of evidence demonstrates that genetic alterations of non-coding regulatory elements (NCREs) such as enhancers can cause human disorders (Carullo and Day, 2019; Perenthaler et al., 2019; Spiess and Won, 2020), it seems likely that a considerable amount of missing heritability in human genetics might be caused by non-coding genetic variation. To resolve this missing heritability, it is crucial to functionally annotate NCREs, characterize their activity in various cell types and to develop novel approaches that help to predict the impact of genomic variants on NCRE function.

Here we employed the massively parallel reporter assay ChIP-STARR-seq (Carullo and Day, 2019) to functionally annotate NCREs at high throughput and measure their activity in neural stem cells (NSCs), as a model system for developing brain cells. Our in-depth characterisation revealed that increased NCRE activity is associated with increased expression and likelihood of loss-of-function intolerance of the target genes. Highly active NSC NCREs are characterized by increased sequence constraint and are enriched for TF binding motifs including YY1 and p53 family members and primate specific MER61 and LTR10 transposable elements, providing insights into the mechanisms underlying gene regulation in NSCs. Functional examination of the same genomic regions in human embryonic stem cells (ESCs) identified differential NCRE activity between the two cell types that is accompanied by a rewiring of the epigenome landscape, and unravelled NCREs that are primed for activation upon neural differentiation. Finally, we developed and functionally validated BRAIN-MAGNET, a convolutional neural network model trained on the experimentally measured NCRE activity data to predict nucleotides and motifs required for NCRE activity solely based on DNA

composition. We applied BRAIN-MAGNET to prioritize the effects of genomic variants in the context of brain related neurological traits from GWAS data and to identify likely disease-causing non-coding variants in individuals with rare diseases, including those from the Genomics England 100,000 Genomes Project (Caulfield, 2017). Together, the herein created NCRE atlas will facilitate reducing missing heritability in human genetics.

Results

ChIP-STARR-seq in neural stem cells identifies NCREs with different activity levels

To functionally identify NCREs genome-wide in human ESCs, we previously used the massively parallel reporter assay (MPRA) ChIP-STARR-seq (Barakat et al., 2018), herein referred to as ESC ChIP-STARR-seq. In this assay, putative NCREs obtained upon chromatin immunoprecipitation (ChIP) are cloned in the STARR-seq reporter plasmid (Arnold et al., 2013), downstream of a minimal-promoter driven GFP and upstream of the polyadenylation signal. The minimal promoter alone is insufficient to drive GFP expression. However, if the cloned sequence is a functional NCRE, it will enhance GFP transcription and, being located upstream of the polyadenylation signal, it will also be self-transcribed. Thus, performing RNA-seq on GFP-containing mRNAs synthesised upon plasmid cell transfection allows the identification of sequences with functional NCRE activity in the tested cell type.

Here, we applied an adapted ChIP-STARR-seq approach (see **Methods, Figure 1A, Figure S1,** and **Supplementary Note 1** for details) in human NSCs, referred to as NSC ChIP-STARR-seq. These cells are multipotent stem cells reminiscent of the stem cell population in the ventricular zone during human brain development (Zhao and Moore, 2018) and can proliferate and differentiate into multiple neural cell types. Characterizing the functionally active NCRE landscape of NSCs can thus likely identify NCREs involved in early stages of human neurodevelopment, which might be relevant for human neurodevelopmental disorders. For ChIP experiments and plasmid library generation, we focussed on the TFs YY1 and SOX2, that are important in NSCs (Miyagi et al., 2009; Zurkirchen et al., 2019), and on the histone modifications H3K4me1 and H3K27ac which are associated with putative NCREs (Rada-Iglesias et al., 2011). This allowed us to enrich for sequences that likely reflect active chromatin at the endogenous loci in NSCs even when tested on an episomal MPRA.

To identify active NCREs, we first generated an analysis scaffold consisting of all merged ChIP-seq and plasmid DNA-seq peaks (148,198 genomic regions in total, see **Methods**). Large merged sequences were split in scaffold regions of a maximum of 1000 bp, which were approximately equally divided over all human chromosomes, except for the Y chromosome, not present in the female cell line used for the experiments (**Figure S2A**). Subsequently, we counted ChIP-STARR-seq RNA and plasmid reads in each of these scaffold regions. The ratio of normalized RNA-seq and DNA-seq (plasmid) reads per scaffold region was then used as a measure of NCRE activity of that region. This allowed ranking of NCREs based on their activity level in different categories (**Figure 1B**) (Glaser et al., 2021). Visual inspection of RNA and plasmid tracks at multiple loci confirmed the presence of NCREs, as testified by the loci surrounding the *TKT* and the *CHD8* genes, that showed NCREs that ranked amongst the top-

10% activity category (**Figure 1C, S2B**). Independent testing of 15 NCREs confirmed their NCRE activity (**Figure S2C**). To assess whether NCRE activity correlates with the expression of putative target genes in NSCs, we first linked NCREs to their presumed target genes by combining the closest gene method and available Hi-C data obtained from the germinal zone of human foetal brain at gestation weeks 17–18 (Won et al., 2016), representing the anatomical location of NSCs in the developing human brain. We observed that genes linked to NCREs with higher activity show a significantly higher expression level compared to genes linked to regions with lower NCRE activity, reaching the highest expression level when focussing on the top-10% most active NCREs (n=14,818 NCREs, n=7,752 unique genes) (**Figure 1D**). Gene expression further significantly increased when focussing on the subset of genes involved in human disease according to the *Online Mendelian Inheritance in Man* (OMIM) catalogue (**Figure 1D**). As several NCREs with different activity levels can be linked to the same gene, we performed the same analysis using only expression data for those genes uniquely linked to each NCRE activity group and observed similar findings (**Figure S2D**).

Loss of function (LoF) of human genes has been studied as a driver of human disease. We investigated the probability of LoF intolerance (pLI) of the genes linked to the different NCRE activity categories and observed that genes regulated by highly active NCREs, besides being expressed at a higher level, are also more intolerant to LoF, suggesting they might play important biological roles (**Figure 1E**). Gene ontology analysis using Enrichr (Chen et al., 2013; Kuleshov et al., 2016) of the protein-coding genes linked to the various NCRE activity groups revealed that genes linked to the most active NCREs are mainly involved in the regulation of transcription and neurodevelopment, whereas genes linked to NCREs with lower activity were enriched in processes associated with later developmental stages such as the generation of neurons and axons, cell migration and axon guidance (**Figure 1F**).

Next, to investigate whether the presence of a certain histone modification (H3K27ac or H3K4me1) or the binding of one of the tested TFs (YY1, SOX2) would largely predict NCRE activity, we analysed the ChIP-seq signal enrichment across the different NCRE activity categories. We observed that increased NCRE activity was associated with an increased H3K27ac, YY1 and SOX2 ChIP-seq signal, while the H3K4me1 signal was rather constant across all the activity categories (**Figure 1G**). Looking at the ratio between the overlap of each NCRE category with each ChIP-seq and the total overlap of the 148,198 scaffold regions with each ChIP-seq, both normalized for the group size, we observed the same trend (**Figure S2E**).

We conclude that by employing NSC ChIP-STARR-seq, we can rank NCREs according to their activity, and show that NCRE activity correlates with expression of presumed NCRE target genes of cell-type relevance.

Active NCREs in neural stem cells show distinguishing sequence characteristics

To further characterize NCREs active in NSCs, we focused on their sequence characteristics and investigated a variety of *in silico* metrics, including GC content. The non-coding essential regulation (ncER) score measures the essentiality of each nucleotide within a given genome sequence, ranging from 0 (non-essential) to 1 (putative essential) (Wells et al., 2019); the phastcons score indicates the probability of each nucleotide of a given sequence to be conserved among multiple species (Siepel et al., 2005); the Orion score is based on the depletion of variation in the human population (Gussow et al., 2017); and finally the CADD score takes into account the likelihood of a given nucleotide variant to be deleterious (Kircher et al., 2014). We observed that NCRE activity as measured by NSC ChIP-STARR-seq is positively associated with an increase in all these scores except for GC content (**Figure 1H**), the latter in line with previous findings showing that cell-type specific NCREs have lower GC content compared to more broadly active NCREs (Colbran et al., 2017; Lecellier et al., 2018). When focussing on the top-10% most active NCREs linked to genes associated with known OMIM phenotypes, we noticed a further increase in these scores (**Figure 1H**). We previously performed a large-scale integrative computational analysis of epigenome datasets during foetal brain development (Yousefi et al., 2021) and observed that genes linked to differentially active enhancers (as measured by variability of epigenome marks over time) are less tolerant to LoF, whereas non-differentially active regions are more tolerant to LoF. We therefore wondered whether NCRE activity measured by NSC ChIP-STARR-seq would also correlate with intolerance to LoF. To investigate this, we used previously published LoF tolerance scores determined from structural variants of whole genome sequences (Xu et al., 2020). We observed that the investigated scaffold regions largely show intolerance to LoF, without a clear difference emerging between the different NCRE activity groups (**Figure S2F**). This possibly reflects the fact that the investigated scaffold regions were selected for being endogenously marked by H3K27ac, H3K4me1, or bound by YY1, or SOX2 and thus might already be enriched for regions with intolerance to LoF, independent of their NCRE activity in NSCs. Next, we investigated the localization of NCREs relative to the transcription start site (TSS) of genes and found that NCREs belonging to all the categories of activity show a similar distance distribution (**Figure 1I**) and location as annotated by Homer (**Figure S2G**). It is important to note that the design of the STARR-seq reporter construct, with the NCRE being positioned downstream of the reporter open reading frame, does not allow to measure promoter activity. Thus, even when an active region is located in proximity to a TSS, the observed activity reflects NCRE activity (i.e., enhancer activity) and not promoter activity.

To understand whether NCREs of the various activity categories are differentially bound by TFs that might determine their activity, we assessed the top-10 TF binding motifs in each

NCRE activity category. Lower activity categories were mainly enriched for motifs of various SOX family TFs important for neural development, and motifs of the RFX family, involved in the regulation of differentiation of various cell types (Sugiaman-Trapman et al., 2018), while the top-10% most active NCRE category was characterized by p53 family motifs (p53 and p73), that play a crucial role in regulating the development of the central nervous system (Marin Navarro et al., 2020; Talos et al., 2010) and the YY1 motif (**Figure 1J**). Also enriched in the top-10% most active NCREs, was the Fli1 motif, belonging to the ETS family that regulates a wide variety of processes (Seth and Watson, 2005).

Finally, we assessed the presence of transposable elements (TEs) at the identified NCREs. TEs are present throughout the human genome and are enriched in TF binding sites (Chuong et al., 2017). Moreover, in our previous work, we showed that some classes of TEs, including satellite repeats and some LTR family elements, are enriched in active NCREs in ESCs (Kunarso et al., 2010). The most enriched TEs in the top-10% most active NCREs in NSCs belong to the primate-specific subfamilies of MER61 and LTR10, both members of the ERV1 type of long terminal repeats (LTRs) (**Figure S2H**). These TEs have previously been associated with p53 binding sites (Kaller et al., 2017; Wang et al., 2007), in line with our observation that p53 binding sites are enriched in highly active NCREs (**Figure 1J**). On the other hand, TEs enriched in the classes with low NCRE activity are mostly simple repeats (**Figure 1J**). Importantly, we did not observe any overlap between the 10 most enriched TEs in NSCs and the 10 most enriched TEs we identified previously in ESC ChIP-STARR-seq (Barakat et al., 2018), suggesting that the NSC NCRE landscape is shaped mainly by a different set of TEs. However, our top hits in NSCs, MER61C and MER61E, were included in the top-20 most enriched TEs in active NCREs in ESCs.

Together this indicates that NCREs from different NCRE activity categories, despite being located at similar genomic locations relative to TSSs, show differences in sequence characteristics including sequence constraint, and enrichment for TF motifs and TEs.

Comparative ChIP-STARR-seq using NSC derived plasmid libraries transfected in ESCs identifies NCREs which are primed in ESCs for future NCRE activity

Reasoning that NCREs with a role during neural development might show higher NCRE activity specifically in NSCs whereas NCREs regulating genes with broader biological roles might show activity in multiple cell types, we next investigated cell-type specificity of NCRE activity. Therefore, we transfected ESCs with the same NSC-derived ChIP-STARR-seq plasmid libraries (**Figure 2A, Figure S1I, J**), which we refer to as comparative ChIP-STARR-seq. In NSCs, these sequences are marked by H3K27ac, H3K4me1, YY1 or SOX2 resulting in their capture into the NSC plasmid library. Testing the activity of these NSC plasmids during comparative ChIP-STARR-seq in an earlier developmental cell type such as ESCs, where these sequences are not necessarily endogenously enriched for active chromatin marks, might additionally allow direct assessment of the role of the nucleotide composition itself driving NCRE activity. As done for NSCs, we ranked the 148,198 NSC-derived scaffold regions based on the NCRE activity in ESCs (**Figure 2B**). As expected we found active NCREs, given that some NCREs show activity in multiple cell types, and because 56.35% of the NSC-derived scaffold overlapped with the ESC-derived scaffold and thus genomic regions tested in ESC ChIP-STARR-seq (Barakat et al., 2018), which focussed on chromatin marked by OCT4, NANOG, H3K27ac and H3K4me1 in ESCs (**Figure S3A**). First, comparing for each of the 148,198 scaffold regions the changes in NCRE activity groups between NSC ChIP-STARR-seq and comparative ChIP-STARR-seq, we noticed that the majority of regions remained in the same or in an adjacent higher or lower activity category, and only the minority showed extreme changes (e.g., going from category 5 to category 1 and vice versa) (**Figure 2C**). To further interpret comparative ChIP-STARR-seq findings, we linked NCREs to presumed target genes using proximity and available HiC data in H9 ESCs (Freire-Pritchett et al., 2017). Contrary to NSCs, we observed that increased NCRE activity of the NSC-derived genomic regions transfected in ESCs is not accompanied by a major increase in gene expression of their presumed target genes in ESCs (**Figure 2D**)**.** Like in NSCs, an increased intolerance to LoF of the presumed ESC target genes was observed (**Figure 2E**). Focussing on sequence characteristics, we did not notice a consistent stepwise increase of the various scores following increased comparative ChIP-STARR-seq NCRE activity as found in NSCs, with the exception of the phastcons and CADD score. Rather, we observed a decrease in essentiality (ncER), in Orion and in GC score, the latter mirroring our findings in NSCs (**Figure 2F**). Next, we investigated the localization of NCREs relative to the TSS of genes and noticed that NCREs with increased activity in comparative ChIP-STARR-seq in ESCs tended to be more distally located (**Figure S3B**). When performing TF binding motif analysis, whereas the top-10% most active NCRE category in NSCs was dominated by motifs for p73, p53 and YY1

(**Figure 1J**), these motifs were no longer found amongst the top-10 most enriched motifs in any of the NCRE activity categories from comparative ChIP-STARR-seq in ESCs (**Figure S3C**). Several other TF motifs, including SOX2, SOX3, SOX10, RFX and RFX2 were enriched in the top-10% category in ESCs, like in NSCs, whereas ERG and ETV2, other members of the above mentioned ETS family, were only found enriched in the top-10% most active NCRE category in ESCs but not in NSCs (**Figure S3C**). Interestingly, with the exception of the OCT4-SOX2-TCF-NANOG motif only found in NCRE category 5 of comparative ChIP-STARR-seq in ESCs, we did not observe major differences in TF motif enrichment between NCREs ranked in the 5 activity categories in ESCs, as they all showed SOX and RFX family TFs (**Figure S3C**). When investigating the 10 most enriched TEs in the different categories of NCRE activity in ESCs, we observed similar results as for the NSCs (**Figure S3D**), with a large overlap of TEs identified in highly active NCREs in both cell types.

Gene ontology analysis of the protein-coding genes linked to the various classes of ESC NCRE activity showed that genes involved in the regulation of transcription are enriched amongst all ESC NCRE activity categories (**Figure 2G**). Categories with higher NCRE activity in comparative ChIP-STARR-seq were also enriched for processes involved in nervous system development (**Figure 2G**). This was surprising given that these experiments were performed in ESCs where we would have expected to find a more ESC related gene ontology for genes linked to the most active NCREs. Together with the lack of a major increase in target gene expression in ESCs upon increased NCRE activity from comparative ChIP-STARR-seq in ESCs (**Figure 2D**) and the absence of typical ESC TF and TE enrichments that we previously observed in ESC ChIP-STARR-seq with ESC-derived libraries (Barakat et al., 2018), this might indicate that the NCREs showing episomal activity in comparative ChIP-STARR-seq in ESCs might not all be active at the endogenous loci in ESCs. This could possibly reflect the NSC origin of the plasmid libraries and the fact that these sequences in the ESC experiments are uncoupled from their cell type specific environment and chromatin landscape. Thus, they might not be active at the endogenous loci in ESCs, in contrast to NSCs where these NCREs are endogenously marked by H3K27ac, H3K4me1, YY1 or SOX2. Indeed, when determining the ChIP-seq signal enrichment obtained from ESC data for the NCREs from the various activity categories from comparative ChIP-STARR-seq in ESCs, we only noticed a limited enrichment of H3K27ac and H3K4me1 in ESCs at these sites, not correlating with their measured NCRE activity (**Figure 2H**). Likewise, enrichment of YY1, SOX2, OCT4 and NANOG at these sites in ESCs was limited. Both of these trends were slightly increased when assessing only those scaffold regions showing an overlap between the NSC-derived scaffold used herein, and the previously used ESC-derived scaffold from ESC ChIP-STARR-seq (Barakat et al., 2018) (**Figure S3E**), but still dramatically reduced

when compared to the enrichment of H3K27ac, H3K4me1, YY1 and SOX2 in NSCs over these regions (**Figure 2H, S3E**). Surprisingly, despite the lack of H3K27ac and H3K4me1 in ESCs, these regions where still located in accessible chromatin in ESCs as determined using ATAC-seq and depleted for the repressive mark H3K27me3 in ESCs (**Figure 2H**). Additionally, these same regions are enriched for H3K4me2 and H3K4me3, with the latter being more enriched at the lower active NCRE categories from comparative ChIP-STARR-seq in ESCs (**Figure 2H**). Previously, so called primed enhancers have been shown to lack H3K27ac and are characterized by H3K4 methylation (Heinz et al., 2015). Together, the epigenome profile indicates that some of the NCREs identified by comparative ChIP-STARR-seq in ESCs using NSC-derived plasmid libraries belong to such a class of primed NCREs in ESCs, which are primed for activity at later developmental stages. An extensive differential NCRE activity analysis and epigenome profiling (**Supplementary Note 2, Figure S4**), comparing NCREs with high activity in the one cell type and low activity in the other, and vice versa, likewise pointed to priming of NCREs for later activity along the trajectory of neural differentiation, although future studies will be required to further resolve the underlying mechanisms.

BRAIN-MAGNET, a convolutional neural network model, predicts NCRE activity based on DNA sequence and facilitates interpretation of biological consequences of non-coding variants

The NSC ChIP-STARR-seq experiments performed herein provide a comprehensive atlas of annotated NCREs ranked by activity in a cell type relevant for early brain development. We hypothesized that such an NCRE atlas might aid the interpretation of non-coding genetic variants encountered in neurodevelopmental disorders. The diagnostic yield for most of these disorders using current genomic technologies which mainly focus on coding sequences is well under 50% (Clark *et al.*, 2018; Smith *et al.*, 2019), suggesting that some causes of missing heritability might be located in NCREs. If this would be the case, one would expect that disease relevant NCREs would be relatively depleted of rare genomic variation in the healthy population compared to other genome sequences. Indeed, using the recently released gnomAD v4 data, containing 76,215 genomes from individuals that did not present with early onset neurodevelopmental phenotypes (Chen *et al.*, 2024), we found that NCREs with higher activity in NSCs harboured less rare variants (minor allele frequency (MAF) <0.1%) compared to NCREs with lower activity (**Figure S5A**). An even stronger depletion of rare variants was observed upon increasing NCRE activity from comparative ChIP-STARR-seq in ESC (**Figure S5A**). As expected, the coding sequences (CDS) of protein coding genes showed the strongest depletion of rare variants. Nevertheless the relative depletion of rare variants in the

highly active NCREs in both NSCs and ESCs was in a similar range as for other sets of NCREs, including VISTA enhancers (Visel et al., 2007), candidate *cis* regulatory elements from ENCODE (Consortium et al., 2020), differentially active enhancers (DAEs) and non-differentially active enhancers (nDAEs) from an epigenome study of human foetal brain (Yousefi *et al.*, 2021), and for the 5' and 3' UTRs of genes (**Figure S5A**). Although this relative depletion of rare variants is in line with the hypothesis that the identified NCREs from our atlas could be disease relevant, the 148,198 NCREs assessed, encompassing a total size of 112,314,966 nucleotides, still harboured 28,486,102 rare variants (MAF <0.1%) in healthy individuals. Most of these variants are likely not disease relevant as they will not interfere with NCRE function, but it is currently difficult to assess their potential pathogenicity, especially given the lack of a genetic code for NCREs contrary to that of the protein coding exome. If a rare variant would affect a nucleotide directly relevant for the NCRE function and consequently downstream gene expression, it would be more likely to be disease relevant. This underscores the necessity to develop tools that can predict the function of individual nucleotides within NCREs, enabling prioritizing variants that are more likely to be pathogenic.

To be able to interpret the potential effects of variants in NCREs, we set out to build an artificial intelligence based prediction model. As identifying non-coding variants that affect NCRE function is reminiscent of searching for needles in a haystack, and needles are more likely to be identified using a magnet, we refer to the model as BRAIN-MAGNET (*BRain-focussed Artificial INtelligence Method to Analyse Genomes for Non-coding regulatory Element mutation Targets*). BRAIN-MAGNET is a sequence-based single-task convolutional neural network (CNN) model (LeCun et al., 2015), that allows predicting NCRE activity for NSCs and ESCs individually (**Figure 3A, Figure S5B,C, Methods**). This model uses the genomic sequences assessed by ChIP-STARR-seq as input and links to their measured activity as output, determining what is causing a given sequence to have a given activity. BRAIN-MAGNET trained on both NSC ChIP-STARR-seq and comparative ChIP-STARR-seq in ESCs could precisely predict NCRE activity (**Figure 3B, Figure S5B,C**), with the ESC trained model slightly outperforming the NSC trained model (**Figure S5B,C**). Possibly this reflects the fact that some NCREs identified by comparative ChIP-STARR-seq in ESCs might be primed, and these NCREs might be more depending on their sequence composition for episomal activity which might be more amenable for training by the sequence-based model.

To further interpret the predictions of BRAIN-MAGNET, we next utilized the explainable AI DeepExplainer framework (Lundberg and Lee, 2017; Shrikumar et al., 2018) to calculate BRAIN-MAGNET contribution (cb) scores. These cb scores indicate for each of the 112,314,966 nucleotides in our NCRE atlas its relative contribution to NCRE activity. Next, we identified functionally important motifs in sequences with high cb scores using TF-MoDISco-

lite (see **Methods**) (**Figure 3C**). For example in the NSC NCREs we found prominent enrichment for the TP53 and TP73 motifs when using BRAIN-MAGNET trained on NSC ChIP-STARR-seq (**Figure 3C**). Another motif frequently identified was ZFP42 (REX1). REX1 originated from a duplication event of YY1 in the eutherian lineage (Kim et al., 2007), thus it utilizes the same TF binding motif. Since REX1 is not expressed in NSCs, the observed enrichment for the ZFP42 motif likely represents YY1 bindings sites. Indeed, the YY1 motif is the third redundant motifs called by the algorithm at the ZFP42 motif sites. Hence, the results from BRAIN-MAGNET are in line with the previous motif discovery analysis (**Figure 1J**) that showed enrichment of the same TFs at active NCREs, but additionally indicate that the binding of these TFs might also be functionally important for NCRE activity given the high cb scores of these motifs. To functionally validate BRAIN-MAGNET findings and to confirm that the prioritized motifs were indeed required for NCRE activity, we selected 15 NCREs highly active in NSCs that displayed a significant TF motif with high cb score (**Figures 3D, E, F, S6-S11**). We cloned these NCREs in the STARR-seq reporter plasmid and generated mutations lacking 30 bp centred on the TF motif. Upon transfection in NSCs, wild type NCREs displayed the expected NCRE activity, while the deletion mutants significantly impacted on the NCRE activity of 14 out of 15 tested NCREs (**Figure 3D, S6-11**). Likewise, generating point mutations disrupting the prioritized motifs of 4 NCREs similarly affected NCRE activity (**Figure 3E, S6, S7**). For example, introducing 4 nucleotide changes in the prioritized TP53 motif of an NCRE upstream of *ACTB* (**Figure 3F**), a gene of which mutations cause a syndromic neurodevelopmental disorder known as Baraitser-Winter syndrome (OMIM #243310), significantly reduced NCRE activity (**Figure 3E, F**). The same holds true when introducing point mutations in the TP53 motif of an NCRE interacting with *OAT*, a gene of which mutations are linked to gyrate atrophy of the choroid and retina (OMIM #258870), and in NCREs of *PAFAH1B1* and *ASH1L* (**Figure 3E, S6-7**), genes playing a role in lissencephaly (OMIM #607432) and intellectual disability (OMIM #617796), respectively.

