Evolution is in the details: Regulatory differences in modern human and Neanderthal

Harlan R. Barker [1,2*], Seppo Parkkila [1,2], Martti E.E. Tolvanen [3]

1 Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland.

2 Fimlab Ltd., Tampere University Hospital, 33520 Tampere, Finland.

3 Department of Information Technology, University of Turku, 20014 Turku, Finland.

*To whom correspondence should be addressed.

Abstract

Transcription factor (TF) proteins play a critical role in the regulation of eukaryote gene expression by sequence-specific binding to genomic locations known as transcription factor binding sites.

Here we present the TFBSFootprinter tool which has been created to combine transcription-relevant data from six large empirical datasets: Ensembl, JASPAR, FANTOM5, ENCODE, GTEX, and GTRD to more accurately predict functional sites. A complete analysis integrating all experimental datasets can be performed on genes in the human genome, and a limited analysis can be done on a total of 125 vertebrate species.

As a use-case, we have used TFBSFootprinter to study sites of genomic variation between modern humans and Neanderthal promoters. We found significant differences in binding affinity for 86 transcription factors, groups of which are both highly expressed, and show correlation of expression, in immune cells and adult and developing neural tissues.

Introduction

North of Eden

Humans and their hominin relatives have been leaving Africa in waves for the past two million plus years. Various environmental and cultural pressures have impacted each diaspora in different ways, subsequently producing adaptations reflected in physiology, immunity, and brain size. In relatively recent history, the discovery and sequencing of DNA from remains of Neanderthal[1-5] and Denisovan[6, 7] has now allowed direct comparison of DNA from modern and ancient hominids. In the observed genomic variations between modern humans and Neanderthals, a limited number have been identified which occur in gene coding regions. Some of these are found in genes known to affect cognition and morphology (cranium, rib, dentition, shoulder joint)[1], pigmentation and behavioral traits[2], and brain development[3]. However, as has been noted before, there is a paucity of coding variations to explain the differences between related species; the genome of the Altai Neanderthal reveals just 96 fixed amino acid substitutions, occurring in 87 proteins. Unsurprisingly, a much larger set of variants are observed in intergenic regions, owing not only to the fact that these are comparatively much larger regions but also to the expectation of lesser conservation in what until recently was often termed "junk DNA". While variants in coding regions can directly affect protein structure, those found in intergenic regions may affect regulation of gene expression, through alternative binding of transcription factors in promoters and enhancers and expression of non-coding RNAs. What may surprise some, is the cumulative effect of numerous small – and large – changes to gene expression arising due to these manifold intergenic changes, and which may ultimately serve as the engine of speciation. Using a new computational tool we have created, which incorporates numerous transcription-relevant genomic features, we sought to reveal how the comparative differences in these regions may effect regulatory differences between modern human and Neanderthals. Specifically, our aim is to identify those gene-regulating transcription factors whose binding to DNA may vary between these species of hominid and thus drive the differences between them.

Transcription factors drive gene regulation

As early as 1963, Zuckerkandl and Pauling began addressing the apparent disparity in the fact that species with obvious differences could have proteins which look so similar. At the time they posited that it could be explained by the idea that "in some species certain products of structural genes inhibit certain sets of other genes for a longer time during development than in other species."[8]. Since this early exposition, the idea was more formally proposed in a publication by King and Wilson[9], and now the importance of regulatory regions in evolutionary adaptation has been further explored and accepted[10, 11].

Transcription factor (TF) proteins play a critical role in the regulation of eukaryote gene expression by sequence-specific binding to genomic locations known as cis-regulatory elements (CREs) or, more simply, transcription factor binding sites (TFBSs). TFBSs can be found both proximal and distal to gene transcription start sites (TSSs), and multiple TFs often bind cooperatively towards promotion or inhibition of gene transcription, in what is known as a cis-regulatory module (CRM). Because of the role these proteins play in transcription, discovery of TFBSs greatly furthers understanding of many, if not all, biological processes. As a result, many tools have been created to identify TFBSs. Owing to the time and material requirements of individual mutation studies needed for

experimental verification, many of these tools are computational. However, at issue in both the synthetic and experimental approach are two distinct problems. First is identification of where TF proteins bind to DNA. Experimental tools like ChIP-Seq are rapidly revealing the landscape of TFBSs for individual TFs in various cell types, and under various conditions. Computational tools often leverage this new data to build TFBS models which are increasingly accurate at making those predictions in silico. The second issue is how to determine if an experimental site (ChIP-Seq peak) or putative computational prediction (statistically likely) are actually biologically relevant, such as contributing to gene expression/repression, chromatin conformation, or otherwise. This second problem is by far the more demanding of the two, and it is here where we seek to levy the inclusion of transcription-relevant data.