Together this indicates that BRAIN-MAGNET can predict NCRE activity based on DNA sequence composition and can be exploited to identify which nucleotides and motifs within NCREs can impact on NCRE activity when mutated.

Application of the NCRE atlas and BRAIN-MAGNET to prioritize genomic variants found in common disease

To test the utility of our NCRE atlas and BRAIN-MAGNET to prioritize genomic variants, we first applied them to genome-wide association studies (GWAS). Numerous GWAS have identified risk loci for neuropsychiatric disorders, of which the far majority is located in the non-coding genome (Maurano *et al.*, 2012). It has been tremendously challenging to decipher the mechanisms through which SNPs identified in GWAS might confer

these disease risks, as the majority of SNPs is in linkage disequilibrium (LD) with other adjacent SNPs. Thus, it remains unclear which of these genomic variants underscores the allele-specific biological effects causing the increased disease risks. Recently, MPRAs have been employed to fine-map GWAS loci and assess how the lead SNPs and the adjacent SNPs in LD would affect function of NCREs at single nucleotide resolution. In one such application, Guo et al. assessed 2,221 non-coding variants associated with 10 neuropsychiatric disorders, with an average of 5 SNPs per disease risk loci being tested (Guo et al., 2023). Comparing the activity of matched risk versus reference allele pairs, they identified 892 SNPs that showed differential activity (daSNVs) in their MPRA depending on the genotype of the given SNP. As the BRAIN-MAGNET cb scores enable to predict which nucleotides within a given NCRE have the highest likelihood of impacting on the NCRE activity and daSNVs were already shown to affect NCRE activity in an allele-specific manner (Guo *et al.*, 2023), we hypothesised that daSNVs that overlapped with the NCREs measured in our study should have on average a higher BRAIN-MAGNET cb score than non-daSNVs. In agreement with this hypothesis, we observed a significant difference in normalized cb score percentile between daSNVs and non-daSNVs (p=0.0374, t-test) (**Figure S12A**). Since GWAS typically nominates a lead SNP, with other SNPs in LD, we next asked whether BRAIN-MAGNET cb scores could prioritize which SNP in a given GWAS locus would be more likely to be a daSNV. We therefore investigated GWAS associated regions where multiple SNPs were previously functionally tested (Guo *et al.*, 2023) and for which multiple NCREs were present in our data set. One such locus on chromosome 6, associated with the rs401754 SNP that was previously shown to be associated with an increased risk for schizophrenia, contains in total 7 SNPs in LD which are located on 4 NCREs from our atlas (**Figure 4A**). Amongst these 7 SNPs, rs200483 has relatively the highest normalized cb score compared to the other 6 SNPs (**Figure 4B, S12B**), and indeed rs200483 was previously identified as daSNVs (Guo *et al.*, 2023). Similarly, at a locus on chromosome 1 associated with major depression, comparing the relative normalized cb score amongst the 4 candidate SNPs in LD rightly pointed out to rs301806 as the daSNV (**Figure 4C, D, S12C**). Of note is that other nucleotides within these NCREs, not overlapping with common SNPs, have higher cb scores, and alterations of those nucleotides might thus

impact on the activity of these sequences more severely than the daSNVs. The likely dramatic effect of such genomic variants might not be compatible with the expected moderate risk effects of GWAS loci. Therefore, those nucleotides might not harbor common SNPs associated with common disease, but perhaps might harbor rare variants causing other related monogenic disorders. Together this indicates that comparing cb scores of candidate SNPs at GWAS loci can help to prioritize those SNPs that have the relatively highest likelihood to affect NCRE activity compared to the reference allele.

Application of the NCRE atlas and BRAIN-MAGNET to prioritize genomic variants found in rare disease

Finally, we investigated the utility of the NCRE atlas to investigate non-coding variants in WGS data from individuals with rare disease. We first applied the NCRE atlas to a patient from our clinic, which was clinically suspected of Mowat-Wilson syndrome, a severe neurodevelopmental disorder linked to dysfunction of the *ZEB2* gene. Previous extensive genetic investigations, including trio whole exome and genome sequencing did not reveal a protein coding mutation in *ZEB2*, or any other likely disease-causing variant, but identified a chr2:144469837C>T variant in intron 2 of *ZEB2*. This variant overlapped with a highly active NCRE from NSCs (category 5), was absent in gnomAD v4, had a CADD score of 20.7, an ncER score of 99.6 and was predicted by the RegulomeDB v.2 (Dong et al., 2023) to contain a TF binding motif (**Figure S13A**). According to LINSIGHT (Huang et al., 2017), an algorithm that aims to predict which non-coding variants are likely to have deleterious fitness effects, the chr2:144469837C>T variant was predicted to be likely pathogenic (score 0.98). In contrast, according to BRAIN-MAGNET, this nucleotide had a low cb score (65.7th percentile within the NCRE (cb_each); 52.4th percentile within all NCREs (cb_all)), while a different ONECUT1 TF motif located 36 nucleotides upstream is predicted to mostly impact on activity of this NCRE. To investigate this functionally, we cloned the NCRE in the STARR-seq plasmid and introduced by site directed mutagenesis either the variant identified in the patient or a 27 bp deletion centred on the ONECUT1 motif. Upon transfection in NSCs, the ONECUT1 deletion completely abolished the NCRE activity, while the patient variant did not majorly affect NCRE activity, in agreement with the BRAIN-MAGNET predictions (**Figure S13B**). In addition, clinical RNA-seq of patient derived fibroblasts did not reveal any *ZEB2* expression changes and a diagnostic Episignature did not reveal evidence of a Mowat-Wilson syndrome specific DNA-methylation pattern (data not shown). Even though we cannot completely exclude that the intronic *ZEB2* variant could have an effect in other cell types, based on the current functional testing the variant was clinically classified as a variant of unknown significance which was considered unlikely to contribute to the phenotype which currently remains unexplained.

We next applied BRAIN-MAGNET to the data from the Genomics England 100,000 Genomes Project (GEL) (Caulfield, 2017). If genetic variants at nucleotides with high cb scores would contribute to unexplained neurodevelopmental disorder phenotypes, we should observe an increased burden of *de novo* variants overlapping such nucleotides in genetically unexplained patients with neurodevelopmental disorders. To investigate this, we assessed 4,415 individuals from the unexplained neurology and neurodevelopmental cohort (NDD) of the v18 GEL data release for which 13,276 *de novo* variants overlapped with our NSC NCREs Indeed, at population level, we observe a tendency towards an increased number of *de novo* variants at nucleotides with higher cb scores in NCREs from higher NSC ChIP-STARR-seq activity groups (**Figure 5A**), with similar findings for the NCREs identified by comparative ChIP-STARR-seq in ESCs (**Figure 5B**). Likewise, we found a similar trend when investigating 4,558 genomes from individuals with a variety of other rare disease categories (**Figure S13C**).

We next screened the GEL data, searching for variants that were overlapping with 20 bp fragments centred on the high confidence motifs found in the most active (category 4 and 5) NSC NCREs for which BRAIN-MAGNET predicts a major impact when disturbed. We found 705 rare variants (<10 times present in GEL) in 824 individuals with unexplained NDDs that were located within 20 bp centred on 440 motifs in NCREs that are linked to known OMIM genes. We randomly chose 4 variants for functional validation which were absent in gnomAD v4. Three out of four variants significantly affected NCRE activity, with the fourth variant in an NCRE linked to *GRIA4* (gene linked to autosomal dominant neurodevelopmental disorder with or without seizures and gait abnormalities, OMIM #617864), found in an individual with a complex neurological phenotype including ataxia and spasticity, not showing a significant difference (**Figure 5C, D**). Surprisingly, the activity of an NCRE linked to *KPTN* (gene linked to autosomal recessive intellectual developmental disorder 41, OMIM #615637) increased upon introduction of the rare variant that was identified in two unrelated individuals affected by epilepsy and ataxia. This indicates that genomic variants overlapping nucleotides with high cb scores do not necessarily only diminish NCRE activity but could also cause gain-of-function. The heterozygous variant in an NCRE linked to *MN1* (gene linked to autosomal dominant CEBALID syndrome, OMIM #618774, that presents with dysmorphic features, intellectual disability, and brain malformations), significantly reduced the NCRE activity. This variant was found in an unexplained individual reported to have a complicated epilepsy phenotype with amongst other features abnormalities of the cerebral cortex possibly presenting a (partial) phenotypic match. Finally, the variant which disrupts a ZFP42/YY1 binding motif in an NCRE linked to *RAB7A* significantly reduced NCRE activity (**Figure 5C,D**). As *RAB7A* is linked to an autosomal dominant type of Charcot-Marie-Tooth disease (type 2B, OMIM #600882, (Saveri et al., 2020), and the currently genetically unexplained individual

harbouring this variant was clinically diagnosed with Charcot-Marie-Tooth disease, it is currently considered that this non-coding variant is causative for the disease phenotype and clinical follow-up is currently initiated.

Finally, we screened through an independent set of 249 trio WGS of unsolved rare disorder patients. Focussing solely on *de novo* variants overlapping NCREs with a cb score above the 95th percentile, we identified 12 variants in NCREs which target genes could potentially provide a clinical match.

Together, this indicates that the NCRE atlas presented herein coupled with BRAIN-MAGNET can facilitate the search for possibly disease implicated rare variants identified by WGS in rare disease cohorts. Furthermore, BRAIN-MAGNET can predict the impact of such variants on functional NCRE activity (**Figure 6**).

Discussion

Here we used ChIP-STARR-seq to build an extensive atlas of NCREs functionally active in NSCs. Besides informing on the biological mechanisms regulating gene expression in NSCs and on NCREs distinguishing sequence characteristics, we provide evidence that a CNN model trained on the experimental data allows to predict NCRE activity solely based on NCRE nucleotide composition, similar to what other recent MPRA studies have found (Bravo Gonzalez-Blas et al., 2024; de Almeida et al., 2022; Smith et al., 2023). BRAIN-MAGNET can be used to predict the functional effects of genomic variants overlapping with NCREs, and such predictions occur with high confidence as testified by our functional validation. We show how BRAIN-MAGNET can be used to interpret effects of common SNPs identified in GWAS, as well as rare variants identified by WGS in affected individuals with currently unexplained neurological and neurodevelopmental phenotypes. We thus anticipate that this resource might be useful to diminish so-called missing heritability of non-coding variants in the field of human genetics.

Our study inherently has several limitations. First, MPRAs measure NCRE activity in an episomal context, outside of the natural chromatin environment, hence, results might not always reflect the endogenous NCRE activity. Notwithstanding this potential limitation, previous work has already extensively shown that multiple MPRA findings can be reproduced when altering NCREs at the endogenous locus (Barakat *et al.*, 2018; Lim et al., 2024; Smith *et al.*, 2023). To minimize potentially confounding effects, we generated ChIP-STARR-seq plasmid libraries from chromatin marked in NSCs with histone modifications associated with active NCREs. In addition, the episomal approach might also reveal insights in mechanisms that are alternatively difficult to study at the endogenous locus, such as the potential priming of NCREs in ESCs for later activation that we observe here when using comparative ChIP-STARR-seq. Second, results of MPRAs are obtained in a given cell type, and it remains to be investigated how cell type specific findings and rules identified by CNN models can be extrapolated to other cell types and tissues. Third, although we show the utility of this resource to prioritize non-coding genomic variants for functional effects on NCRE activity, it remains challenging to draw final conclusions regarding pathogenicity of such variants and causality in causing rare disease. One of the largest conceptual challenges in this regard is the fact that the clinical phenotype of an NCRE disrupting variant does not necessarily has to equal the phenotype of a protein coding pathogenic variant affecting the gene that the NCRE is regulating. This is illustrated by tissue-specific effects of mutations in the *PTF1A* enhancer, where such NCRE mutation only causes part of the phenotype compared to that of the *PTF1A* disease gene (Weedon et al., 2014). Hence, when finding a variant that strongly impacts on NCRE function, its associated phenotype might be similar to the known phenotype of the target

gene, such as is likely the case of the *RAB7A* example that we present, but might also cause phenotypes that represent only “partial known syndromes”, or novel phenotypes which might complicate the clinical interpretation of such variants. Future studies will be required to more clearly dissect correlations between cell type specific effects of NCREs and phenotypes that NCRE dysfunction might cause. The NCRE atlas presented herein and BRAIN-MAGNET will form a solid starting point for such studies.

Acknowledgements

We thank all members of the Barakat lab, the Clinical Genetics Discovery Unit, Innovation Team and Whole Genome Sequencing Implementation and Research Task Force at Erasmus MC for helpful discussions. Shimriet Zeidler (Erasmus MC) is acknowledged for sharing clinical information on the *ZEB2*. Some of the authors are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA. RD was supported by a China Scholarship Council (CSC) PhD Fellowship (201906300026 to RD) for her PhD studies at the Erasmus Medical Center, Rotterdam, The Netherlands. KL was supported by a ZonMw PSIDER Doorbraken grant (grant 10250042110005), a Brain and Behavior Research Foundation Young Investigator award (grant 30787) and a NWO Veni grant (grant 501100003246). GR was supported by the ZonMw Veni grant 1936320. Part of this research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK, and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. Some of the analysis involved external data generated by the ENCODE and Roadmap projects, that received funding from the National Institutes of Health (NIH) (grants U01ES017166, U54HG004570, U41HG006992 and U01ES017155). The Barakat lab was supported by the Netherlands Organisation for Scientific Research (ZonMw Veni, grant 91617021; ZonMw Vidi, grant 09150172110002), a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation, an Erasmus MC Fellowship 2017, and Erasmus MC Human Disease Model Award 2018, and acknowledges other ongoing support for rare disease research from Stichting 12q, EpilepsieNL, CURE Epilepsy and the Spastic Paraplegia Foundation, Inc. Funding bodies did not have any influence on study design, results, and data interpretation or final manuscript.

Data availability

Genome sequencing data from the 100,000 Genomes Project and limited phenotype information are available in the National Genomic Research Library from the 100,000 Genomes Project for which researchers can apply for access at Genomics England. Due to privacy regulations and given consent under which the in house patient was recruited, raw patient genomic sequencing data cannot be made available. All other raw data are prior to the peer reviewed publication available upon reasonable request from the corresponding author.

Declarations:

All authors declare no conflict of interest

Author contribution

RD performed all computational and genomic analysis, with help of SY, RS, MM, JP, MS, TH, GVR and EM. EP performed all multi-omics experiments. AN and KL performed validation experiments. MJP provided clinical data. WvIJ and Genomics England Research Consortium performed next generation sequencing and genome sequencing. TSB conceived and supervised the work. RD, EP and TSB wrote the manuscript with input from all authors. All authors approved the final version of the manuscript.

Supplementary Note 1

Adapted ChIP-STARR-seq approach

Compared to our previous application of ChIP-STARR-seq in human embryonic stem cells (ESCs) (Barakat *et al.*, 2018) that focussed on NCREs enriched for the TFs NANOG, OCT4, and the histone modifications H3K4me1 and H3K27ac (herein referred to as ESC ChIP-STARR-seq), we performed a number of updates on the experimental procedures, aiming to reduce experimental noise, increasing reproducibility and reducing experimental complexity and costs (**Figure S1A**).

First, to enrich for genomic regions with putative NCRE function in NSCs, we performed ChIP in duplicates for the TFs SOX2 and YY1, that are important in NSCs (Miyagi *et al.*, 2009; Zurkirchen *et al.*, 2019), and the histone modifications H3K27ac and H3K4me1, which are associated with putative NCREs (Rada-Iglesias *et al.*, 2011). ChIP-qPCR showed an enrichment at known NCREs surrounding the *FGFR1*, *PAX6*, *SOX2* and *NES* loci (**Figure S1B**). Sequencing of the ChIP material (ChIP-seq) showed a good correlation between replicates (Pearson correlation coefficient > 0.88) (**Figure S1C,D**). Next, using the same ChIP-DNA, ChIP-STARR-seq plasmid libraries were generated from both ChIP replicates and pooled. To further decrease the number of transfections and simultaneously increase the number of individual plasmids covering each putative NCRE, theoretically reducing experimental noise, we pooled together the plasmid libraries derived from ChIP for the histone modifications H3K27ac and H3K4me1, and for the TFs SOX2 and YY1. Hereafter, we refer to these combined plasmid libraries as HIST and TF libraries, respectively, and the application of these libraries for ChIP-STARR-seq in NSCs or ESCs as NSC ChIP-STARR-seq or comparative ChIP-STARR-seq, respectively. Visual inspection of sequencing results to assess whether the pooled plasmid libraries are a good representation of the initial ChIP material, showed that peaks from ChIP-seq experiments were still recognisable in the sequencing tracks of the pooled plasmids (**Figure S1E**), and analysis of normalized sequencing reads confirmed a high correlation between initial ChIP-seq peaks and corresponding pooled plasmid libraries (**Figure S1C,D**). On average, 93.04% of HIST ChIP-seq peaks and 96.86% of TF ChIP-seq peaks were covered by more than 5 different plasmids and 46.91% of HIST peaks and 67.54% of TF peaks by more than 20 plasmids, with a mean of 27 and 43 distinct plasmids per HIST and TF ChIP-seq peak, respectively. Together this indicates an increased complexity of the generated plasmid libraries compared to our earlier study using ChIP-STARR-seq in ESCs (Barakat et al., 2018).

Second, to overcome the variability between samples that we previously observed in ESC ChIP-STARR-seq (Barakat et al., 2018), we implemented the following changes in the

experimental strategy. We performed 5 independent transfections per plasmid pool, each starting with ~3 million NSCs seeded 48 hours before the transfection (or ~6 million ESCs transfected in suspension, see **Methods**). 24 hours post-transfection, we collected and pooled together approximately 30 million cells from these 5 independent transfections (**Figure S1A**). FACS analysis showed that 26.5% and 30% of NSCs transfected with the HIST or TF libraries, respectively, were GFP positive (**Figure S1F**).

Third, RNA was isolated from these pooled, unsorted cells and subsequently split and treated as two replicates, which were subjected to DNase treatment and cDNA synthesis. To limit PCR artefacts, the first PCR amplification step which suppresses residual plasmid contamination by the use of primers spanning a synthetic intron in the STARR-seq plasmid, was performed in 10 parallel reactions for each of the two replicates. These 10 PCRs per replicate were then pooled, purified and further amplified with the index primers required for sequencing, using 5 independent PCR reactions and two independent index primers for each replicate. Thus, we sequenced a total of four replicates of STARR-seq RNA for each pooled HIST or TF library, with an average of 20 million reads per sample (**Figure S1A**). As the two technical replicates amplified with independent index primers showed a high correlation (Pearson correlation coefficient >0.96) (**Figure S1G,H,I,J**), they were merged for further downstream data analysis.

Together this indicates that our revised experimental strategy generated highly complex plasmid libraries capturing genomic regions marked by H3K27ac, H3K4me1 or bound by SOX2 and YY1 in NSCs, which upon transfection and RNA sequencing results in highly correlated replicate data.

Supplementary Note 2

Differential NCRE activity analysis in NSCs and ESCs

To identify differences in NCRE activity between the two cell types, we next performed a differential NCRE activity analysis (Glaser *et al.*, 2021), comparing NCREs captured in the same NSC-derived plasmid library with high activity in the one cell type and low activity in the other, and vice versa (**Figure S4A**). To this end, we first selected those NCREs that ranked in the top-10% of NSC ChIP-STARR-seq activity in NSCs and in the bottom-90% of comparative ChIP-STARR-seq in ESCs, and those NCREs that ranked in the top-10% of comparative ChIP-STARR-seq activity in ESCs and the bottom-90% in NSC ChIP-STARR-seq. With these settings, we identified 6,850 out of 14,818 NCREs (46.2%) overlapping between the top-10% of NSCs and ESCs (referred to as common-high) (**Figure S4B)**, possibly explaining part of the overlap between TF motifs and TEs in NSC and ESC active NCREs (**Figure 1J, S3C**), and 7,968 out of 14,818 NCREs (53.8%) per cell type that were either NSC- or ESC-high. Visual inspection of the tracks surrounding the *CHD2* locus shows both common-high NCREs and differentially active NCREs (**Figure S4C**). Investigating the distance distribution of these regions from the TSS, we observed that part of the NSC-high NCREs is located in closer proximity to TSSs, while common and ESC-high NCREs were spatially distributed in a more similar manner (**Figure S4D**). A deeper investigation revealed that all the three classes or NCREs are enriched mainly at intronic (NSC-high: 46.4%; ESC-high: 55.2%; common-high: 53.7%) and intergenic (NSC-high: 29.6%; ESC-high: 36.7%; common-high: 35.9%) regions with an increase at promoter-TSS regions for the NSC-high ones (NSC-high: 17.1%; ESC-high: 4.4%; common-high: 6.3%), in agreement with their tendency to be located in closer proximity to the TSS (**Figure S4E**). Next, we wondered whether cell-type specific highly active NCREs were also more susceptible to LoF. As we observed previously (**Figure S2F, S3B**), the majority of the investigated regions are intolerant to LoF, with NSC-high NCREs showing a larger enrichment amongst intolerant regions, possibly indicating that alterations of these regions might be negatively selected in the human population or might be implicated in disease (**Figure S4F**). Investigating the pLI of the genes linked to these NCREs groups, we did not find any significant difference (**Figure S4G**).

Next, we aimed at understanding whether these NCREs might regulate cell-type specific gene expression. We observed that NSC-high NCREs are linked to genes expressed at a significantly higher level in NSCs but also in ESCs (**Figure S4H**), whereas genes linked to ESC-high and common-high NCREs did not show major expression differences. Gene ontology analysis identified for all three categories of NSC-high, ESC-high and common-high NCREs terms related to transcriptional regulation (**Figure S4I**). Common-high NCREs were

also enriched for terms related to nervous system development and generation of neurons and axons. Like previously seen for all NCREs active upon comparative ChIP-STARR-seq in ESCs, also ESC-high NCREs were enriched for nervous system development and generation of neurons, rather than other terms that would be expected for pluripotent stem cells, despite that these genes are not highly expressed in ESCs. This observation further provides support for the hypothesis that these ESC-high NCREs, despite showing episomal activity in comparative ChIP-STARR-seq in ESCs when using plasmid libraries obtained from NSCs, might not be active at the endogenous chromatin in ESCs. Rather, these ESC-high NCREs might be primed for future activation at the endogenous locus, and this priming might not be recapitulated in an episomal plasmid-based system.

To further investigate this hypothesis, we assessed various chromatin profiling data at the common and cell-type specific highly active NCREs. As expected, since all the investigated putative NCREs are derived from H3K27ac, H3K4me1, YY1 and SOX2 ChIP in NSCs, all the common and differentially active NCREs (ESC-high, NSC-high) show central enrichment of these chromatin marks across NSC ChIP-seq data (**Figure S4J,K**). To further characterize the 7968 NSC-high NCREs at their endogenous locus, we retrieved H3K4me3 and H3K27me3 from previously published data of H9-derived NSCs (Chan et al., 2013) and performed ATAC-seq in NSCs to assess chromatin accessibility. We observed that the NSC-high NCREs are enriched for accessible chromatin, as expected given the chromatin immunoprecipitation in this cell type, and a strong enrichment of the active chromatin mark H3K4me3 (**Figure S4J,K**) overlapping the enrichment of YY1 and H3K27ac. H3K4me3, as well as YY1, is highly enriched at promoters (Santos-Rosa et al., 2002), suggesting this fraction of loci might partially include the 17% of NSC-high NCREs found in proximity to promoters (as previously discussed, the STARR-seq design does not allow to test promoter activity and thus despite being located in proximity to promoters these sequences are NCREs). Compatible with the immunoprecipitation of markers associated with NCRE activity used to identify ChIP-STARR-seq NCREs, we did not see enrichment of the inactive chromatin mark H3K27me3 at these NCREs.

To investigate the hypothesis that ESC-high NCREs are not enriched in active marks at the endogenous chromatin loci, we collected H9 ESC epigenome data from publicly available sources and from our previously generated H3K27ac, H3K4me1, OCT4 and NANOG ChIP-seq (Barakat *et al.*, 2018). Furthermore, we generated ChIP-seq data for YY1 and SOX2 in ESCs. ESC-high NCREs, even though located in accessible chromatin in ESCs, show no enrichment of the active chromatin marks H3K27ac and H3K4me1 in ESCs (**Figure S4J,K**). The majority of the ESC-high NCREs from comparative ChIP-STARR-seq in ESCs show a very mild enrichment of SOX2 and NANOG binding, but interestingly, a subset of the ESC-

high NCREs (n=1,100 out of the 7,968 ESC-high NCREs) had an accumulation of the histone marks H3K4me2 and H3K4me3, with especially the latter mark showing a bi-modal distribution with a valley over the NCRE centre. Primed NCREs are indeed characterized by methylation of H3K4 and no acetylation of H3K27 (Heinz *et al.*, 2015). Therefore, this observation is compatible with the hypothesis that these NCREs, despite being highly active in the episomal setting, are not active at the endogenous locus but primed for activation at a later stage. Further pointing to this scenario, we noticed that ESC-high NCREs primed with H3K4 methylation, gain H3K27 acetylation at the NSC stage, although to a lower level compared to NSC-high NCREs, compatible with the lower activity of these regions in NSCs (**Figure S4J,K**). Interestingly, these regions maintain H3K4me3 also at later stages of differentiation, in neurons and in astrocytes, where they also display H3K27ac. The same scenario described for ESC-high NCREs applies to common-high NCREs, for which we could observe a fraction of regions (n=1,200 out of 6,850 common-high) with H3K4me2/3 accumulation and no H3K27ac in ESCs, but with H3K27ac in NSCs. H3K4me3 regions have been previously reported to be located mainly at the TSS (Santos-Rosa *et al.*, 2002). In line with this, we observed that the majority of the ESC-high and common-high NCREs marked by H3K4me3 were located in proximity of a TSS, within 5kb down- or upstream (**Figure S4L**), but as outlined as the STARR-seq design does not allow to capture promoter activity these are thus NCREs in proximity of the TSS. Moreover, when investigating the enrichment of TF binding motifs, the TATA-box was among the most enriched motifs in the H3K4me2/3 positive subset of both the ESC-high and common-high NCREs (**Figure S4M**). Further supporting the hypothesis that this subset of regions is primed for activation at a later stage, we noticed that the majority of them (>75%) in NSCs belong to the NCRE activity categories 4 and 5 (**Figure S4N**).

Analysing further the TF motifs enriched in these regions, we noticed that the most enriched in the common-high subset positive for H3K4me2/3 in ESCs were YY1, p73 and p53, possibly explaining their activity in NSCs where these regions are marked by H3K27ac. We also identified the OCT4 and OCT4-SOX2-TCF-NANOG (Wang et al., 2012) motifs at this subset of ESC-high NCREs, possibly contributing to the episomal activity of these NCREs. Interestingly, the repressor protein ZNF281 was enriched at a small subset of ESC-high, H3K4me2/3 positive NCREs (**Figure S4N**). ZNF281 is a known repressor protein expressed at high levels in ESCs but downregulated upon neural differentiation (Pieraccioli et al., 2018), and it will therefore be interesting to investigate whether ZNF281 or related factors might contribute to explain silencing of some of the ESC-high NCREs at the endogenous chromatin.

Experimental procedures

Cell culture

H9-derived human Neural Stem Cells (NSC, Gibco) were cultured as previously described (Perenthaler et al., 2020). Briefly, cells were seeded onto Geltrex (Gibco) coated plates and cultured in KnockOut DMEM/F12 (Gibco) supplemented with 2 mM L-Glutamine (Gibco), 2% StemPro neural supplement (Gibco), 20 ng/ml EGF (Peprotech), 20 ng/ml b-FGF (Peprotech), 100 U/ml penicillin and 100 µg/ml streptomycin. Cells resulted negative at mycoplasma testing.

H9 ESCs were cultured as previously described (Perenthaler *et al.*, 2020) on Matrigel (Corning) coated plates in MTeSR-1 medium (STEMCELL Technologies).

Chromatin immunoprecipitation

For chromatin immunoprecipitation, $5x10^7$ NSC were harvested in 9 mL of PBS and dual cross-linked first with 2mM Di(N-succinimidyl) glutarate (Sigma) for 45 min followed by 10 minutes 1.1% buffered formaldehyde (50mM Hepes-KOH pH7.6, 100mM NaCl, 1mM EDTA pH8.0, 0.5mM EGTA pH8.0, 11% formaldehyde). Formaldehyde was then quenched with 125 mM glycine (Sigma) for 10 min. Cross-linked cell pellets were snap-frozen and stored at -80°C.

ChIP was performed as described (Barakat *et al.*, 2018). Briefly, cell nuclei were sonicated at 4°C with Diagenode Bioruptor for 45 cycles (30 sec on, 30 sec off), prior to overnight incubation with antibody-coupled beads. The antibodies used were rabbit-anti-H3K4me1 (ab8895, Abcam), rabbit-anti-H3K27ac (ab4729, Abcam), rabbit-anti-YY1 (ab109237, Abcam) and goat-anti-SOX2 (AF2018, R&D systems). Immunoprecipitated chromatin and total input control were de-cross-linked in a final volume of 100 µL, by addition of 3 µL of 10% SDS and 10 µL proteinase K (10 mg/ml, Roche) and 5µl RNase (11119915001, 500 ng/µl, Roche) and incubation overnight at 65°C at 1400 RPM on a shaking thermomixer block. Following Phenol:Chloroform:IAA (Sigma) extraction and ethanol precipitation, the concentration of ChIP-DNA was determined by Qubit (Invitrogen) and the sonication was assessed by gel-electrophoresis of total input DNA (target fragment size between 200 and 600 bp).

ChIP-qPCR

For qPCR, 10µl of ChIP material and total input material were diluted 6 times. 2 µL of DNA were used per qPCR reaction, using iTaq universal SYBR Green Supermix in a CFX96RTS thermal cycler (Bio-Rad), following manufacturer's instructions. Fold enrichment was

determined following the ΔΔct method. All data shown are averages of at least 2 biological replicates and 2 technical replicates.

ChIP-seq, ChIP-STARR-seq plasmid library preparation

ChIP-seq and ChIP-STARR-seq plasmid libraries were generated as described (Barakat *et al.*, 2018). Briefly, 10 ng of ChIP DNA were end-repaired, dA tailed and adaptor ligated using NEBNext ChIP-seq library preparation kit (E6240, NEB), according to manufacturer's instructions. Adaptor ligated DNA was eluted into 32 µL of 0.1xTE, 25 µL were used for ChIP-seq library preparation and 5 µl were processed for ChIP-STARR-seq plasmid libraries preparation.

For ChIP-seq library preparation, purified adaptor ligated DNA was PCR amplified with Illumina index primers (E7335 and E7500, NEB) according to manufacturer's instructions. ChIP-seq libraries were assessed on an Agilent Tapestation. All sequencing occurred on an Illumina HiSeq 2500 platform, using 50bp single-end sequencing.

For details about the plasmid library preparation see Barakat *et al* (Barakat *et al.*, 2018). Briefly, 5 µL of purified adaptor ligated DNA for each replicate were separately amplified with primers 147 STARRseq libr FW and 148 STARRseq libr RV (Arnold *et al.*, 2013), and used in a Gibson assembly reaction with an AgeI-SalI digested STARR-seq plasmid and a home-made Gibson reaction buffer (100mM Tris-HCl, 10mM MgCl2, 0.2 mM dNTP, 0.5U Phusion DNA polymerase (NEB), 0.16U 50 T5 exonuclease (NEB)). Gibson reactions for the ChIP replicates were pooled together, purified by ethanol precipitation and used for electroporation into electrocompetent MegaX DH10b *E. coli* bacteria (C640003, Invitrogen), according to manufacturer's instructions. A total of 5 electroporation reactions per library were performed with 2 µl of DNA each. After recovery, bacteria from the 5 reactions were pooled together and 100 µL of a 1:100 and 1:10000 dilution was plated on Ampicillin containing Agar plates to enable estimation of the number of transformants. The remaining 5 mL of bacteria culture were incubated in 1 L of LB-media supplemented with Ampicillin over-night, and the plasmid DNA was isolated using a Maxiprep kit (Macherey-Nagel) according to manufacturer's instructions.

Transfection of plasmid libraries

For each transfection in NSC ChIP-STARR-seq, approximately 3 million NSC were seeded in a 10 cm dish and 48 hours later transfected with 6.8 µg of plasmid library DNA and 27.5 µl Lipofectamine Stem Transfection Reagent (STEM00015, Invitrogen) following manufacturer's instructions. In total, 5 dishes of transfected cells were used. For ESCs transfections in ESC

ChIP-STARR-seq, approximately 6 million cells were seeded in a 10 cm dish and transfected in suspension with an identical transfection mix to NSCs in the presence of 5µM ROCK-inhibitor (Y27632, Millipore). After 24h of transfection, cells were collected, and RNA was extracted with Tri-reagent (Sigma) following the manufacturer's instruction. A small fraction of cells was kept for estimating the fraction of GFP positive cells by flow cytometry.

ChIP-STARR-seq RNA and plasmid sequencing

ChIP-STARR-seq RNA libraries were prepared as described (Barakat *et al.*, 2018), with minor modifications. Briefly, RNA was isolated with Trizol (Sigma), following manufacturer's instruction and split into two replicates. The mRNA fraction was captured from each replicate of total RNA using Oligo (dT) 25 beads (61002, Life Technologies), DNaseI treated (18068-015, Life Technologies), and reverse transcribed with a GFP-mRNA specific primer (149 STARRseq rep RNA cDNA synth) following Superscript III protocol (18080-044, Life Technologies). To avoid plasmid DNA contamination, cDNA was PCR amplified in 10 parallel reactions for each replicate using primers spanning a synthetic intron (152 STARR reporter specific primer 2 fw and 153 STARR reporter specific primer 2 rv), as previously described (Arnold *et al.*, 2013). 35 ng of PCR amplified DNA were then used in a second round of PCR to add Illumina index primers (E7335, E7500, NEB). PCR was performed in 5 parallel reactions and using 2 independent index primers for each of the two replicates. In total, we sequenced 4 replicates of STARR-seq RNA for each pooled HIST or TF plasmid library. Corresponding plasmid libraries were similarly amplified using two independent Illumina index primers for each plasmid library. Quantity and quality of generated sequencing libraries was assessed on an Agilent Tapestation. All sequencing occurred on an Illumina HiSeq 2500 platform, using 50bp paired-end sequencing.

ATAC-seq

The Assay for Transposase Accessible Chromatin coupled to high-throughput sequencing (ATAC-seq) was performed as previously described (Buenrostro et al., 2013) with slight modifications. Briefly, approximately 50.000 single cells were lysed in Resuspension buffer containing 0,1% IgePal, 0,1% Tween-20 and 0,01% Digitonin. Immediately after lysis, the buffer was washed out by Resuspension buffer with only Tween-20. Nuclei were centrifuged at 500xg for 10 min at 4°C. Nuclei were re-suspended in 25 µl 2x TD buffer (Illumina), 2.5 µl TDE1 (transposase, Illumina), 16.5 µl PBS, 0.5 µl Tween-20, 0.5 µl Digitonin, and incubated for 30 min at 37°C. Afterward, the sample was purified using the MinElute PCR Purification Kit (QIAGEN) according to the manufacturer's protocol and eluted with 10 µl elution buffer. Samples were amplified using the NEBNext High Fidelity PCR master mix (New England Biolabs) and afterwards purified using a MinElute PCR purification kit (Qiagen) according to

the manufacturer's instructions. One microliter was loaded on an Agilent Technologies 2100 Bioanalyzer using a DNA 1000 assay to determine the library concentration and for quality check.

Cluster generation was performed according to the Illumina TruSeq SR Rapid Cluster kit v2 (cBot) Reagents Preparation Guide (www.illumina.com). Briefly, for sequencing libraries were pooled together to get a stock of 10 nM. One microliter of the 10 nM stock was denatured with NaOH, diluted to 10 pM and hybridized onto the flowcell. The hybridized products were sequentially amplified, linearized and end-blocked according to the Illumina Single Read Multiplex Sequencing user guide. After hybridization of the sequencing primer, sequencing-by-synthesis was performed using the HiSeq 2500 with paired end 50-cycle protocol followed by dual index sequencing. ATAC-seq libraries were sequenced to a depth of $20x10^6$ reads.

RNA-seq

For RNA sequencing of ESCs, we used our previously published data (Perenthaler *et al.*, 2020). For RNA sequencing of NSCs, RNA was isolated from two independent cultures following TRI reagent (Sigma) manufacturer's instruction. RNA was further purified using column purification (Qiagen, #74204). mRNA capture, library prep and barcoding were performed according to standard procedures of the Erasmus MC Biomics facility. Libraries were then single end sequenced, 50bp, on an Illumina HiSeq2500 machine at a depth of approximately 20 million reads per sample.

Data processing

The FASTQ files of RNA-, ChIP-, ATAC- and STARR-seq data were trimmed using Trimmomatic (version 0.39) (Bolger et al., 2014) to remove possible adaptor contamination and remove low quality reads. The trimmed RNA- and ATAC-seq data were mapped to the GRCh38/hg38 human genome build using the HISAT2 aligner (version 2.2.1) (Kim et al., 2015; Kim et al., 2019). For RNA-seq, aligned reads were counted for each gene using htseq-count (version 0.12.4) (Anders et al., 2015) and further processed using edgeR (Robinson et al., 2010) to normalize the data and perform downstream analysis. Bowtie2 (version 2.4.2) (with --very-sensitive parameter) (Langmead and Salzberg, 2012) was used to align both ChIP- and STARR-seq data against the GRCh38/hg38 human genome build. Only properly and uniquely mapped reads, with mapping quality more than 30 (MAPQ >=30), were kept followed by removing possible duplicated reads (only for ChIP-seq data) using Picard's

MarkDuplicates (version 4.0.1.1) (http://broadinstitute.github.io/picard/). The aligned reads were converted to bigwig format using ”bamCoverage” function with --binSize 10 and --normalizeUsing CPM parameters, the read coverage was then computed using “multiBigwigSummary” and pearson correlation between replicates was calculated using “plotCorrelation” via deepTools (version 3.4.3) (Ramirez et al., 2016). Peak calling was performed using MACS2 (version 2.1.2) (with -q 0.05) (Zhang et al., 2008) for each replicate of ESCs and NSCs separately, and took sequencing data of total input as background. Peaks overlapping blacklist regions defined by the ENCODE project (Accession: ENCFF356LFX, Date: 05.05.2020) were removed. The overlapping peaks between replicates were merged as common peaks using intersectBed (version2.30.0) (Quinlan, 2014; Quinlan and Hall, 2010).

NCRE activity analysis

To generate a non-redundant set of scaffold regions to assess activity of NCREs, ChIP-seq and RNA-plasmid peaks (H3K4me1, H3K27ac, YY1 and SOX2 of ChIP-seq and histone and TF plasmid DNA-seq libraries of STARR-seq) were merged after extending the peak summit to a width of 1000 bp (500 bp from each side). All regions larger than 1000 bp were split in half until all regions were at most 1000 bp in length, avoiding large regions and preserving high genomic resolution for later analysis. All further analysis for both NSC ChIP-STARR-seq and comparative ChIP-STARR-seq was performed on these scaffold regions.

To define active regions, reads under the defined scaffolds in both NSC and comparative ChIP-STARR-seq in ESCs were counted by FeatureCount (Liao et al., 2014). Read counts for each scaffold region were normalized by the total number of aligned reads in each library, after removing low coverage regions with less than 20 reads in at least two samples. The average number of normalized reads between replicates was taken for computing NCRE activity followed by the ratio of RNA-seq and DNA-seq (plasmid) reads as measurement of NCRE activity. The scaffold regions were ranked based on the average activity of RNA-seq samples in 5 different, equally sized, categories and the top-10% regions were defined as the most active NCREs in each cell line.

As NSC ChIP-STARR-seq we define those experiments that used NSC-derived plasmid libraries transfected in NSCs. As comparative ChIP-STARR-seq we refer to those data obtained from the transfection of NSC-derived plasmid libraries in ESCs. When discussing ESC ChIP-STARR-seq, we refer to our previous study that generated data obtained from transfections of ESC-derived plasmid libraries in ESCs (Barakat *et al.*, 2018). Furthermore, we defined cell-specific regions as the overlapping regions between active regions (top-10%)

of a given cell line and regions with less activity (bottom90%) of another cell line (i.e., from NSC ChIP-STARR-seq or comparative ChIP-STARR-seq in ESCs). We also considered the active regions that were in the top-10% of both NSC ChIP-STARR-seq and comparative ChIP-STARR-seq as common highly active regions.

NCRE-gene interaction

To integrate gene expression data, NCREs were linked to their target genes using HiC data and distance to the closest TSS. Previously published HiC data from the germinal zone (GZ) at gestation weeks 17–18 of 3 human foetal brains were used (Won *et al.*, 2016) to link NSC NCREs to target genes. This data provides 10 kb resolution bins for gene loop interactions and 40 kb resolution for TADs. Moreover, to link NCREs identified in comparative ChIP-STARR-seq in ESCs to their target genes, we used published HiC data from H9 human ESCs with 25 kb resolution bins for both gene loop interactions and TADs (Freire-Pritchett *et al.*, 2017).

NCREs were then intersected with pre-calculated significant interactions using intersectBed to define NCRE-gene interaction. NCRE-gene interactions located within the same TAD were considered for downstream analyses. The coordinates of HiC data were liftedOver to the GRCh38/hg38 human genome build before intersecting with regions. In addition, NCREs were annotated for target genes based on the distance to nearest TSS using the “annotatePeaks.pl” function of HOMER (version 4.11) (Heinz et al., 2010). Finally, all NCRE-gene interactions either using HiC or distance to closest TSS were considered for further analysis. In total, we found 28,514 target genes (coding and non-coding) for NSCs and 20,347 target genes for ESCs, of which 19,739 target genes are identified based on the closest TSS and the other target genes are based on HiC.

We only included protein-coding genes in our analysis. Gene expression levels were plotted and statistical comparison was performed using Wilcoxon signed rank test in R. The *Online Mendelian inheritance in Man* (OMIM) gene list (updated 21-04-2022) was downloaded from the OMIM database to further assist in annotation of target genes.

Gene ontology analysis

The web interface of Enrichr (Chen *et al.*, 2013; Kuleshov *et al.*, 2016) (https://maayanlab.cloud/Enrichr/) was used for functional enrichment analysis using the default settings, and the whole genome was set as background. Wordcloud (version 2.6) (Fellows et al., 2018) was used to visualize the GO enrichment.

Functional enrichment

NCREs were annotated for genome features using the “annotatePeaks.pl” function of HOMER (version 4.11) (Heinz *et al.*, 2010). Motif enrichment analysis was performed using the “findMotifsGenome.pl” function of HOMER (Heinz *et al.*, 2010) with default setting and whole genome as background. Significant enriched motifs were selected based on p-value ≤ 0.01.

Sequence characteristics analysis

To determine whether different DNA sequence features distinguish defined regions between NSCs and ESCs, we considered the following features: (i) the non-coding essential regulation (ncER) score (https://github.com/TelentiLab/ncER_datasets/; updated 06-03-2019) (Wells *et al.*, 2019); (ii) GC content, as determined by the GCcontent R packages based on BSgenome. Hsapiens.UCSC.hg38 (version 1.4.3); (iii) conservation score for each NCRE, as derived from the gscores R packages based on phastCons100way.UCSC.hg38 (version 3.7.2) (Siepel *et al.*, 2005); (iv) Orion scores (Gussow *et al.*, 2017); (v) CADD scores (Kircher *et al.*, 2014); (vi) Haploinsufficiency scores (Xu *et al.*, 2020) and (vii) probability of loss-of-function intolerance (pLI) score (Lek et al., 2016). The overlaps between DNA sequence features and regions were defined using intersectBed. The coordinates of data were converted to the GRCh38/hg38 human genome build by liftOver prior to intersections. The scores of the genes linked to the regions were plotted to show pLI distribution. Statistically significant differences between groups were determined using the Wilcoxon signed rank test in R.

Transposable element enrichment

The RepeatMask (GRCh38/hg38, updated 03-09-2021) was downloaded from the UCSC table browser and intersected to the regions. To determine enrichment of transposable elements in NCREs, we followed a strategy previously used for investigating active NCREs in human embryonic stem cells (Barakat *et al.*, 2018). The number of overlaps of each type of repeat (n_overlaps) with all regions (n) was used to calculate the relative frequency (f_all = n_overlaps/n). Multiplication of the relative frequency with the number of regions (n_test, e.g., ESC, NSC etc.) in any tested group yields the expected frequency (E). This number was compared with the actual observed frequency in the subgroups (f_test = (n_overlap, test)/n_test = O) to calculate the observed versus expected ratio (O/E). We considered repeats with O/E > 2 as enriched. For the subsequent data interpretation, we only focused on transposable elements that were present multiple times (n_overlap > 15).

Epigenome profiling of NCREs

To find the enrichment of endogenous marks around NCREs in ESC and NSC, we collected various ChIP-seq data from previously published studies (Barakat *et al.*, 2018; Borsari et al., 2021; Chan *et al.*, 2013) and from the ENCODE project (Zhang et al., 2020).

The FASTQ files of ChIP-seq data was re-analyzed along with ATAC-seq data (from this study) based on what we described in the "*Data processing*" section.

Heatmaps for differentially active regions across the endogenous marks were created using deeptools (version 3.4.3) (Ramirez *et al.*, 2016). We first converted bam files of endogenous marks into bigwig format using the "bamCoverage" function with --binSize 10 and --normalizeUsing RPKM parameters. The bigwig files were then used to compute the number of reads across 6 kb centered on NCREs using "computeMatrix" with --referencePoint center, --upstream 3000 and --downstream 3000 parameters. Finally, the line plots and heatmaps were generated using the "plotProfile" and "plotHeatmap" functions, respectively.

Construction of BRAIN-MAGNET

To generate BRAIN-MAGNET, we first removed from the analysis scaffold of 148,198 genomic regions assessed in ChIP-STARR-seq a small number of patch sequences resulting in a final number of 148,114 genomic sequences. These ranged from 500 to 1000 base pairs in length, and were augmented with their reverse complement to yield a total of 296,228 examples. A random 10% of the augmented sequences were held out for both validation and testing purposes. The DNA sequences were extracted from the human genome using bedtools getfasta (version 2.30.0) and their reverse complements were obtained using revseq of EMBOSS (version 6.6.0.0) (Rice et al., 2000). The input sequences were then converted into an one-hot encoded matrix of length 1000, with each nucleotide being represented by a binary vector (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]). Sequences shorter than 1000 bp were padded with zeros. The corresponding sequence activities served as outputs of the model. We subsequently used input and output in BRAIN-MAGNET, a single task convolutional neural network, to predict either activities of NCREs identified by NSC ChIP-STARR-seq or by comparative ChIP-STARR-seq in ESCs. The model is composed of three convolutional layers (filters = 128, 256, 512; size = 11, 9, 7), each followed by batch normalization, a ReLU non-linearity, and MaxPool (size = 2). After the convolutional layers, there are two fully connected layers, each with 1024 neurons, followed by batch normalization,

a ReLU non-linearity, and dropout where the fraction is 0.4. The final layer is mapped to either NSC or comparative ChIP-STARR-seq outputs.