The law of large numbers

Computational prediction of TFBSs seeks to enlist experimental data in the quest for the best possible specificity and sensitivity. However, computational modeling has deficits rooted in the law of large numbers; that is, because of the large size of a target genome, any method of prediction is bound to produce a large number of spurious results and filtering or thresholding results often means lost true positives. This can be compounded by the fact that biologically relevant TFBSs can be weakly binding [12].

Depending on the approach, the extent of incorporation of relevant experimental data varies widely. From very early on, the position weight matrix (PWM) has been used to represent and predict the binding of proteins to DNA. PWMs use a single, but very relevant, type of experimental data, that derived from observed binding events [13]. To create a PWM, a count is made of each of the four nucleotides at each position in the experimentally determined binding sites, known as a position frequency matrix (PFM). With the PFM, and using some contextual information about the target genome, a PWM probability model is generated which represents the binding preferences of a TF (detailed in Supplementary Methods). The PWM can then be used to arrive at a likelihood score for a target DNA region, which thus represents the likelihood of a TF binding to that DNA sequence. The accuracy of this method can continue to improve solely due to the large, and increasing, amounts of TFBS sequencing data provided by newer experimental technologies; like chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-Seq)[14], high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)[15, 16], and protein binding microarrays (PBMs)[17].

Old game new tricks

Updates to the traditional PWM have appeared over the years. One addresses the fact that traditional PWMs assume that the binding preference of a TF is independent at each position of the sequence it binds, resulting in both dinucleotide[18, 19] and position flexible[20-22] models. These become relevant when a TF has significantly different binding modes due to changes in conformation, structure, or splicing[23]. Ultimately, however, the difference between position independent and dependent models has been shown to be minimal, and TFs with more complex binding specificities can be accounted for using multiple position-independent PWMs[24, 25].

New statistical and computational algorithms, as they have come in vogue, have been brought to bear on the problem of TF-DNA binding: regression, Monte Carlo simulations, Markov models, machine learning, and deep learning, to name the most prominent. An important distinction is that

these, and many other, approaches seek to improve what their creators viewed as the primary aspect of prediction of TFBSs, the binding model itself. There has been great success in targeting this aspect, and while it is critical to know, the location of binding alone is not indicative of function[26]. Because of the incremental gains which have been achieved with new computational modelling of TF-DNA binding events, it can be argued that further work in this space is best served in discovery of binding sites for TFs which are not currently cataloged or modeled. Indeed, this is often a key difference to be noted when choosing between traditional PWMs and some of the later computational tools/models. The existence of a greater number of TF binding models is a strong reason to choose to incorporate the more extensively used traditional PWMs in binding prediction.

In the search for increased accuracy, other new models have improved TFBS prediction by instead incorporating other relevant biological data, for example: 3D structure of DNA [27-31], chromatin accessibility/DNAse hypersensitivity sites[32, 33], overlap in gene ontology[34], amino acid physicochemical properties[35], and gene expression and chromatin accessibility[36, 37]. These alternative models often match or outperform strictly sequence-based models[22, 31].

TFBSFootprinter incorporates transcription-relevant data

We sought to identify multiple sources of experimental data relevant to gene expression and transcription factor binding, and to incorporate it into a comprehensive model in order to improve prediction of functional TFBSs. Specifically, clustering of TFBSs has been shown to be an indicator of functionality[26, 38, 39]; conservation of genetic sequence across genomes of related species is one of the most successfully used attributes in identification of TFBSs[39, 40]; proximity to TSS is strongly linked to TFBS functionality[41]; correlation of expression between a transcription factor and another gene is an indication of a functional relationship[36, 42, 43]; variants in non-coding regions have a demonstrated effect on gene expression[44-46] and variants affecting gene expression are enriched in TFBSs[47]; open chromatin regions (ascertained by ATAC-Seq or DNAse-sensitivity) correlate with TF binding[48]; and finally, as previously mentioned, significant effort has gone into identifying the actual composition of the binding sites themselves through the use of sequencing of TFBSs (e.g., ChIP-Seq and HT-SELEX)[15, 21, 49].