Hyperparameters were manually adjusted to optimize the performance on the validation set. The model was implemented and trained in Pytorch (version 1.13.1 with CUDA version 11.7) (Paszke et al., 2019) using the Adam optimizer with a learning rate of 0.01, mean squared error (MSE) as the loss function, a batch size of 128, and early stopping criterion with a patience of 20 on 100 epochs. Model training, hyperparameter tuning, and performance evaluation were conducted on different sets of genomic regions.

The performance of the model was evaluated separately for NSC and ESC predictions on the held-out test sequences. We used the Pearson and Spearman correlation coefficient (R and Rho) across all bins for a quantitative genome-wide evaluation, the area under the precision-recall and receiver operating characteristic curve (PR and ROC; calculated using pr.curve and pr.roc functions from R package PRROC (version 1.3) (Grau et al., 2015) for NCRE classification (Top10 vs. Bottom10 from the test set).

Motif discovery using TF-MoDISco-lite

DeepExplainer (Lundberg and Lee, 2017), a DeepSHAP implementation of DeepLIFT, was used to compute contribution (cb) scores for all nucleotides of the assessed NCREs, which reflects the individual contribution of each nucleotide to the NCRE activity in either NSCs or ESCs. 100 dinucleotide-shuffled versions of each input sequence were used as reference sequences (the fork from https://github.com/kundajelab/shap/blob/master/shap/explainers/deep/deep_pytorch.py). The obtained hypothetical importance scores were then multiplied by the one-hot encoded matrix of the sequences to derive the actual nucleotide cb scores, which were visualized using the ggseqlogo (version 0.2) (Wagih, 2017). The cb scores of each NCRE can be visualized using the R-shinny app for which the code is available in our github repository.

The hypothetical importance scores, along with the corresponding one-hot encoded matrix of the sequences, were subjected to TF-MoDISco-lite (https://github.com/jmschrei/tfmodisco-lite), a more efficient and user-friendly version of TF-MoDISco (Shrikumar *et al.*, 2018), to identify the most relevant motifs associated with changes of each nucleotide. The TF-MoDISco-lite algorithm in default settings was used to extract and cluster similar seqlets across all selected sequences, the identified seqlets were matched against JASPAR 2022 CORE vertebrate non-redundant database (https://jaspar.genereg.net/download/data/2022/CORE/JASPAR2022_CORE_non-redundant_pfms_jaspar.zip) by Tomtom (version 5.5.1) (Bailey et al., 2009) within TF-

MoDISco-lite (Shrikumar *et al.*, 2018). To further facilitate the interpretation of the cb scores across the genome, we calculated normalized cb scores taking either a normalisation of all NCREs (cb_all) or normalisation within each NCRE (cb_each) into account, and calculated percentile scores for these.

GWAS analysis

To test the utility of BRAIN-MAGNET cb scores for the functional interpretation of GWAS loci, we made use of a recent MPRA data set provided by Guo et al, which tested 2,221 non-coding variants associated with 10 neuropsychiatric disorders by MPRA, with an average of 5 SNPs per disease risk loci being tested (Guo *et al.*, 2023). Comparing the activity of matched risk versus reference allele pairs, the authors identified 892 SNPs that showed differential activity in their MPRA depending on the genotype of the given SNP, which the authors refer to as daSNVs. We retrieved the daSNVs and non-daSNVs from Supplementary Data 3 and 5 of Guo et al (Guo *et al.*, 2023) and intersected these with NCREs and cb scores from our data set. We subsequently plotted the distribution of cb scores (>60 cb_all) over the daSNV and non-daSNVs, and used the BRAIN-MAGNET cb scores to fine map two previously identified GWAS loci for which multiple SNPs in LD were assessed in the Guo et al study and present in multiple NCREs from our data set.

Genomic variant analysis

For the gnomAD variant analysis, the curated VCFs of gnomAD v4 were downloaded from https://gnomad.broadinstitute.org/downloads#v, which contains 909,084,110 short variants mapped to the GRCh38 genome build obtained from whole-genome sequencing of 76,215 individuals. bcftools (v1.8) was used to keep high-quality variants ("PASS" flag) and rare variants (AF < 0.1%). tabix (v1.9) was used to extract the variants overlapping our NCREs. The total number of high-quality rare variants was calculated without chromosome Y and mitochondrial DNA (chromosome M), as these sequences were not covered in our NCRE analysis scaffold. The VISTA enhancers, ENCODE_cCREs, fiveUTR, threeUTR and codingExon regions (GRCh38/hg38, updated 14-02-2024) were downloaded from the UCSC table browser. DAE and nDAE regions were retrieved from Supplementary Table 3 of Yousefi et al (Yousefi *et al.*, 2021). The total number of filtered rare variants (MAE <0.1) genome wide (variant_genome) was normalized by the size of the genome (size_genome) and used to calculate the expected frequency (E = n_variant/size_ genome). The number of filtered rare variants located in the target regions (listed in **Figure S5A**) (variant_region) was normalized by the size of the target region (size_region) to calculate the observed frequency (O =

variant_region/size_region). The O/E ratio was calculated by comparing the observed frequency over the expected frequency.

For the *de novo* variant analysis from the Genomics England 100,000 Genomes Project (GEL), *de novo* variants of GEL data version 18 (main-programme_v18_2023-12-21) were retrieved by Labkey from the *de novo*_flagged_variants table in the GRCh38 genome build. Next, high confidence *de novo* variants (stringent_filter = 1) were kept. Then, solved cases, labelled with “yes”, “unknown” and “partically” in the gmc_exit_questionnaire table, were removed. Finally, 4,415 individuals with unexplained neurology and neurodevelopmental disorders were kept, for which 13,276 *de novo* variants overlapped with our NSC NCREs. Also we analysed 4,558 genomes from individuals with a variety of other rare disease categories (including cardiovascular disorders, hearing an ear disorders, ultra-rare disorders, dermatological disorders, gastroenterological disorders, renal and urinary tract disorders, respiratory disorders, endocrine disorders, dysmorphic and congenital abnormality syndromes, growth disorders, skeletal disorders, ophthalmological disorders, metabolic disorders, hematological and immunological disorders and rheumatological disorders), harbouring 13,564 *de novo* variants overlapping NSC NCREs. The *de novo* variants were further annotated by the cb scores, and the cb_each distribution of NCREs were plotted.

For the GEL analysis using variants overlapping with high-confidence motifs, we used 20 bp sequences with coordinates centred on the high-confidence motifs called by TF-MoDISco (q-values < 0.05) from the NSC NCRE category 4 and 5. An aggregation of gVCFs (AggV2, only available in version 10) comprising 78,195 genomes from GEL harbouring over 722 million variants was intersected with these high-confidence motif coordinates. Only variants marked with the “PASS” flag, indicating high quality, were retained, resulting in 11,788 variants overlapping 2,980 motifs. We next retained only variants in unsolved individuals (55,379 genomes, 9,787 variants) and subsequently individuals with neurodevelopmental disorders (10,536 genomes, 4,259 variants). Next we retained only variants with allele count (AC) <= 10 (2,700 genomes, 2,684 variants). Finally, only variants located in NCREs with target genes recorded as OMIM disease genes were kept, resulting in a final list of 705 variants from 824 genomes that overlapped 440 motifs.

For screening of an independent WGS trio data set, we used 330 trio WGS from individuals affected by rare disorders from the Institute of Medical Genetics and Appliced Genomics (IMGAG) database from Tübingen, Germany. Removing those individuals in which already a definitive molecular diagnosis was found (n=81) left 249 individuals, of which 60 harbored a candidate disease explaining variant. In those 249 individuals, a total of 16,437 *de novo* variants was found. Intersecting these variants with the NCRE atlas and filtering for variants

with a cb score >95th percentile resulted in 70 *de novo* variants that were manually curated for potential clinical phenotype matches.

Experimental validation of wild type and mutant NCREs

For experimental validation, we randomly selected 15 NCREs of which the target gene was expressed in NSCs. NCREs were amplified from genomic DNA and cloned into the STARR-seq reporter plasmid (kind gift of A.Stark) (Arnold *et al.*, 2013) as previously described (Barakat *et al.*, 2018). Subsequently, site-directed mutagenesis was used to generate NCRE deletions or point mutations in the generated plasmids. NSCs were then transfected with wild type or mutant NCRE plasmids using polyethylenimine (PEI, Sigma) or Lipofectamine™ Stem Transfection Reagent (Thermo Scientific) respectively. Twenty-four hours post transfection cells were collected, stained with Hoechst dye and the NCRE activity was measured by FACS analysis (20,000 cells per sample). GFP-positive cells within the mCherry-positive population were quantified to assess NCRE activity compared to an empty STARR-seq vector control. Two independent transfection experiments were performed, each in duplicates. Statistical analysis was performed using a one-way ANOVA test followed by multiple comparison test (Fisher's LSD test). Calculations were conducted in GraphPad Prism (version 8).

Figures and legends

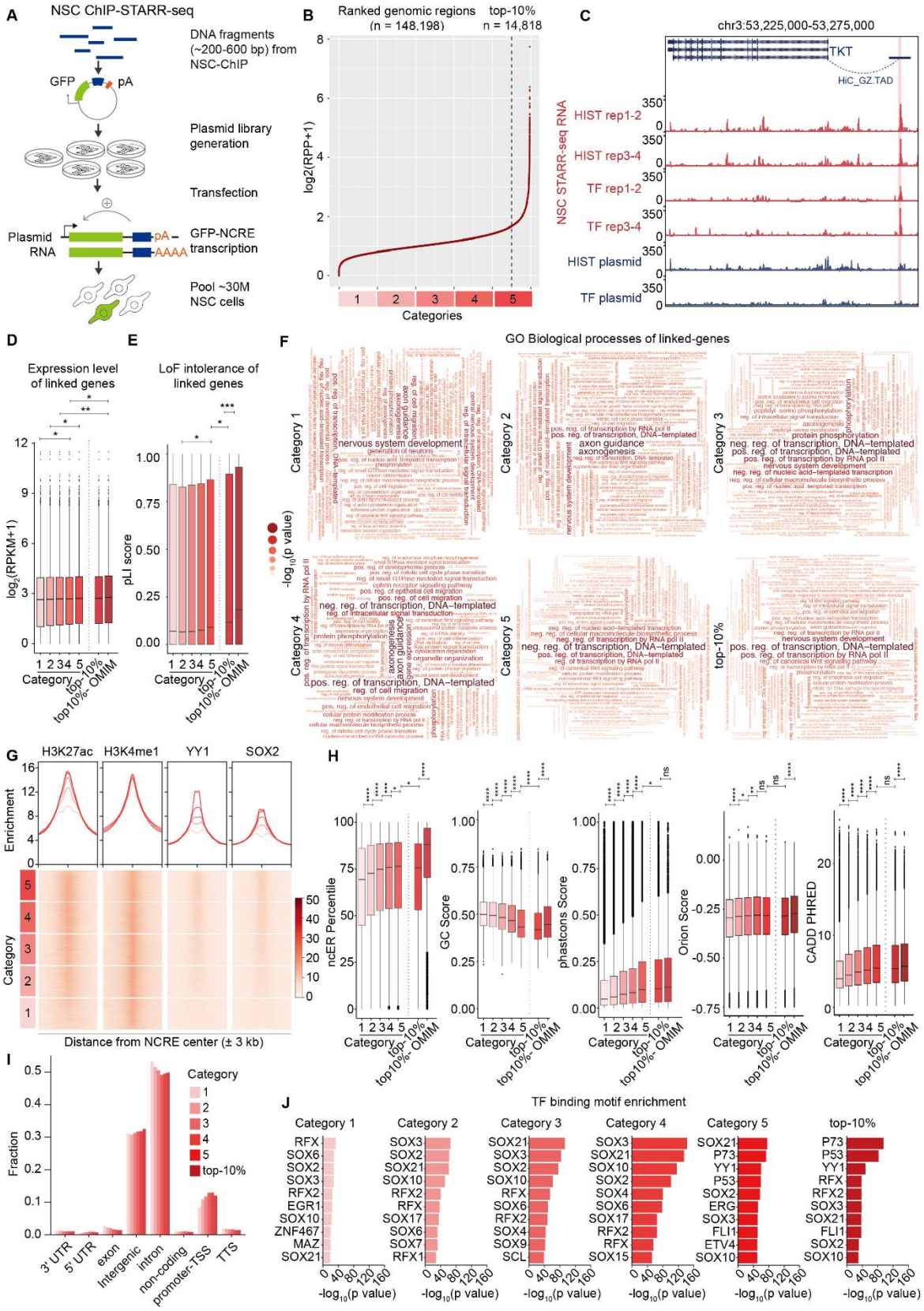


Figure 1: ChIP-STARR-seq in neural stem cells identifies NCREs with different activity levels and sequence characteristics

A) Schematic representation of the experimental strategy used in this study.

B) Line plot showing NCRE activity (enrichment of ChIP-STARR-seq RNA over plasmids per region of the scaffold; $\log_2$) of the 148,198 assessed scaffold regions, ranked from lowest to highest activity. In gradient of red are depicted the 5 categories (groups of 20 percentiles) of increasing NCRE activity detected by NSC ChIP-STARR-seq used throughout this study. The dotted line demarcates the top-10% most active NCREs.

C) Genome browser tracks surrounding the *TKT* locus showing a distal NCRE showing activity by NSC ChIP-STARR-seq and interaction of this NCRE with the *TKT* gene confirmed by HiC from foetal brain. Blue tracks: sequencing of plasmid DNA. Red tracks: sequencing of ChIP-STARR-seq RNA.

D) Expression level ($\log_2$(RPKM+1)) of the genes linked to each of the 5 NCRE activity categories, to the top-10% most active NCREs and of the OMIM disease-related genes linked to the top-10% most active NCREs. Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent outliers. (* $p < 0.05$; ** $p < 0.01$, Wilcoxon test).

E) As in D) showing the pLI score of the genes linked to the different NCRE categories. (* $p < 0.05$; *** $p < 0.001$ Wilcoxon test).

F) Word cloud representation of gene ontology analysis using Enrichr of biological processes for the protein-coding genes linked to the NCREs of the 5 categories and the top-10% most active NCREs.

G) Top: line plot showing the average enrichment of ChIP-seq signal from NSCs across the genomic regions encompassed by NCREs from the 5 categories of NCRE activity (bins of 3 kb up and down centred at the middle of the NCRE). Bottom: corresponding heatmaps of H3K27ac, H3K4me1, YY1 and SOX2 ChIP-seq signals in NSCs across the 148,198 assessed scaffold regions ranked by activity from the highest to the lowest.

H) Box plots, showing from left to right ncER percentile, GC content score, phastcons score, Orion score, and CADD score, for all the 5 NCRE activity categories, the top-10% most active NCREs (top-10%) and the top-10% most active NCREs linked to genes with a known OMIM phenotype. Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent the outliers. (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$; ns not significant; Wilcoxon tests).

I) Bar graphs showing the fraction of NCREs per genomic region using Homer, for all NCRE categories. UTR = untranslated region; non-coding = non-coding RNAs such as snRNA, lncRNA, etc; TSS = Transcriptional start site; TTS = Transcriptional termination site.

J) Bar graph showing the 10 most enriched TF motifs in the NCREs belonging to the different activity categories in NSCs. Plotted is the -log10 p-value.

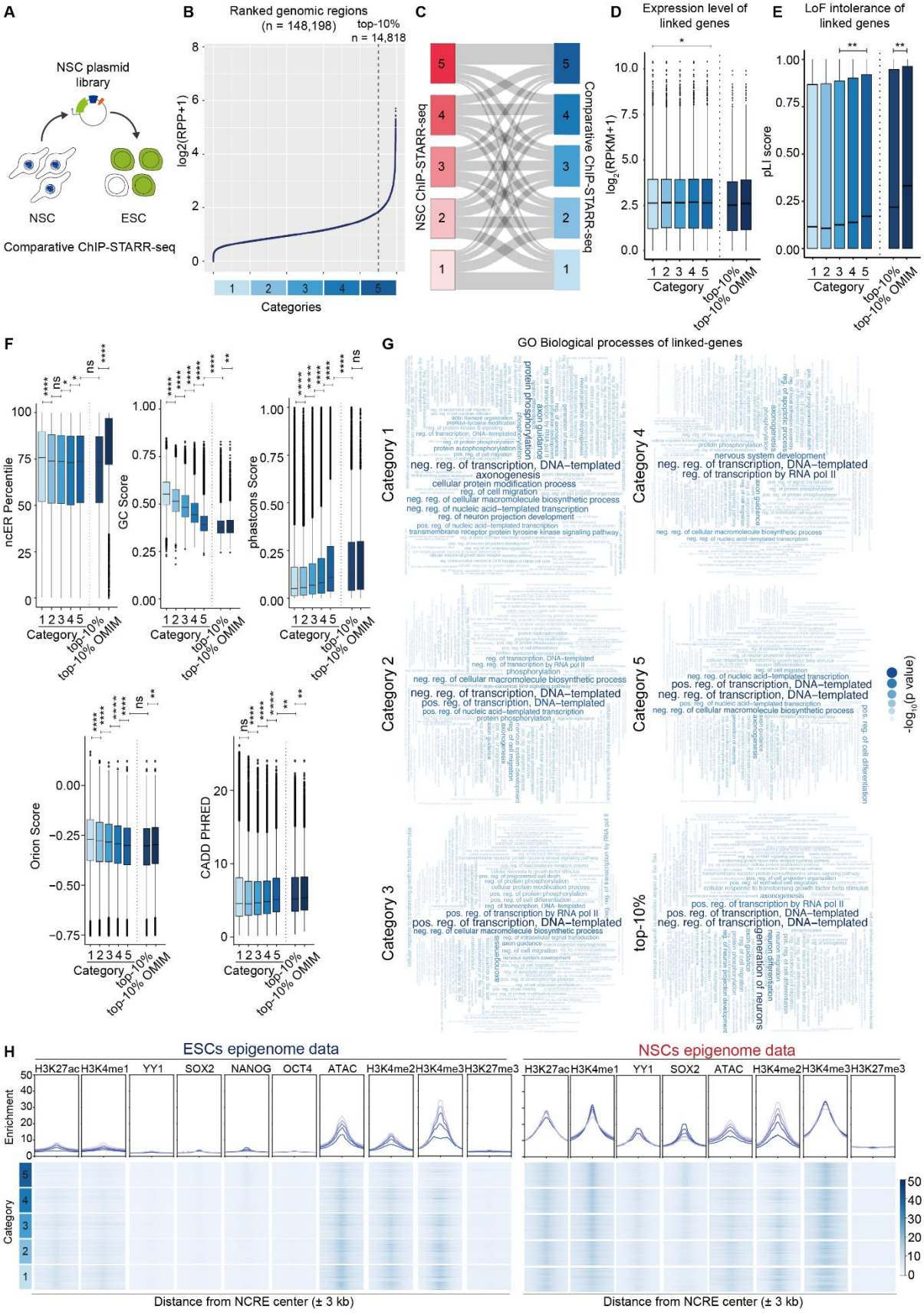


Figure 2: Comparative ChIP-STARR-seq using NSC derived plasmid libraries in ESCs identifies NCREs which are primed for future NCRE activity

A) Schematic representation of the experimental strategy used for transfecting NSC derived ChIP-STARR-seq libraries in ESCs, referred to as comparative ChIP-STARR-seq.

B) Line plot showing NCRE activity (enrichment of ChIP-STARR-seq RNA over plasmids per region of the scaffold; log_2) of the 148,198 assessed scaffold regions, ranked from lowest to highest activity from comparative ChIP-STARR-seq in ESCs. In gradient of blue are depicted the 5 categories (groups of 20 percentiles) of increasing NCRE activity used throughout this study. The dotted line demarcates the top-10% most active NCREs.

C) Sankey plot showing the changes between activity categories for the NCREs measured in NSC ChIP-STARR-seq (red) and in comparative ChIP-STARR-seq (blue). For the latter, we transfected the NSC-derived plasmid libraries in ESCs.

D) Expression level (log_2(RPKM+1)) of the genes linked to each of the 5 NCRE activity categories, to the highly active NCREs (top-10%) and to the highly active NCREs linked to genes with a known OMIM phenotype. Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent outliers. (* $p < 0.05$; Wilcoxon test).

E) As in D) showing the pLI score of the genes linked to the different NCRE categories. (** $p < 0.01$; Wilcoxon test).

F) Box plots as in D), showing from left to right on the upper row ncER percentile, GC content score, phastcons score and on the lower row Orion score, and CADD score. (* $p < 0.05$; ** $p < 0.01$; **** $p< 0.0001$; ns not significant; Wilcoxon tests).

G) Word cloud representation of gene ontology analysis using Enrichr of biological processes for the protein-coding genes linked to the NCREs of the 5 categories and the top-10% most active NCREs.

H) Epigenome profiling of NCREs with comparative ChIP-STARR-seq activity in ESCs (plasmid libraries derived from ChIP experiments in NSCs). Top: line plot showing the average enrichment of ChIP-seq (or ATAC-seq) signal from ESCs (left panel) or NSCs (right panel) across the genomic regions encompassed by NCREs from the 5 categories of NCRE activity from comparative ChIP-STARR-seq in ESCs (bins of 3 kb up and down centred at the middle of the NCRE). Bottom: corresponding heatmaps of H3K27ac, H3K4me1, YY1, SOX2, NANOG, OCT4, H3K4me2, H3K4me3 and H3K27me3 ChIP-seq signals and ATAC-seq in ESCs (left), or H3K27ac, H3K4me1, YY1, SOX2, H3K4me2, H3K4me3 and H3K27me3 ChIP-seq signals and ATAC-seq in NSCs (right) across the 148,198 assessed scaffold regions

ranked by activity as measured by comparative ChIP-STARR-seq in ESCs from the highest to the lowest.

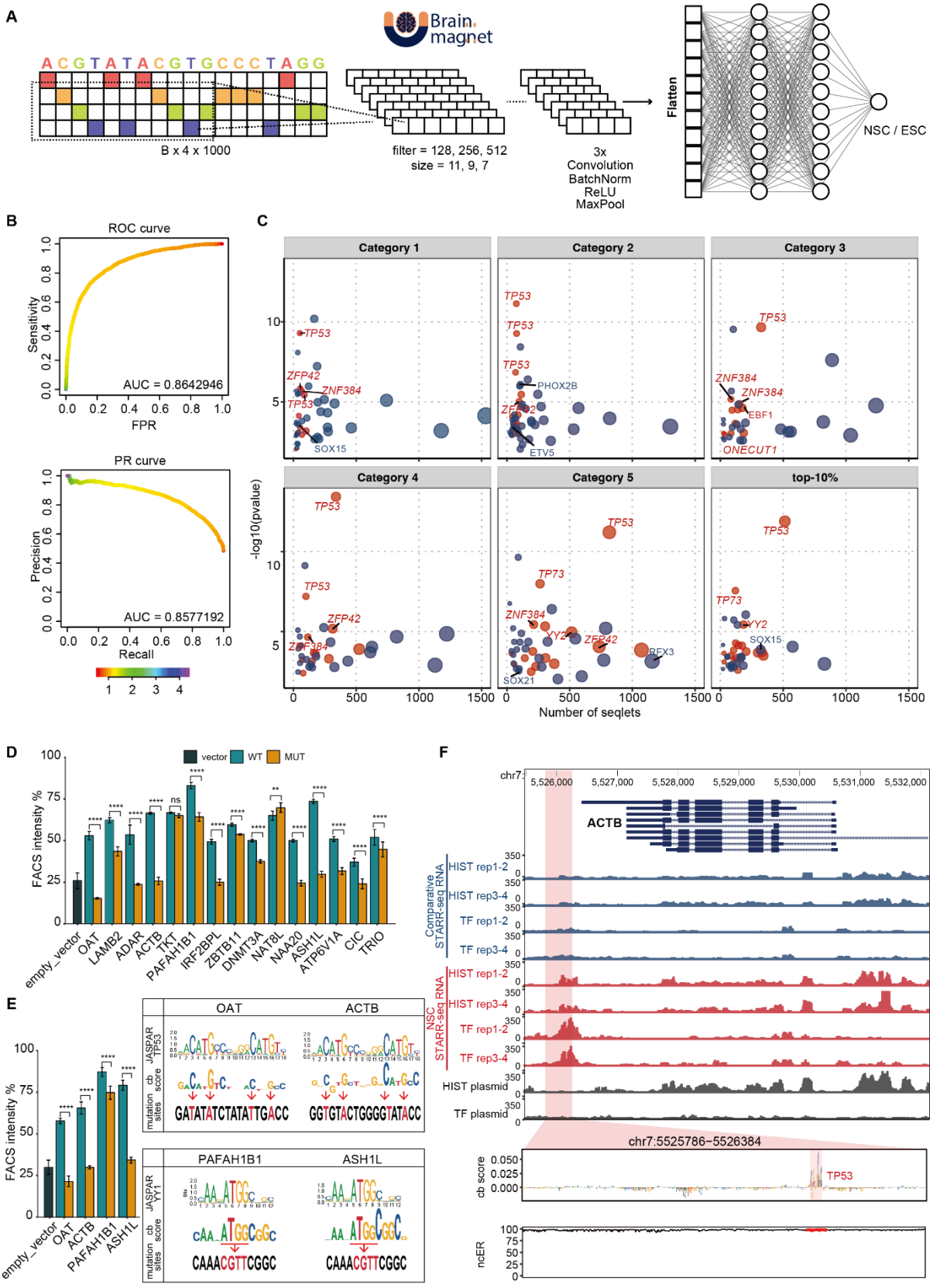


Figure 3: BRAIN-MAGNET, a sequence based convolutional neural network model, predicts NCRE activity and facilitates interpretation of biological consequences of genomic variation

A) BRAIN-MAGNET is a deep convolutional neural network architecture trained to predict NCRE activity from an input sequence of the tested element.

B) Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves of comparing Top10 and Bottom10 NCREs from the test set using the R package PRROC (Grau *et al.*, 2015) for NSCs. The PR curve reflects the precision-recall trade-off, and the ROC curve portrays the true positive and false positive rate trade-off. The color scale maps classification thresholds to points on the curve, providing visual guidance for selecting an optimal threshold. (Recall = TP / (TP + FN), Precision = TP / (TP + FP), True positive rate = TP / (TP + FN), False positive rate = FP / (FP + TN))

C) Enrichment of TF motifs in the various NSC NCRE activity categories discovered by TF-MoDISco-lite (Shrikumar *et al.*, 2018). The number of seqlets compared with significance (-log10 p-value) for the most matched TF motif per motif cluster. Red dots denote enriched motifs, and blue dotes represent depleted motifs. The motifs with FDR < 0.05 are highlighted as *italic*.

D) Bar plot showing the percentage of GFP+ cells in NSCs upon cell transfection experiments with reporter plasmids containing either wild type (green) or motif deleted NCREs (orange) (n=15) or empty reporter plasmid control. NCREs are indicated with the name of the presumed target gene. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

E) As D), but now showing 4 NCREs with point mutations of the predicted functional relevant motifs. Right insert shows the predicted motifs from JASPAR, a visual representation of the cb scores for each nucleotide and the created point mutations of the motifs that were tested.

F) Genome browser view showing the *ACTB* locus and the upstream NCRE tested in panel D and E for which BRAIN-MAGNET predicts a functional motif. Top panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of plasmid (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted.

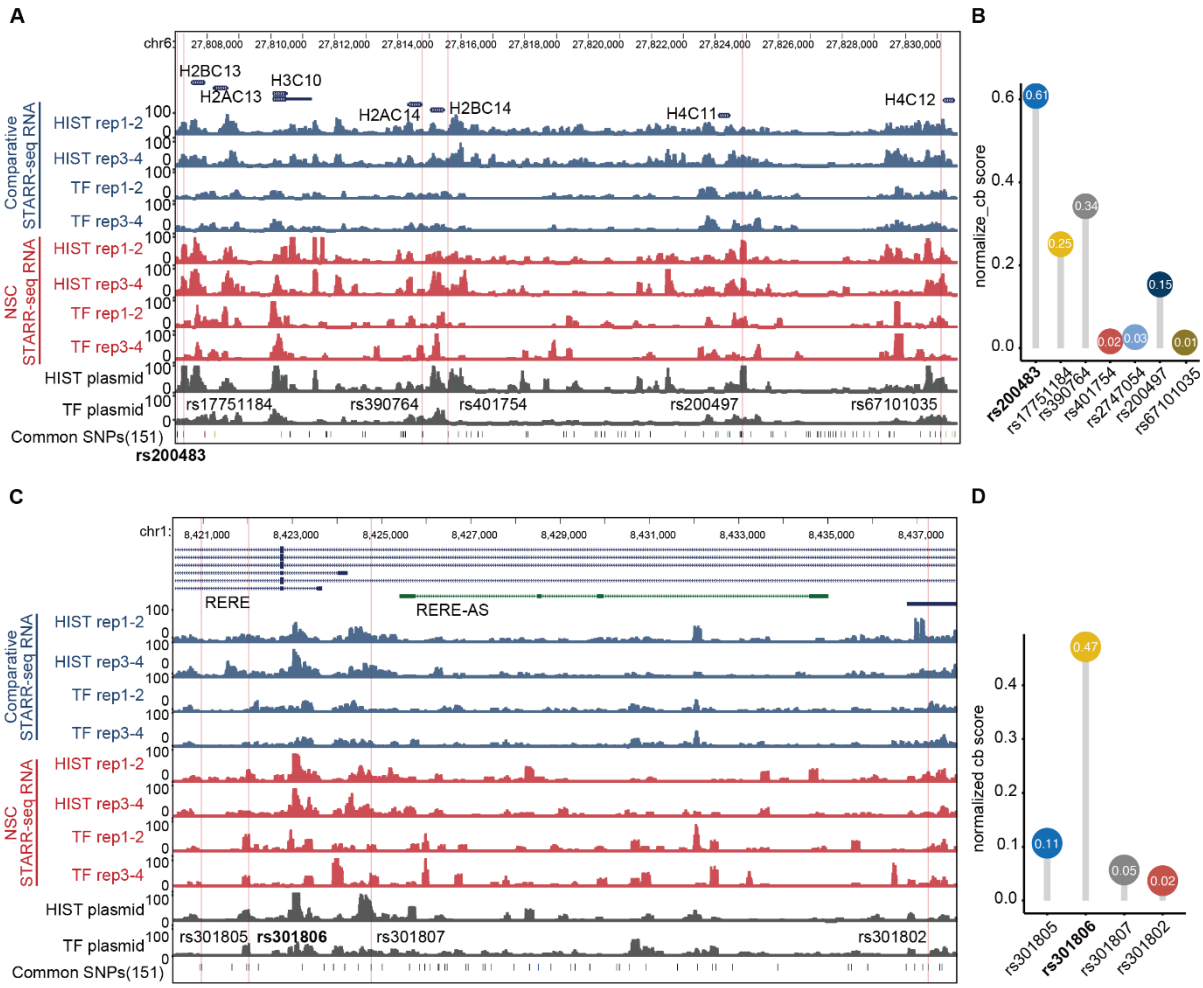


Figure 4: Utility of the NCRE atlas and BRAIN-MAGNET for prioritizing genomic variants found at GWAS associated regions

A) Genome browser view showing ChIP-STARR-seq RNA for ESCs (blue) and NSCs (red) and plasmid library sequencing tracks of a locus on chromosome 6 which GWAS studies have shown to be associated with increased risk for schizophrenia. 7 associated SNPs in LD which were also present on NCREs from our study are highlighted.

B) Line plot showing the BRAIN-MAGNET cb scores (normalized cb_all score) for the 7 SNPs found at the schizophrenia risk locus from A). The rs200483 SNP has relatively the highest normalized cb score compared to the other 6 SNPs, and indeed rs200483 is one of the identified daSNVs from the MPRA studies by Guo et al (Guo *et al.*, 2023).

C) As A), but now showing a GWAS locus associated with major depression on chromosome 1, with 4 GWAS associated SNPs in LD that overlap with NCREs in our study highlighted.

D) As B), but now showing the BRAIN-MAGNET cb scores (normalized cb_all score) for the 4 SNPs found at the major depression risk locus from C). Again, the daSNV (rs301806) found

by Guo et al (Guo *et al.*, 2023) has the relatively highest cb score amongst the 4 candidate SNPs.

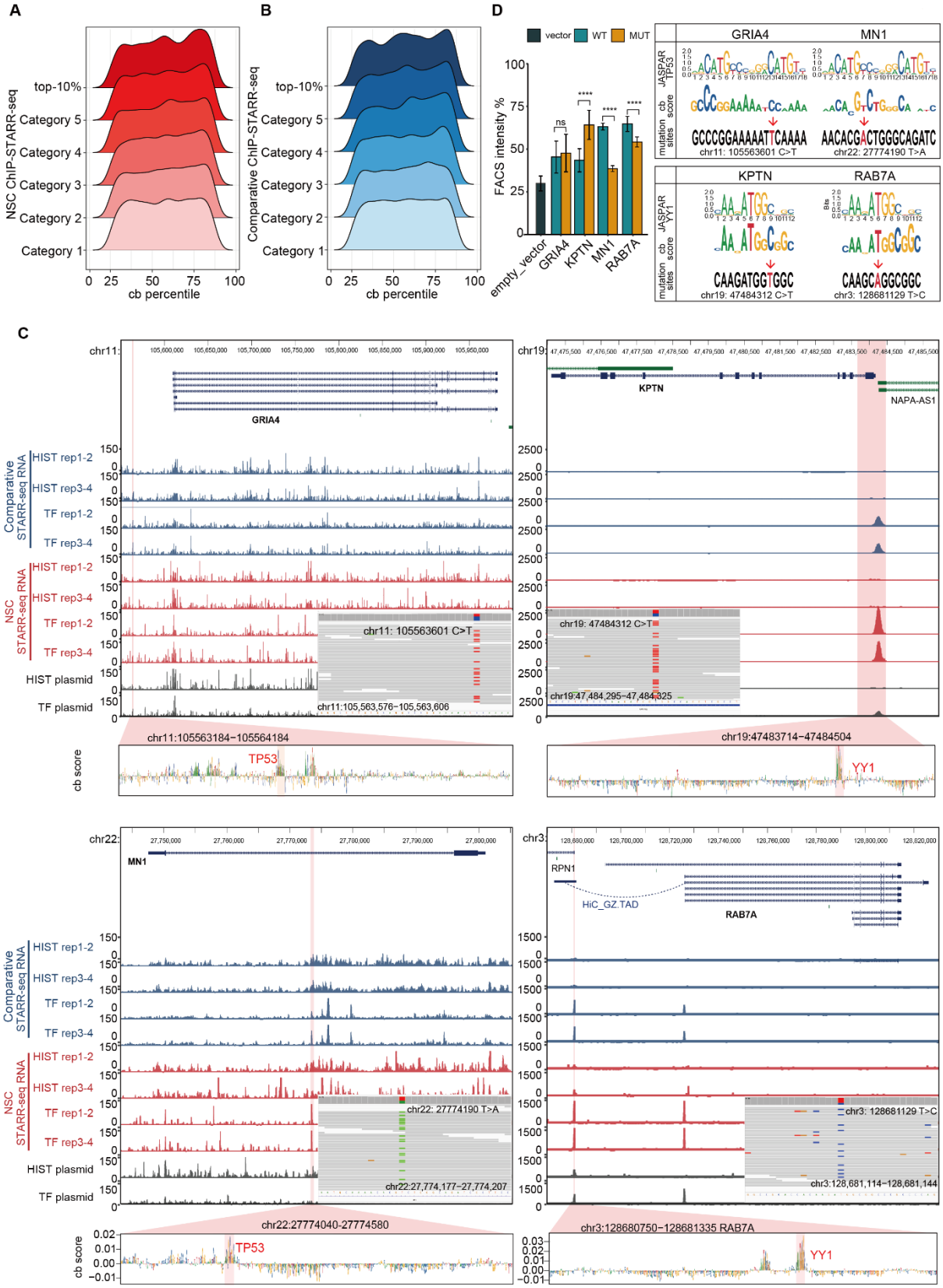


Figure 5: Utility of the NCRE atlas and BRAIN-MAGNET for prioritizing genomic variants found in rare disease

A) Density plots depicting the number of *de novo* variants found in 4,415 genetically unexplained individuals with neurological and neurodevelopmental disorders from the Genomics England 100,000 Genomes project, located in NCREs from the different NCRE categories for NSCs stratified for their BRAIN-MAGNET cb percentile.

B) As A), but now for ESC NCREs.

C) Genome browser views of the loci of *GRIA4* (left, top), *KPNT* (right, top), *MN1* (left, bottom) and *RAB7A* (right, bottom), showing ChIP-STARR-seq RNA for ESCs (blue) and NSCs (red) tracks and plasmid library sequencing results. The identified NCREs are highlighted (with HiC interactions indicated if available) and the zoom-in shows BRAIN-MAGNET cb scores and predicted functional motifs. The inserts show IGV views from heterozygous rare variants that affect the NCRE predicted functional motifs, detected in genetically unexplained individuals affected by neurological and neurodevelopmental disorders from the Genomics England 100,000 Genomes project.

D) Bar plot showing the percentage of GFP+ cells in NSCs upon cell transfection experiments with reporter plasmids containing either empty reporter plasmid control (black), wild type NCREs from C) (green) or the same NCREs with patient mutations found in the Genomics England 100,000 Genomes project (orange). The motifs predicted from JASPAR, the BRAIN-MAGNET cb score and the patient mutation are indicated in the right panel. NCREs are indicated with the name of the presumed target gene. Plotted on the left is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

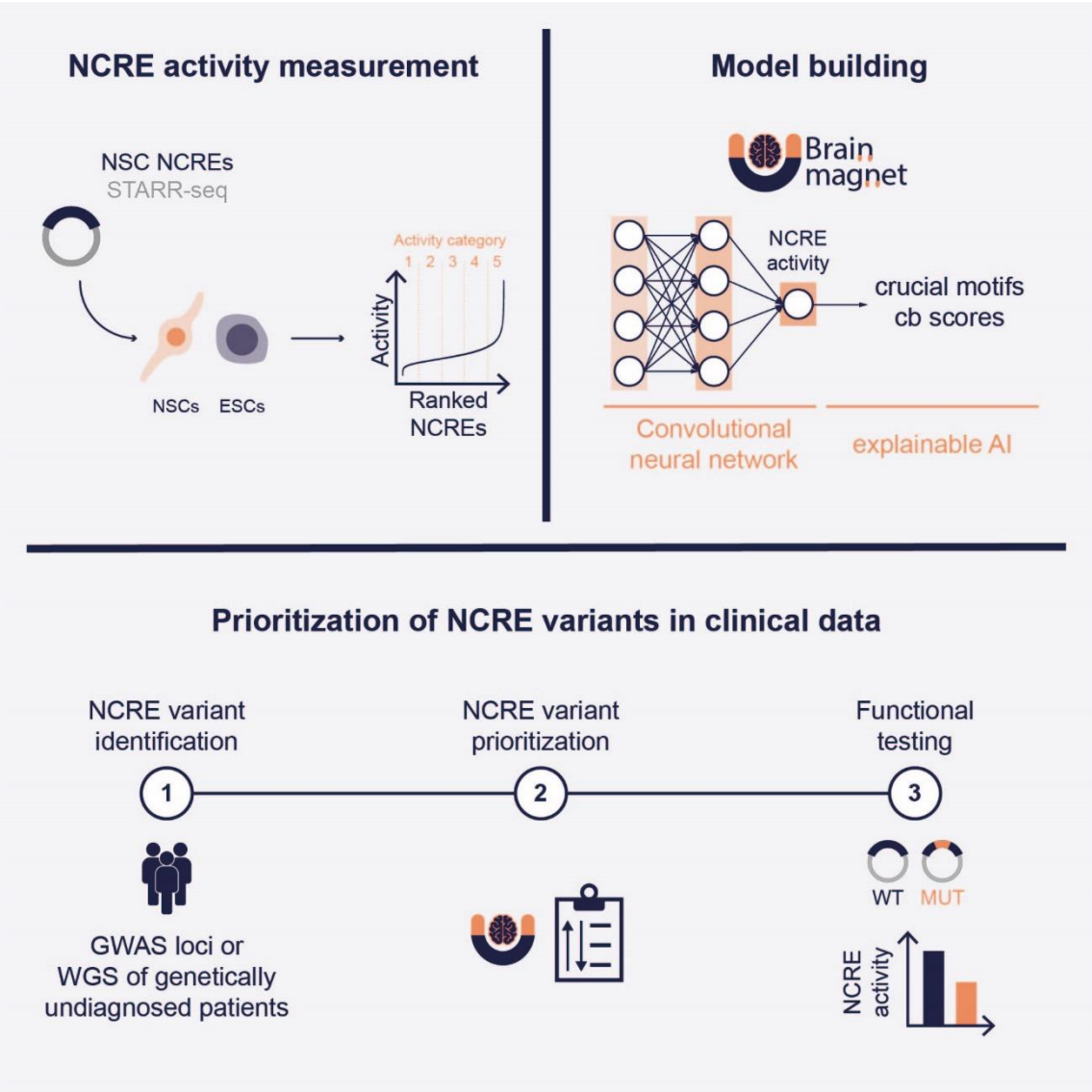


Figure 6: Graphical abstract summarizing BRAIN-MAGNET and its application to prioritize genetic variants in NCREs in rare and common disease

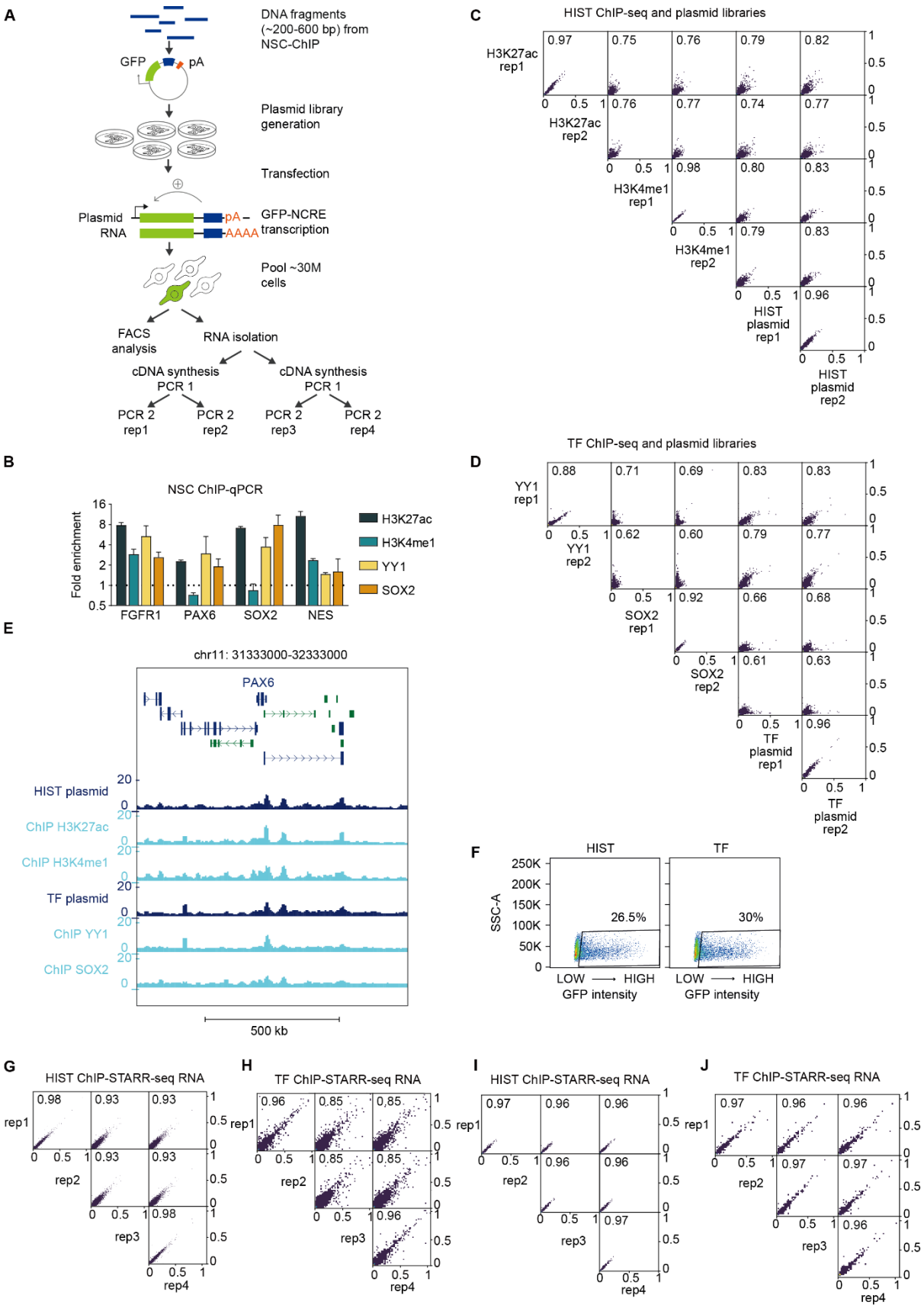


Figure S1: Experimental strategy and quality control for ChIP-STARR-seq.

A) Schematic representation of the experimental strategy used in this study.

B) ChIP-qPCR in NSC showing the average fold enrichment over total input, normalized to a non-bound site in *NCAPD2* (for H3K4me1 and H3K27ac histone ChIP) or in *ACTB* (for YY1 and SOX2 ChIP). Boxes represent the average of 2 replicates. Error bars represent standard deviations.

C) Scatterplot depicting the correlation between read counts (ln(CPM)) per genomic regions between the H3K27ac and H3K4me1 ChIP-seq and the HIST ChIP-STARR-seq plasmid libraries. Reported is the Pearson correlation coefficient.

D) As in C) but now for YY1 and SOX2 ChIP-seq and for the TF ChIP-STARR-seq plasmid libraries.

E) Genome browser tracks surrounding the *PAX6* locus showing good representation of H3K27ac, H3K4me1, YY1 and SOX2 ChIP-seq signals (light blue) in the HIST and TF ChIP-STARR-seq plasmid libraries (dark blue), and correlation between ChIP-STARR-seq RNA tracks (green)

F) FACS plot showing GFP expression of NSCs transfected with the HIST or TF ChIP-STARR-seq plasmid libraries.

G) Scatterplot depicting the correlation between read counts (ln(CPM)) per genomic regions between the HIST ChIP-STARR-seq RNA replicates from transfection of HIST plasmid libraries in NSCs (i.e., NSC ChIP-STARR-seq). Reported is the Pearson correlation coefficient.

H) As in G) but now for TF ChIP-STARR-seq RNA replicates from transfections of TF plasmid libraries in NSCs (i.e., NSC ChIP-STARR-seq).

I) As in G) but now for HIST ChIP-STARR-seq RNA replicates from transfections of HIST plasmid libraries in ESCs (i.e., comparative ChIP-STARR-seq in ESCs).

J) As in G) but now for TF ChIP-STARR-seq RNA replicates from transfections of TF plasmid libraries in ESCs (i.e., comparative ChIP-STARR-seq in ESCs).

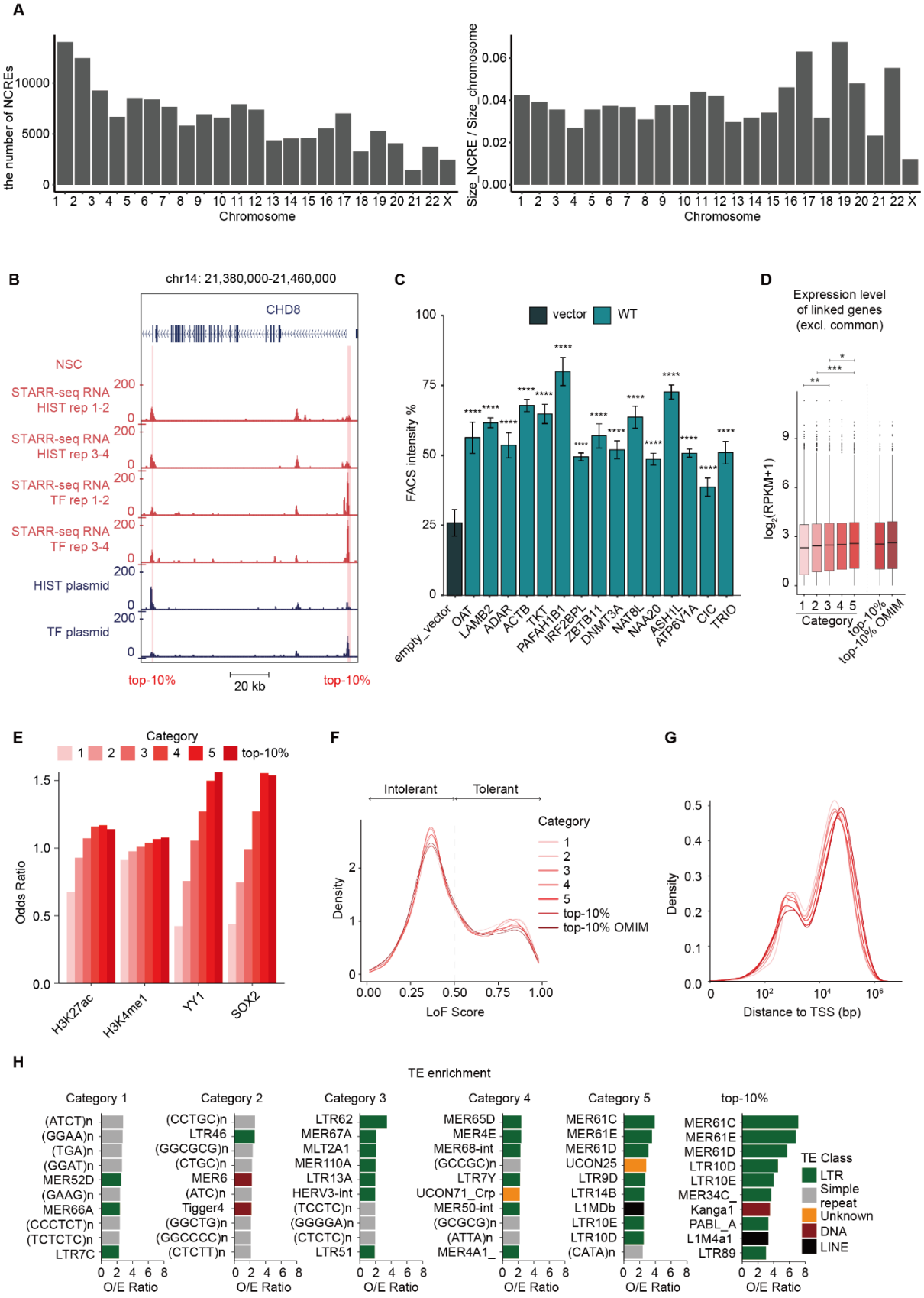


Figure S2: Extended data for identification of active NCREs in NSCs

A) Bar graph showing the number of NCRE scaffold regions per chromosome (left) and bar graph showing the ratio between total NCRE scaffold size divided by chromosome size (right). Both graphs exclude chromosome Y and chromosome M (mitochondrial DNA) which were not included in the analysis scaffold.

B) Genome browser tracks surrounding the *CHD8* locus showing two NCREs belonging to the top-10% category of activity.

C) Bar plot showing the percentage of GFP+ cells in NSCs upon cell transfection experiments with reporter plasmids containing either empty reporter plasmid control (black) or 15 different wild type NCREs (green). NCREs are indicated with the name of the presumed target gene. Plotted on the left is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

D) Expression level ($\log_2$(RPKM+1)) of the genes linked to each of the 5 NCRE activity categories, to the top-10% most active NCREs and of the OMIM disease-related genes linked to the top-10% most active NCREs, considering only the genes unique for each NCRE category. Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent outliers. (* $p < 0.05$; ** $p < 0.01$, Wilcoxon test).

E) Bar plot showing the ratio between the overlap of each NCRE category with each ChIP-seq and the total overlap of the 148,198 scaffold regions with each ChIP-seq, both normalized for the group size.

F) Kernel density plot showing the distribution of loss-of-function tolerance scores for non-coding sequences for all the NCRE categories from NSC ChIP-STARR-seq.

G) Density plot showing the distance in bps to the TSS for all NCRE categories.

H) Bar graph showing the 10 most enriched transposable elements (TEs) overlapping with the NCREs belonging to the different activity categories in NSCs. Plotted is a ratio between the observed (O) number of TEs over the expected (E).

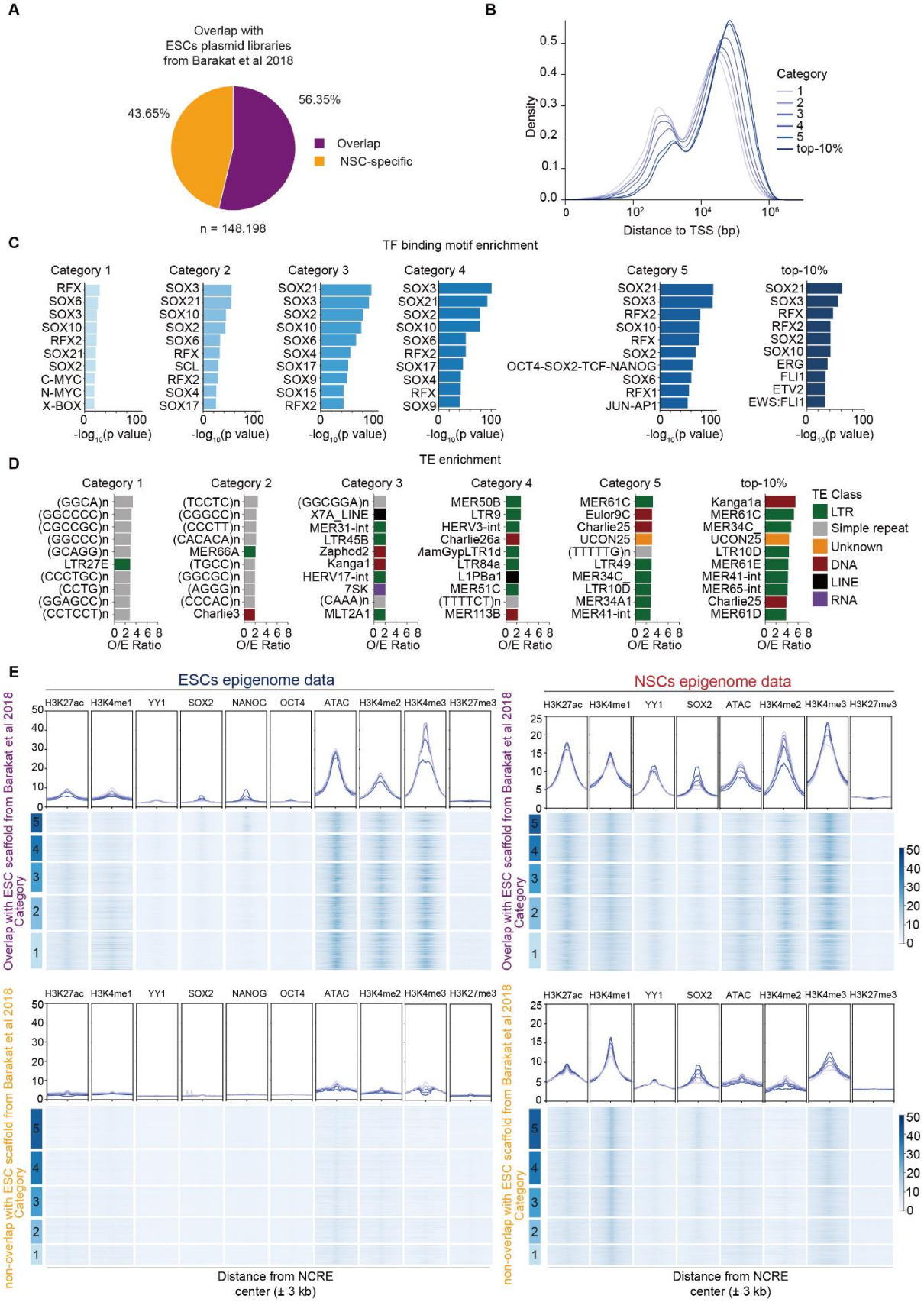


Figure S3: Extended data for comparative ChIP-STARR-seq using NSC derived plasmid libraries in ESCs

A) Pie chart showing the fraction of NSC ChIP-STARR-seq scaffold regions overlapping with our previously studied ESC scaffold regions from ESC ChIP-STARR-seq (Barakat *et al.*, 2018).

B) Density plot showing the distance in bps to the TSS for all NCRE categories.

C) Bar graph showing the 10 most enriched TF motifs in the NCREs belonging to the different activity categories from comparative ChIP-STARR-seq in ESCs. Plotted is the -log10 p-value, as calculated by Homer.

D) Bar graph showing the 10 most enriched transposable elements (TEs) overlapping with the NCREs belonging to the different activity categories from comparative ChIP-STARR-seq in ESCs. Plotted is a ratio between the observed (O) number of TEs over the expected (E).

E) Epigenome profiling of NCREs with comparative ChIP-STARR-seq activity in ESCs (plasmid libraries derived from ChIP experiments in NSCs). Results are shown for the 148,198 scaffold regions that overlap (top row) or do not overlap (bottom row) with the scaffold regions used earlier in ESC ChIP-STARR-seq (Barakat *et al.*, 2018). For each row at the Top: line plot showing the average enrichment of ChIP-seq (or ATAC-seq) signal from ESCs (left panel) or NSCs (right panel) across the genomic regions encompassed by NCREs from the 5 categories of NCRE activity detected by comparative ChIP-STARR-seq in ESCs (bins of 3 kb up and down centred at the middle of the NCRE). For each row at the Bottom: corresponding heatmaps of H3K27ac, H3K4me1, YY1, SOX2, NANOG, OCT4, H3K4me2, H3K4me3 and H3K27me3 ChIP-seq signals and ATAC-seq in ESCs, or H3K27ac, H3K4me1, YY1, SOX2, H3K4me2, H3K4me3 and H3K27me3 ChIP-seq signals and ATAC-seq in NSCs across the assessed scaffold regions ranked by comparative ChIP-STARR-seq activity in ESCs from the highest to the lowest.

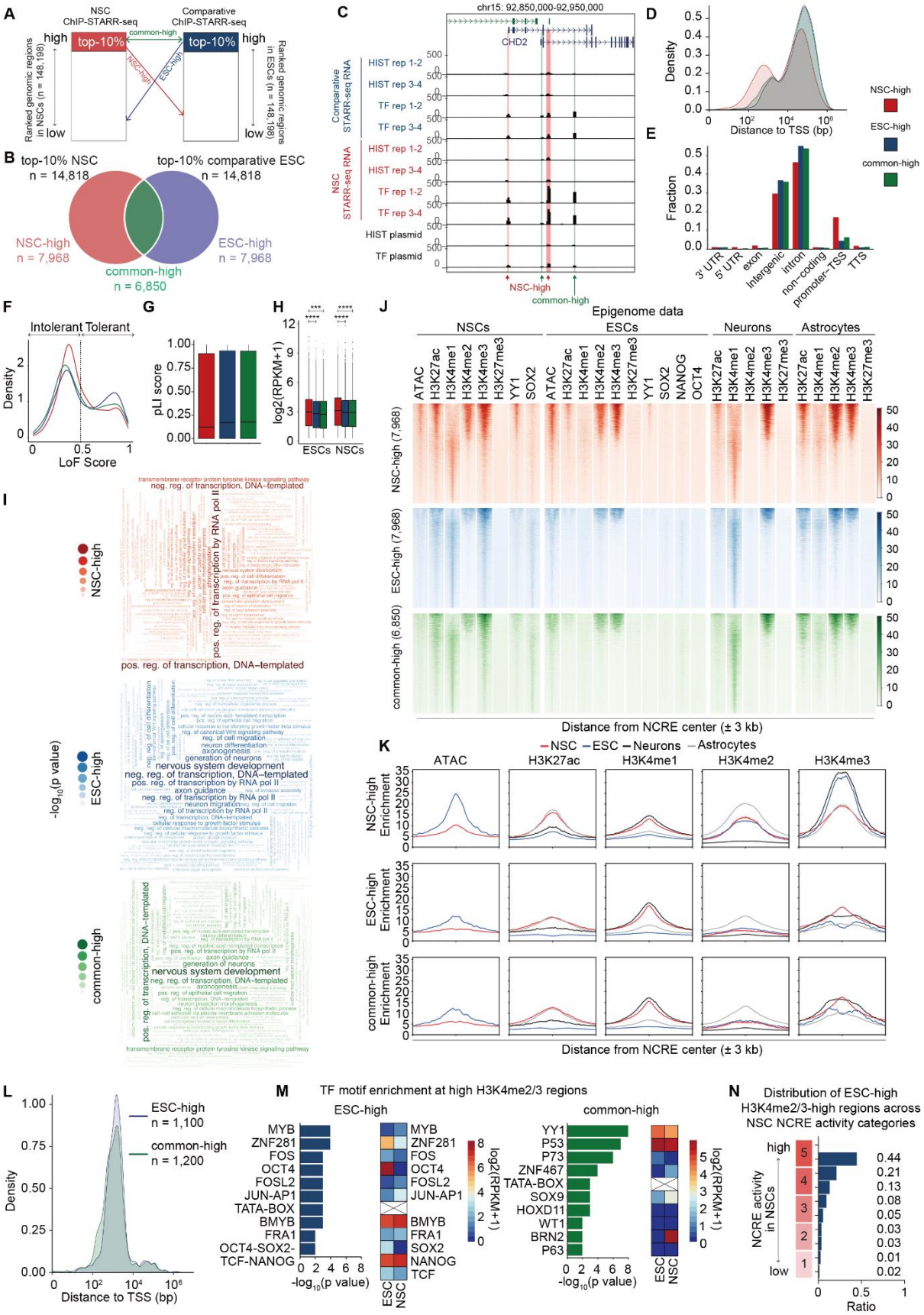


Figure S4 Comparative ChIP-STARR-seq identifies differentially active NCREs in NSCs and ESCs.

A) Schematic overview of the process to define common-high (green arrow), NSC-high (red arrow) and ESC-high (blue arrow) NCREs.

B) Venn diagram showing the overlap between the top-10% most active NCREs in NSCs and ESCs.

C) Genome browser tracks surrounding the *CHD2* locus showing ChIP-STARR-seq RNA-seq and DNA-seq (plasmid) tracks.

D) Density plot showing the distance to the TSS of differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high, green).

E) Bar graphs showing the fraction of differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high, green) per genomic region using Homer. UTR = untranslated region; non-coding = non-coding RNAs such as snRNA, lncRNA, etc; TSS = Transcriptional start site; TTS = Transcriptional termination site.

F) Kernel density plot showing the distribution of loss-of-function tolerance scores for non-coding sequences for differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high, green).

G) pLI score of the genes linked to differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high, green).

H) Expression level ($\log_2$(RPKM+1)) in ESCs and NSCs of the genes linked to differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high). Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent outliers. (* $p < 0.05$; Wilcoxon test).

I) Word cloud representation of gene ontology analysis using Enrichr for the protein-coding genes linked to differentially active NCREs (NSC-high (red) and ESC-high (blue)) and common highly active NCREs in ESCs and NSCs (common-high).

J) Heat maps of ATAC-seq and ChIP-seq signals in NSCs, ESCs, neurons and astrocytes across the 7,968 NSC-high, 7,968 ESC-high and the 6,850 common-high active NCREs

K) Line plots showing the average enrichment of ATAC-seq and ChIP-seq signal for H3K27ac, H3K4me1, H3K4me2 and H3K4me3 in NSCs (red), ESCs (blue), neurons (black) and astrocytes (grey)

L) Density plot showing the distance to the TSS of the 1,100 ESC-high NCREs with high H3K4me2/3

M) Bar graphs showing the 10 most enriched TF motifs in the 1,100 ESC-high NCREs and the 1,200 common highly active NCREs with high H3K4me2/3. Plotted is the -log10 p-value. Flanking the bar graphs is a heatmap showing the expression (log_2(RPKM+1)) of each TF in ESCs and NSCs.

N) Bar graph showing the distribution of the 1,100 ESC-high NCREs with high H3K4me2/3 across the different categories of NCRE activity in NSCs. Indicated is the fraction of regions per category.

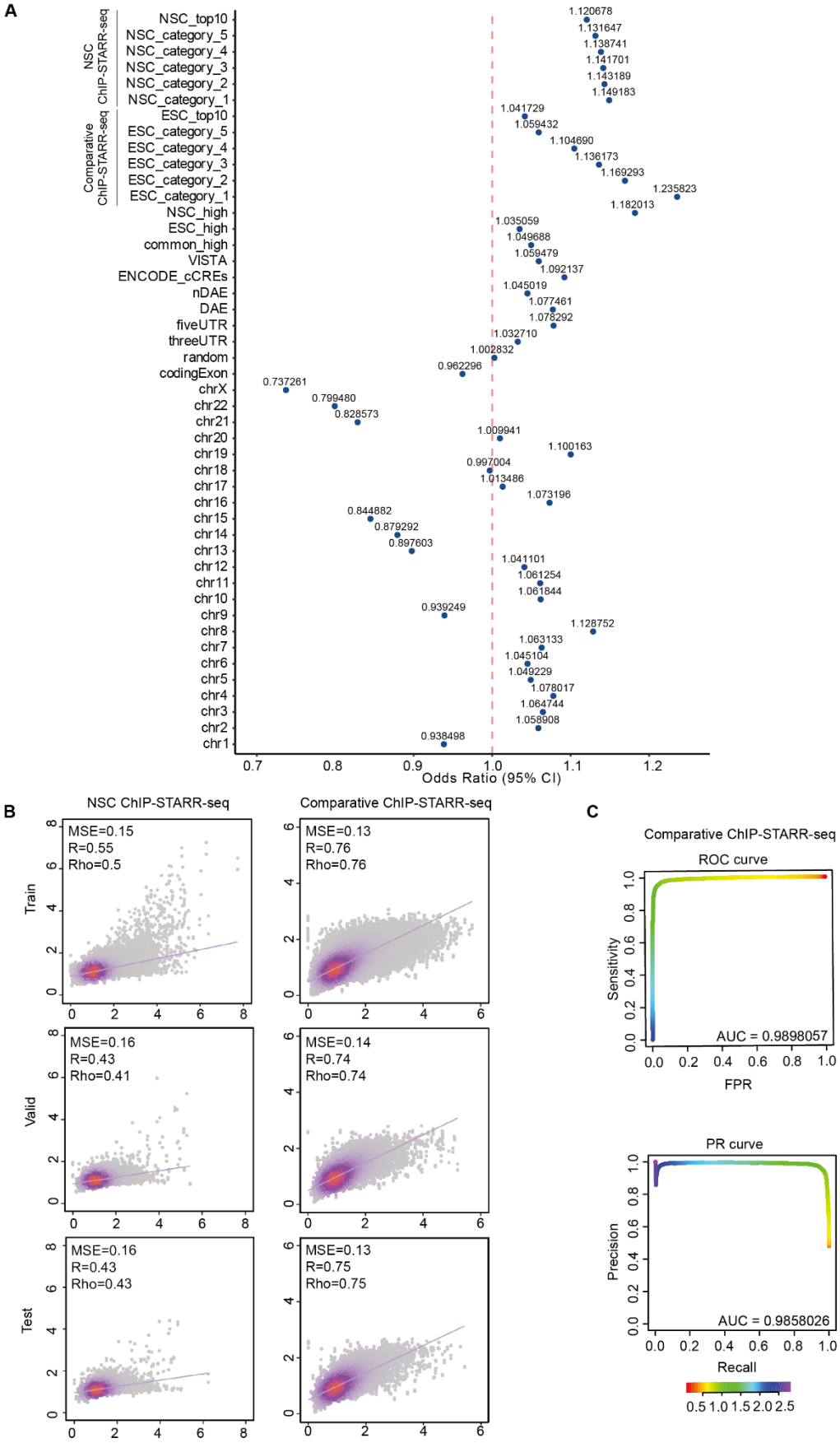


Figure S5: Extended data on rare variants in NCREs and BRAIN-MAGNET

A) Plot showing the odds ratio of observed and expected rare variants (minor allele frequency <0.1%) from 76,215 genomes from individuals that did not present with early onset neurodevelopmental phenotypes from gnomAD v4 (Chen *et al.*, 2024) that overlap with the various categories of genomic sequences: NSC category 1-5 and NSC top-10%, NCRE activity categories in NSCs; ESC category 1-5 and ESC top-10%, NCRE activity categories in ESCs; NSC_high, ESC_high and common_high, NCRE categories from the differential activity analysis (see **Figure S4**); VISTA (Visel *et al.*, 2007), ENCODE candidate cis regulatory elements (Consortium *et al.*, 2020), nDAEs and DAEs (Yousefi *et al.*, 2021), 5' and 3' UTR regions and coding exome obtained from the UCSC Genome browser; 148,114 random regions. For reference, the odds ratio was also plotted for the various individual chromosomes (amongst others, the acrocentric chromosomes and chromosome X are depleted for rare variants).

B) Scatter plot indicating relationship between BRAIN-MAGNET predicted and observed NCRE activity scores for each cell type. Also indicated are the mean square error (MSE), Pearson (r) and Spearman (rho) correlation values after concatenating the observations and predictions for the train, valid or test set of data.

C) Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves of comparing Top10 and Bottom10 NCREs from the test set using the R package PRROC (Grau *et al.*, 2015) for ESCs. The PR curve reflects the precision-recall trade-off, and the ROC curve portrays the true positive and false positive rate trade-off. The colour scale maps classification thresholds to points on the curve, providing visual guidance for selecting an optimal threshold. (Recall = TP / (TP + FN), Precision = TP / (TP + FP), True positive rate = TP / (TP + FN), False positive rate = FP / (FP + TN))

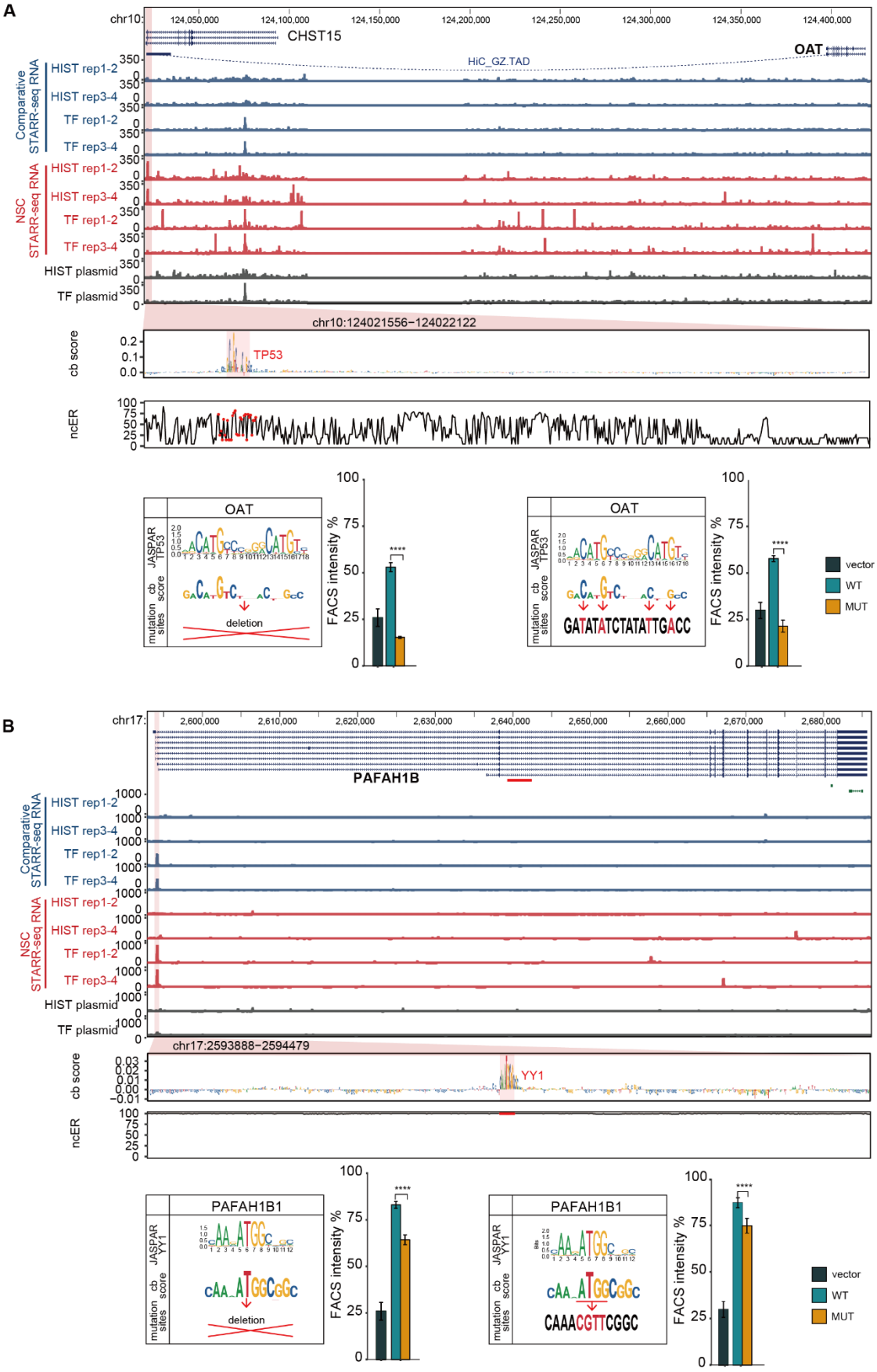


Figure S6: Example loci of *OAT* and *PAFAH1B* for which BRAIN-MAGNET prioritized NCREs and motifs, including functional validation.

A) Genome browser view of the *OAT* and *CHST15* locus, showing an NCRE in an intron of *CHST15* that interacts by HiC in foetal brain with *OAT*. Top panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted. Lower panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion (left) or point mutation (right) that was introduced in the NCRE. Bar plots show the results from experimental testing of the created deletion (left) or point mutation (right). Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

B) As A, but now for an NCRE in *PAFAH1B* for which BRAIN-MAGNET predicts a functional role for a YY1 motif.

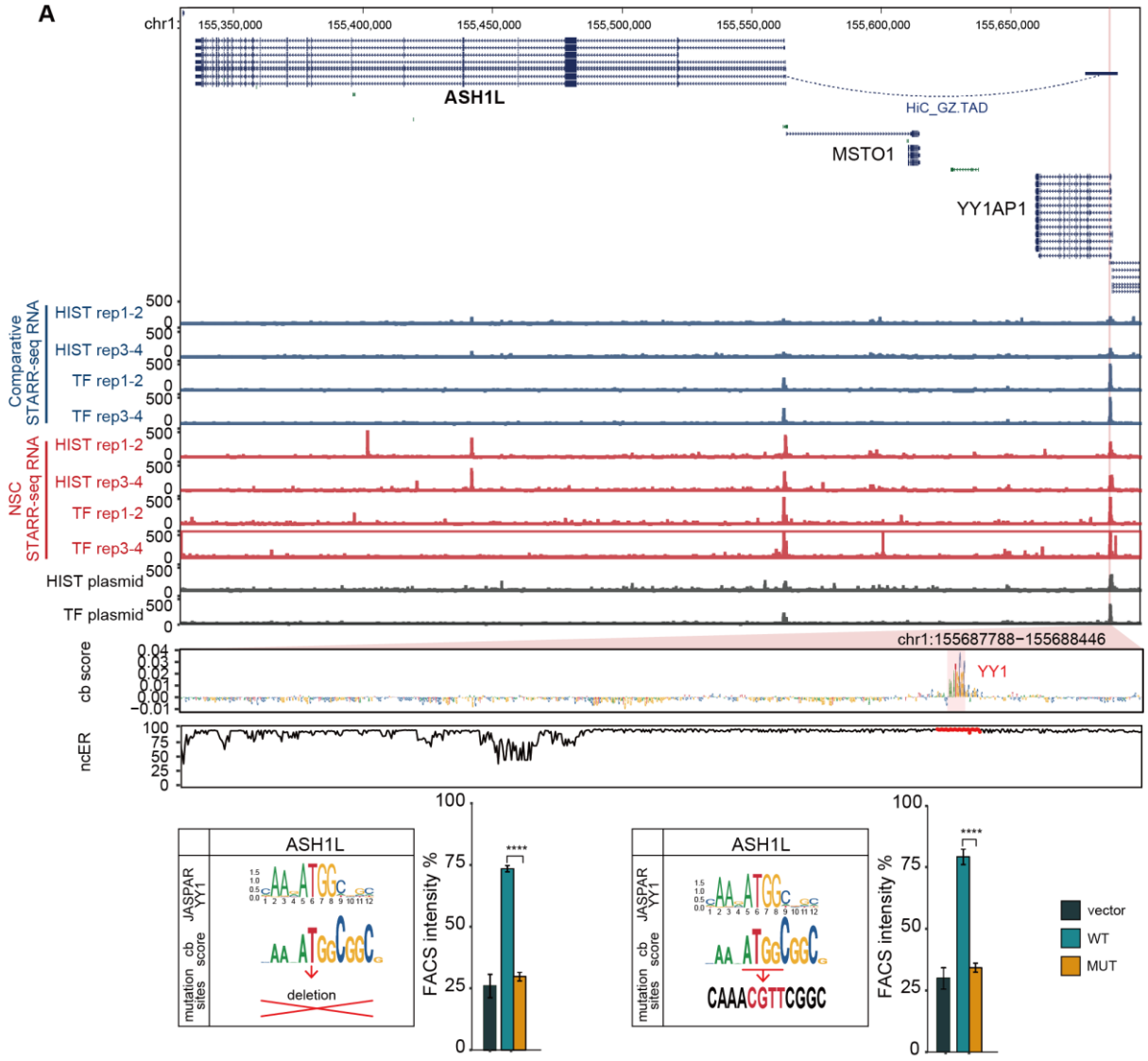


Figure S7: Example locus of *ASHL1*, for which BRAIN-MAGNET prioritized NCRE and motif, including functional validation.

A) Genome browser view of the *ASHL1* locus and flanking genes, showing an NCRE in *YY1AP1* that interacts by HiC in foetal brain with *ASHL1*. Top panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping YY1 motif is highlighted. Lower panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion (left) or point mutation (right) that was introduced in the NCRE. Bar plots show the results from experimental testing of the created deletion (left) or point mutation (right). Plotted is the percentage of GFP+ in cells co-transfected with an mCherry

expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

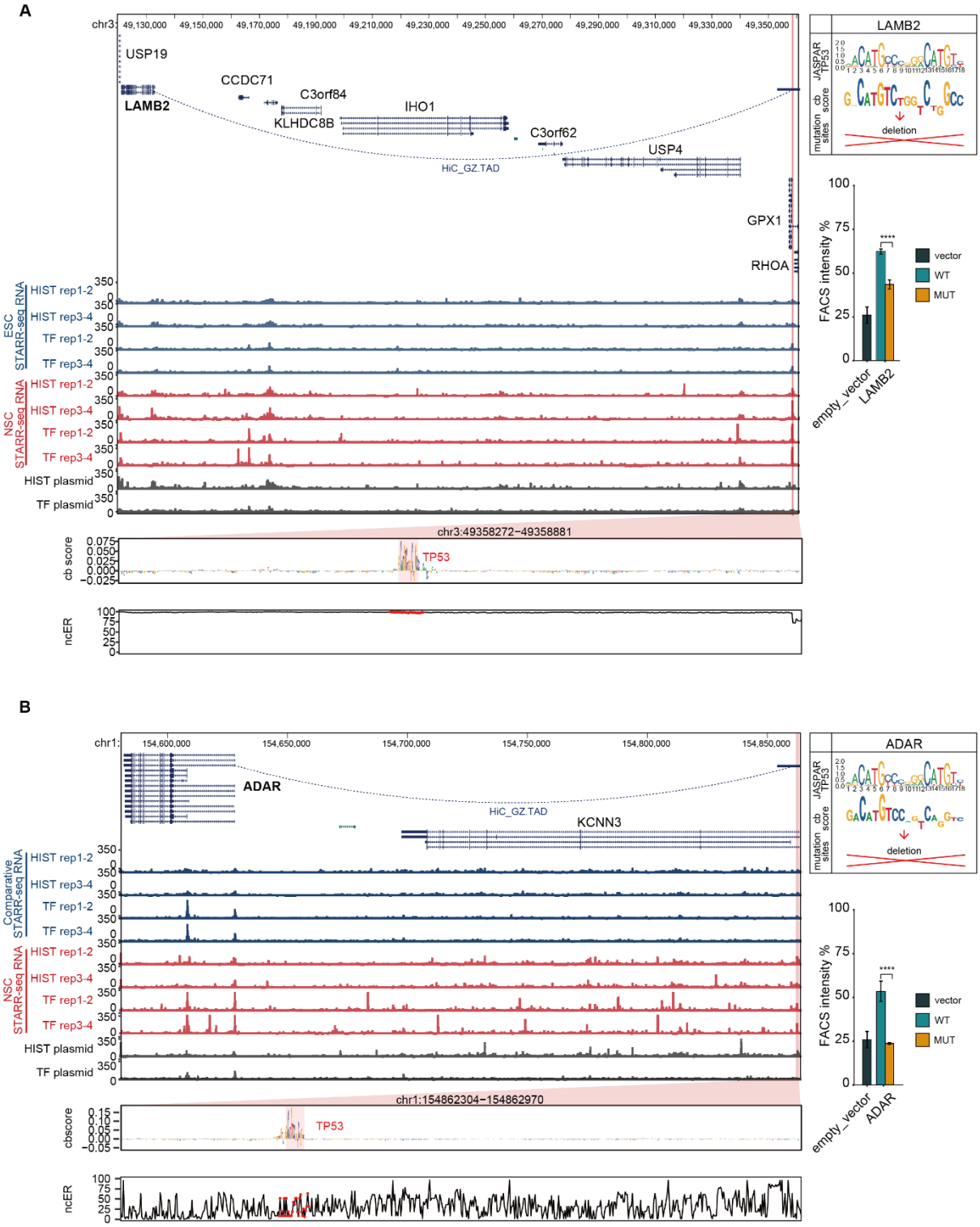


Figure S8: Example loci of *LAMB2* and *ADAR* for which BRAIN-MAGNET prioritized NCREs and motifs, including functional validation.

A) Genome browser view of the *LAMB2* locus and flanking genes, showing an NCRE located between *RHOA* and *GPX1* that interacts by HiC in foetal brain with *LAMB2*. Left panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and

sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted. Upper right panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion that was introduced in the NCRE. Lower bar plots show the results from experimental testing of the created deletion. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

B) As A), but now for an NCRE located in *KCNN3* that interacts by HiC in foetal brain with *ADAR*.

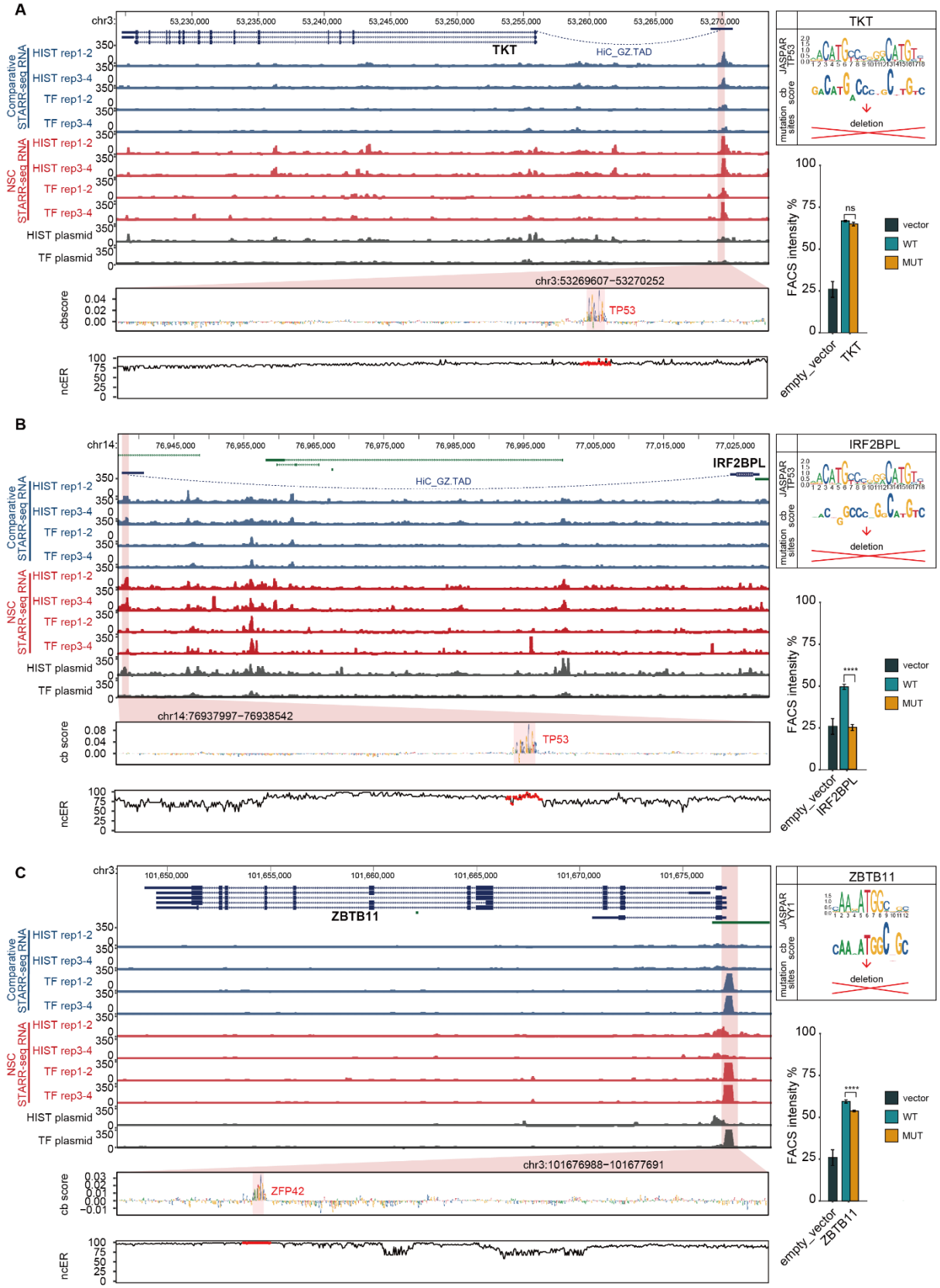


Figure S9: Example loci of *TKT, IRF2BPL* and *ZBTB11* for which BRAIN-MAGNET prioritized NCREs and motifs, including functional validation.

A) Genome browser view of the *TKT* locus, showing a distal NCRE that interacts by HiC in foetal brain with *TKT*. Left panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted. Upper right panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion that was introduced in the NCRE. Lower bar plots show the results from experimental testing of the created deletion. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

B) As A), but now for a distal NCRE that interacts by foetal brain HiC with *IRF2BPL*.

C) As A), but now for an NCRE upstream of *ZBTB11* that contains a ZFP42/YY1 motif.

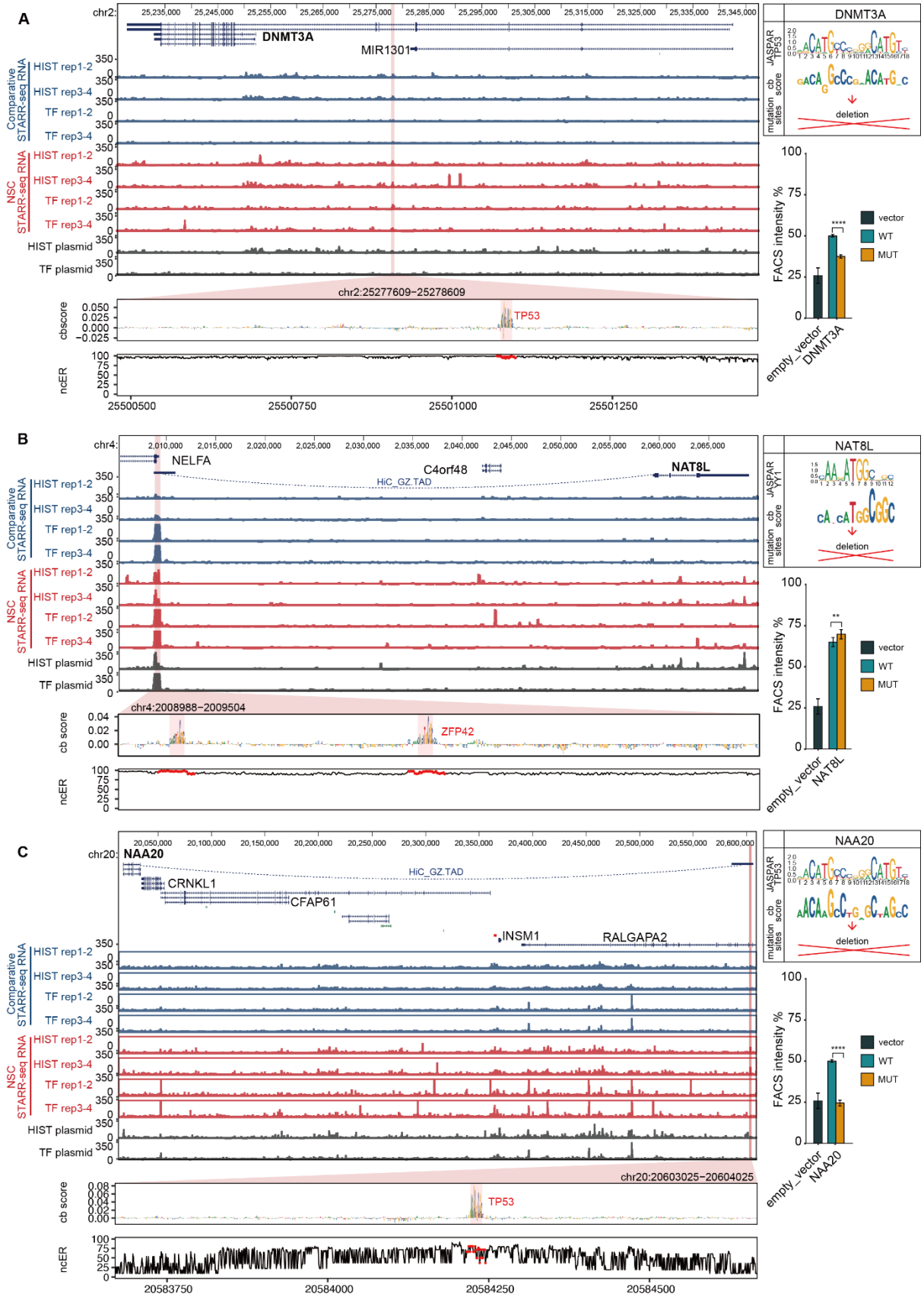


Figure S10: Example loci of *DNMT3A, NAT8L* and *NAA20* for which BRAIN-MAGNET prioritized NCREs and motifs, including functional validation.

A) Genome browser view of the *DNMT3A* locus, showing an intronic NCRE. Left panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted. Upper right panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion that was introduced in the NCRE. Lower bar plots show the results from experimental testing of the created deletion. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

B) as for A), but now for a distal NCRE located near *NELFA* harboring a ZFP42/YY1 motif, that interacts with *NAT8L* as confirmed by HiC in foetal brain.

C) as for A), but now a distal NCRE that interacts by foetal brain HiC with *NAA20*.

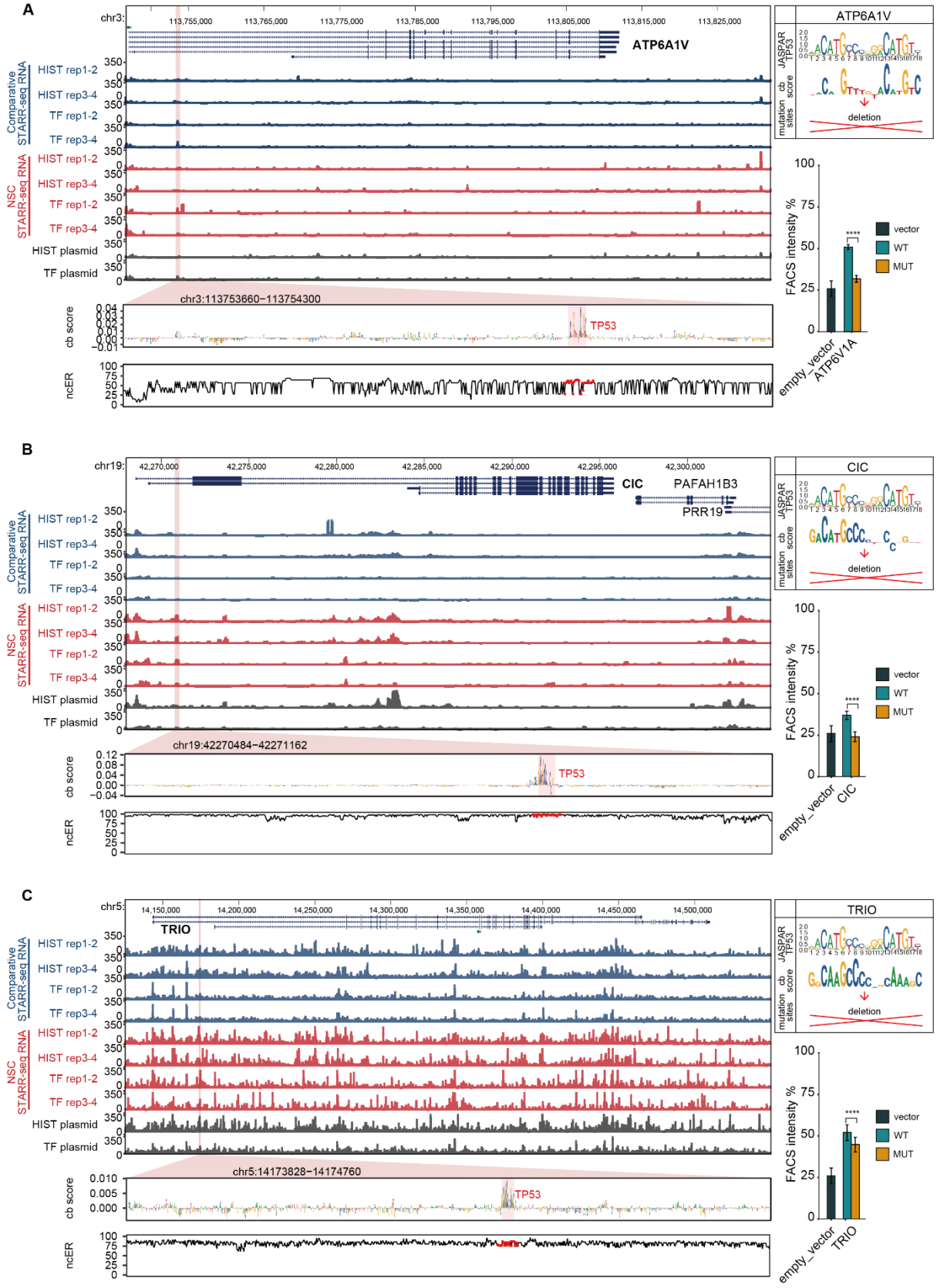


Figure S11: Example loci of *ATP6A1V, CIC* and *TRIO* for which BRAIN-MAGNET prioritized NCREs and motifs, including functional validation.

A) Genome browser view of the *ATP6A1V* locus, showing an intronic NCRE. Left panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping TP53 motif is highlighted. Upper right panel shows the motif from JASPAR, a visualization of the cb score and the generated deletion that was introduced in the NCRE. Lower bar plots show the results from experimental testing of the created deletion. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

B) as for A), but now for an intronic NCRE in *CIC*.

C) as for A), but now for an intronic NCRE in *TRIO*.

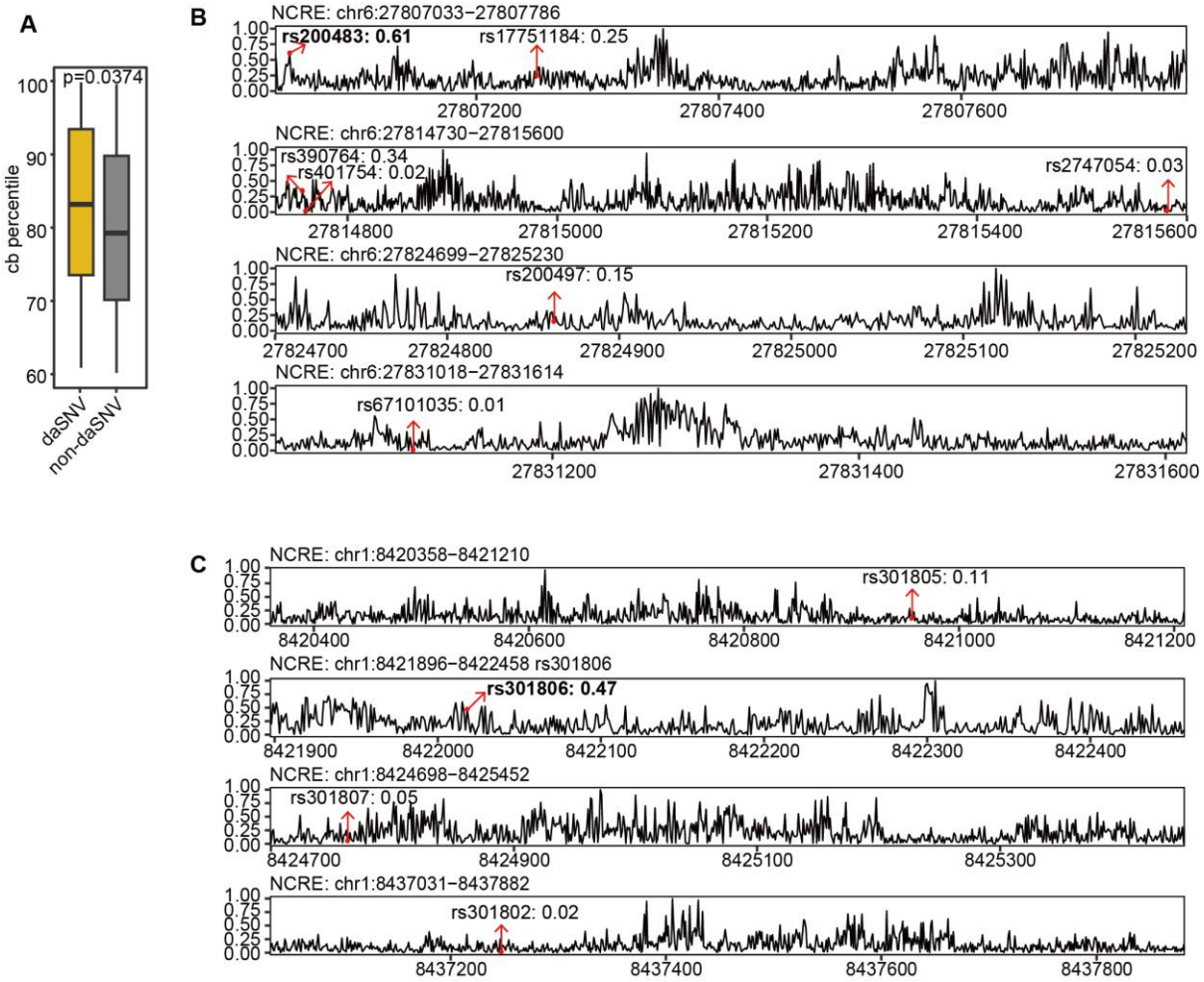


Figure S12: Utility of the NCRE atlas and BRAIN-MAGNET for prioritizing genomic variants to fine map GWAS associated regions

A) Box plot showing the distribution of BRAIN-MAGNET cb scores (normalized cb_all) for daSNVs (yellow) and non-daSNVs (gray) from Guo et al (Guo *et al.*, 2023) that overlap with NCREs assessed in our study. Boxes represent the interquartile range (IQR); lines represent the median; whiskers extend to 1.5 the IQR; dots represent outliers. (* $p < 0.05$; t.test).

B) Zoom-in showing the normalized BRAIN-MAGNET cb scores (cb_all) for the 4 NCREs from our data set that overlap with the GWAS associated risk locus for schizophrenia on chromosome 6. The 7 SNPs in LD and their cb scores are indicated. Of note is that within each NCRE, there are nucleotides which are not common SNPs not assessed in GWAS studies that have higher individual cb scores and might thus impact more on NCRE function if mutant.

C) As B), but now for the 4 NCREs that overlap with a GWAS associated locus for major depression on chromosome 1.

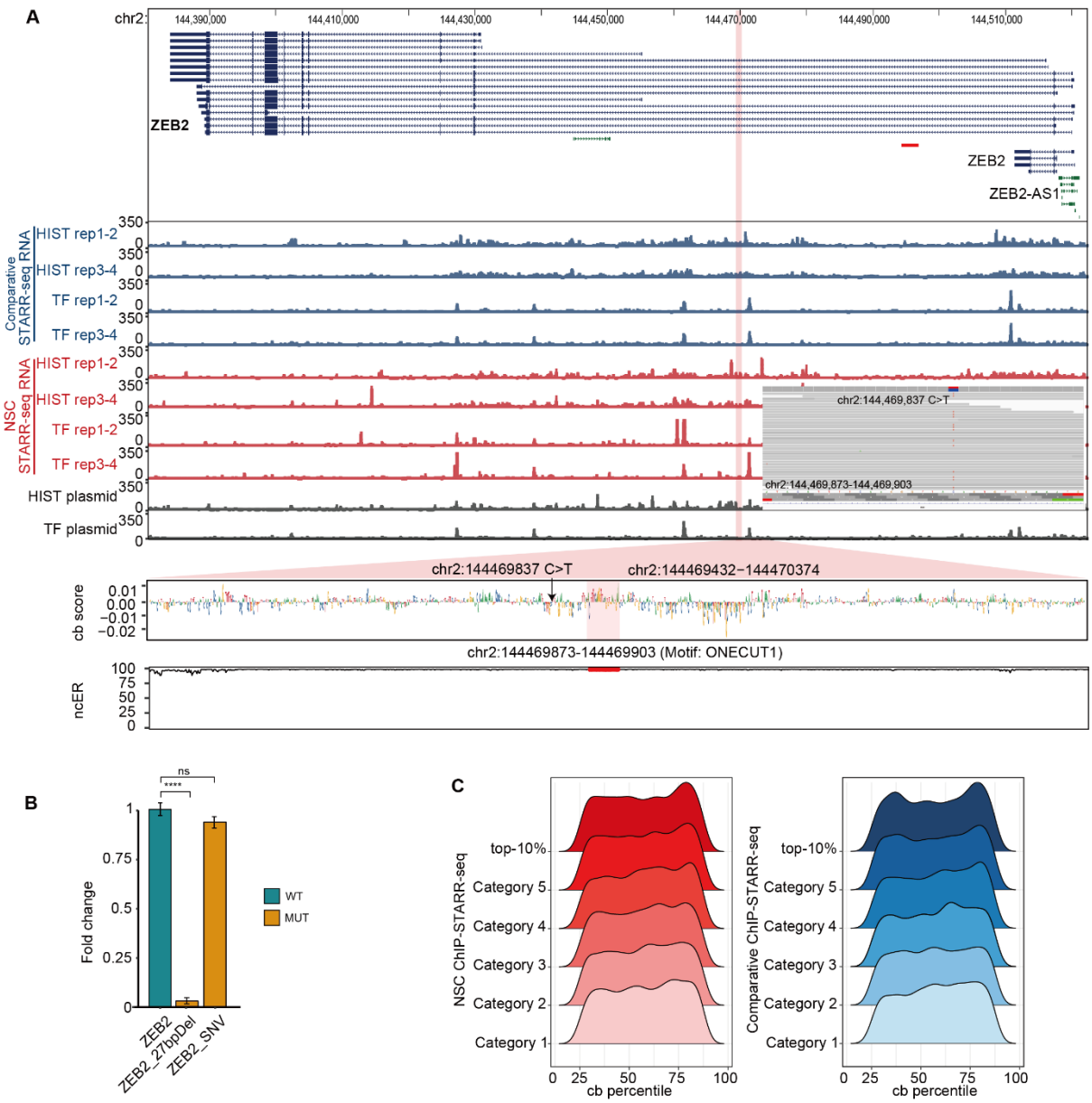


Figure S13: Extended data showing the utility of the NCRE atlas and BRAIN-MAGNET for prioritizing genomic variants found in rare disease

A) Genome browser view of the *ZEB2* locus, showing an intronic NCRE in which a rare variant was found in a genetically unexplained individual clinically suspected of Mowat-Wilson syndrome. Top panel: genome browser tracks showing ChIP-STARR-seq RNA in ESCs (blue) and NSCs (red) and sequencing of ChIP-STARR-seq plasmids (grey); highlighted region in magenta indicates the NCRE. The zoom-in of the NCRE shows the BRAIN-MAGNET contribution (cb) scores from the NSC model and the ncER scores of the same region. A region with high cb scores and an overlapping ONECUT1 motif and the mutation site of the variant found in the affected individual are highlighted. The IGV browser view show the heterozygous variant identified in the individual.

B) Left panel shows the motif from JASPAR, a visualization of the cb score and the generated ONECUT1 deletion that was introduced in the NCRE and the NCRE single nucleotide variant identified in the affected individual. Right bar plot shows the results from experimental testing of the created deletion of the ONECUT1 motif or the NCRE variant identified in the affected individual. Plotted is the percentage of GFP+ in cells co-transfected with an mCherry expressing plasmid, to correct for transfection efficiency. Bars show the average from two independent experiments, with each NCRE tested each in duplicate. Error bars represent standard deviation. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$ (one-way ANOVA test followed by multiple comparison test (Fisher's LSD test).

C) Density plots depicting the number of *de novo* variants found in 4,558 genetically unexplained individuals with other rare disorders (including cardiovascular disorders, hearing an ear disorders, ultra-rare disorders, dermatological disorders, gastroenterological disorders, renal and urinary tract disorders, respiratory disorders, endocrine disorders, dysmorphic and congenital abnormality syndromes, growth disorders, skeletal disorders, ophthalmological disorders, metabolic disorders, hematological and immunological disorders and rheumatological disorders) from the Genomics England 100,000 Genomes project, located in NCREs from the different NCRE categories for NSCs (left) and ESCs (right), stratified for their BRAIN-MAGNET cb percentile.

Reference

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169. 10.1093/bioinformatics/btu638.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074-1077. 10.1126/science.1232542.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208. 10.1093/nar/gkp335.

Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. Cell Stem Cell *23*, 276-288 e278. 10.1016/j.stem.2018.06.014.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120. 10.1093/bioinformatics/btu170.

Borsari, B., Villegas-Miron, P., Perez-Lluch, S., Turpin, I., Laayouni, H., Segarra-Casas, A., Bertranpetit, J., Guigo, R., and Acosta, S. (2021). Enhancers with tissue-specific activity are enriched in intronic regions. Genome Res *31*, 1325-1336. 10.1101/gr.270371.120.

Bravo Gonzalez-Blas, C., Matetovici, I., Hillen, H., Taskiran, II, Vandepoel, R., Christiaens, V., Sansores-Garcia, L., Verboven, E., Hulselmans, G., Poovathingal, S., et al. (2024). Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation. Nat Cell Biol *26*, 153-167. 10.1038/s41556-023-01316-4.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213-1218. 10.1038/nmeth.2688.

Carullo, N.V.N., and Day, J.J. (2019). Genomic Enhancers in Brain Health and Disease. Genes (Basel) *10*. 10.3390/genes10010043.

Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., Hubbard, T., Jostins, L., Maltby, N., Mahon-Pearson, J., et al. (2017). National Genomic Research Library(figshare). doi.org/10.6084/m9.figshare.4530893.v7.

Chan, K.M., Fang, D., Gan, H., Hashizume, R., Yu, C., Schroeder, M., Gupta, N., Mueller, S., James, C.D., Jenkins, R., et al. (2013). The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes Dev *27*, 985-990. 10.1101/gad.217778.113.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128. 10.1186/1471-2105-14-128.

Chen, S., Francioli, L.C., Goodrich, J.K., Collins, R.L., Kanai, M., Wang, Q., Alfoldi, J., Watts, N.A., Vittal, C., Gauthier, L.D., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. Nature *625*, 92-100. 10.1038/s41586-023-06045-0.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet *18*, 71-86. 10.1038/nrg.2016.139.

Clark, M.M., Stark, Z., Farnaes, L., Tan, T.Y., White, S.M., Dimmock, D., and Kingsmore, S.F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genom Med *3*, 16. 10.1038/s41525-018-0053-8.

Colbran, L.L., Chen, L., and Capra, J.A. (2017). Short DNA sequence patterns accurately identify broadly active human enhancers. BMC Genomics *18*, 536. 10.1186/s12864-017-3934-9.

Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699-710. 10.1038/s41586-020-2493-4.

de Almeida, B.P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet *54*, 613-624. 10.1038/s41588-022-01048-5.

Dong, S., Zhao, N., Spragins, E., Kagda, M.S., Li, M., Assis, P., Jolanki, O., Luo, Y., Cherry, J.M., Boyle, A.P., and Hitz, B.C. (2023). Annotating and prioritizing human non-coding variants with RegulomeDB v.2. Nat Genet *55*, 724-726. 10.1038/s41588-023-01365-3.

Fellows, I., Fellows, M.I., Rcpp, L., and Rcpp, L. (2018). Package ‘wordcloud’. R package version *2*, 331.

Freire-Pritchett, P., Schoenfelder, S., Varnai, C., Wingett, S.W., Cairns, J., Collier, A.J., Garcia-Vilchez, R., Furlan-Magaril, M., Osborne, C.S., Fraser, P., et al. (2017). Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. Elife *6*. 10.7554/eLife.21926.
Glaser, L.V., Steiger, M., Fuchs, A., van Bommel, A., Einfeldt, E., Chung, H.R., Vingron, M., and Meijsing, S.H. (2021). Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. Nucleic Acids Res *49*, 12178-12195. 10.1093/nar/gkab1100.
Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics *31*, 2595-2597. 10.1093/bioinformatics/btv153.
Guo, M.G., Reynolds, D.L., Ang, C.E., Liu, Y., Zhao, Y., Donohue, L.K.H., Siprashvili, Z., Yang, X., Yoo, Y., Mondal, S., et al. (2023). Integrative analyses highlight functional regulatory variants associated with neuropsychiatric diseases. Nat Genet *55*, 1876-1891. 10.1038/s41588-023-01533-5.
Gussow, A.B., Copeland, B.R., Dhindsa, R.S., Wang, Q., Petrovski, S., Majoros, W.H., Allen, A.S., and Goldstein, D.B. (2017). Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. PLoS One *12*, e0181604. 10.1371/journal.pone.0181604.
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell *38*, 576-589. 10.1016/j.molcel.2010.05.004.
Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol *16*, 144-154. 10.1038/nrm3949.
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. 10.1038/ng.3810.
Kaller, M., Gotz, U., and Hermeking, H. (2017). Loss of p53-inducible long non-coding RNA LINC01021 increases chemosensitivity. Oncotarget *8*, 102783-102800. 10.18632/oncotarget.22245.
Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Methods *12*, 357-360. 10.1038/nmeth.3317.
Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915. 10.1038/s41587-019-0201-4.
Kim, J.D., Faulk, C., and Kim, J. (2007). Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. Nucleic Acids Res *35*, 3442-3452. 10.1093/nar/gkm235.
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. 10.1038/ng.2892.
Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res *44*, W90-97. 10.1093/nar/gkw377.
Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet *42*, 631-634. 10.1038/ng.600.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359. 10.1038/nmeth.1923.
Lecellier, C.H., Wasserman, W.W., and Mathelier, A. (2018). Human Enhancers Harboring Specific Sequence Composition, Activity, and Genome Organization Are Linked to the Immune Response. Genetics *209*, 1055-1071. 10.1534/genetics.118.301116.
LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436-444. 10.1038/nature14539.
Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291. 10.1038/nature19057.
Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930. 10.1093/bioinformatics/btt656.
Lim, F., Solvason, J.J., Ryan, G.E., Le, S.H., Jindal, G.A., Steffen, P., Jandu, S.K., and Farley, E.K. (2024). Affinity-optimizing enhancer variants disrupt development. Nature *626*, 151-159. 10.1038/s41586-023-06922-8.
Lundberg, S., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874. 10.48550/arXiv.1705.07874.

Marin Navarro, A., Pronk, R.J., van der Geest, A.T., Oliynyk, G., Nordgren, A., Arsenian-Henriksson, M., Falk, A., and Wilhelm, M. (2020). p53 controls genomic stability and temporal differentiation of human neural stem cells and affects neural organization in human brain organoids. Cell Death Dis *11*, 52. 10.1038/s41419-019-2208-7.
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. 10.1126/science.1222794.
Miyagi, S., Kato, H., and Okuda, A. (2009). Role of SoxB1 transcription factors in development. Cell Mol Life Sci *66*, 3675-3684. 10.1007/s00018-009-0097-0.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems *32*.
Perenthaler, E., Nikoncuk, A., Yousefi, S., Berdowski, W.M., Alsagob, M., Capo, I., van der Linde, H.C., van den Berg, P., Jacobs, E.H., Putar, D., et al. (2020). Loss of UGP2 in brain leads to a severe epileptic encephalopathy, emphasizing that bi-allelic isoform-specific start-loss mutations of essential genes can cause genetic diseases. Acta Neuropathol *139*, 415-442. 10.1007/s00401-019-02109-6.
Perenthaler, E., Yousefi, S., Niggl, E., and Barakat, T.S. (2019). Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. Front Cell Neurosci *13*, 352. 10.3389/fncel.2019.00352.
Pieraccioli, M., Nicolai, S., Pitolli, C., Agostini, M., Antonov, A., Malewicz, M., Knight, R.A., Raschella, G., and Melino, G. (2018). ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. Proc Natl Acad Sci U S A *115*, 7356-7361. 10.1073/pnas.1801435115.
Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics *47*, 11 12 11-34. 10.1002/0471250953.bi1112s47.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842. 10.1093/bioinformatics/btq033.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283. 10.1038/nature09692.
Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165. 10.1093/nar/gkw257.
Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277. 10.1016/s0168-9525(00)02024-2.
Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140. 10.1093/bioinformatics/btp616.
Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411. 10.1038/nature01080.
Saveri, P., De Luca, M., Nisi, V., Pisciotta, C., Romano, R., Piscosquito, G., Reilly, M.M., Polke, J.M., Cavallaro, T., Fabrizi, G.M., et al. (2020). Charcot-Marie-Tooth Type 2B: A New Phenotype Associated with a Novel RAB7A Mutation and Inhibited EGFR Degradation. Cells *9*. 10.3390/cells9041028.
Seth, A., and Watson, D.K. (2005). ETS transcription factors and their emerging roles in human cancer. Eur J Cancer *41*, 2462-2478. 10.1016/j.ejca.2005.08.013.
Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. (2018). Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv:1811.00416. 10.48550/arXiv.1811.00416.
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. 10.1101/gr.3715005.
Smith, G.D., Ching, W.H., Cornejo-Paramo, P., and Wong, E.S. (2023). Decoding enhancer complexity with machine learning and high-throughput discovery. Genome Biol *24*, 116. 10.1186/s13059-023-02955-4.
Smith, H.S., Swint, J.M., Lalani, S.R., Yamal, J.M., de Oliveira Otto, M.C., Castellanos, S., Taylor, A., Lee, B.H., and Russell, H.V. (2019). Clinical Application of Genome and Exome Sequencing as a Diagnostic Tool for Pediatric Patients: a Scoping Review of the Literature. Genet Med *21*, 3-16. 10.1038/s41436-018-0024-6.
Spiess, K., and Won, H. (2020). Regulatory landscape in brain development and disease. Curr Opin Genet Dev *65*, 53-60. 10.1016/j.gde.2020.05.007.

Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.M., Mathelier, A., Lauter, G., Misra, S., Daub, C.O., Kere, J., and Swoboda, P. (2018). Characterization of the human RFX transcription factor family by regulatory and target gene analysis. BMC Genomics *19*, 181. 10.1186/s12864-018-4564-6.
Talos, F., Abraham, A., Vaseva, A.V., Holembowski, L., Tsirka, S.E., Scheel, A., Bode, D., Dobbelstein, M., Bruck, W., and Moll, U.M. (2010). p73 is an essential regulator of neural stem cell maintenance in embryonal and adult CNS neurogenesis. Cell Death Differ *17*, 1816-1829. 10.1038/cdd.2010.131.
Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res *35*, D88-92. 10.1093/nar/gkl822.
Wagih, O. (2017). ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics *33*, 3645-3647. 10.1093/bioinformatics/btx469.
Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A *104*, 18613-18618. 10.1073/pnas.0703637104.
Wang, Z., Oron, E., Nelson, B., Razis, S., and Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. Cell Stem Cell *10*, 440-454. 10.1016/j.stem.2012.02.016.
Weedon, M.N., Cebola, I., Patch, A.M., Flanagan, S.E., De Franco, E., Caswell, R., Rodriguez-Segui, S.A., Shaw-Smith, C., Cho, C.H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet *46*, 61-64. 10.1038/ng.2826.
Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat Commun *10*, 5241. 10.1038/s41467-019-13212-3.
Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523-527. 10.1038/nature19847.
Wright, C.F., FitzPatrick, D.R., and Firth, H.V. (2018). Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet *19*, 325. 10.1038/nrg.2018.12.
Xu, D., Gokcumen, O., and Khurana, E. (2020). Loss-of-function tolerance of enhancers in the human genome. PLoS Genet *16*, e1008663. 10.1371/journal.pgen.1008663.
Yousefi, S., Deng, R., Lanko, K., Salsench, E.M., Nikoncuk, A., van der Linde, H.C., Perenthaler, E., van Ham, T.J., Mulugeta, E., and Barakat, T.S. (2021). Comprehensive multi-omics integration identifies differentially active enhancers during human brain development with clinical relevance. Genome Med *13*, 162. 10.1186/s13073-021-00980-1.
Zhang, J., Lee, D., Dhiman, V., Jiang, P., Xu, J., McGillivray, P., Yang, H., Liu, J., Meyerson, W., Clarke, D., et al. (2020). An integrative ENCODE resource for cancer genomics. Nat Commun *11*, 3696. 10.1038/s41467-020-14743-w.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137. 10.1186/gb-2008-9-9-r137.
Zhao, X., and Moore, D.L. (2018). Neural stem cells: developmental mechanisms and disease modeling. Cell Tissue Res *371*, 1-6. 10.1007/s00441-017-2738-1.
Zurkirchen, L., Varum, S., Giger, S., Klug, A., Hausel, J., Bossart, R., Zemke, M., Cantu, C., Atak, Z.K., Zamboni, N., et al. (2019). Yin Yang 1 sustains biosynthetic demands during brain development in a stage-specific manner. Nat Commun *10*, 2192. 10.1038/s41467-019-09823-5.