Ensembl identifier-oriented system of analyses

For our tool, instead of using simple absolute genomic coordinates, the Ensembl transcript ID was chosen as the basic unit of reference. This is useful for several reasons. First, the Ensembl database is one of the most well-maintained biological databases in existence. It is continually updated and expanded and contains a wealth of sequence and regulatory information on a large number of vertebrate species. As a result, the tool we present here - TFBSFootprinter - can offer predictions in 270 vertebrates at the time of writing, from human to humpback whale, which will increase as the Ensembl database itself expands. Second, it allows the inclusion of important datasets which are gene-centric, such as FANTOM TSSs and expression data, GTEx eQTLs, and all annotations which are compiled within Ensembl itself. Finally, the Ensembl transcript ID provides an easy point of reference for a greater audience of scientists, thus increasing the accessibility and utility of the tool.

Non sum qualis eram

This study supports the hypothesis that future advances in prediction offered by incorporation of transcription-related biological data will outshine that which can be achieved by improvements in

modeling of TF-DNA binding alone. We show here that the TFBSFootprinter tool provides a good way to predict TFBSs based on incorporation of a variety of relevant biological data. As a proof of usage, we apply TFBSFootprinter in a comparative analysis of locations of variation in the promoters of modern humans and Neanderthal genomes.

Results

High-scoring TFBSs differ between modern human and Neanderthals

In analysis of 13,233 SNPs occurring in comparison of the modern human and Neanderthal promoteromes – the collection of all human/Neanderthal proximal promoters of protein-coding genes – a total of 85 TF models, representing 86 unique TF proteins, showed a significant difference in scoring between the two human species (Supplementary Table 1). Altogether 50 of the 86 differentially binding (DB) TFs are homeobox genes[50] and a further 8 are forkhead box genes, both of which are TF families well-established as drivers of development (Table 1).

Table 1. Homeobox TFs with differential binding affinity in modern human vs. Neanderthal proximal and semi-distal promoter regions

diff expressed tf	gene name	homeobox family
BARHL2	BarH-like homeobox 2	Barhl
CDX1	caudal type homeobox 1	Cdx
CUX1	cut-like homeobox 1	Cux
EVX1	even-skipped homeobox 1	Evx
HESX1	HESX homeobox 1	Hesx
HNF1A	HNF1 homeobox A	Hnf1
HNF1B	HNF1 homeobox B	Hnf1
HOXB5	homeobox B5	Hox5
HOXD8	homeobox D8	Hox6-8
HOXA10	homeobox A10	Hox9-13
HOXA13	homeobox A13	Hox9-13
HOXC10	homeobox C10	Hox9-13
HOXD11	homeobox D11	Hox9-13
HOXD13	homeobox D13	Hox9-13
HOXD9	homeobox D9	Hox9-13
ISL2	ISL LIM homeobox 2	Isl
LHX3	LIM homeobox 3	Lhx3/4
LMX1B	LMX LIM homeobox 1B	Lmx
NKX2-5	NK2 homeobox 5	Nk4
HMX2	H6 family homeobox 2	Nk5/Hmx
HMX3	H6 family homeobox 3	Nk5/Hmx
NKX6-2	NK6 homeobox 2	Nk6
ONECUT1	one cut homeobox 1	Onecut
ONECUT2	one cut homeobox 2	Onecut
ONECUT3	one cut homeobox 3	Onecut
OTX1	orthodenticle homeobox 1	Otx
OTX2	orthodenticle homeobox 2	Otx
PAX3	paired box 3	Pax3/7
PAX7	paired box 7	Pax3/7
PBX3	pre-B-cell leukemia homeobox 3	Pbx
PHOX2A	paired-like homeobox 2a	Phox
PHOX2B	paired-like homeobox 2b	Phox
PITX1	pituitary homeobox 1	Pitx

POU1F1	POU class 1 homeobox 1	Pou1
POU2F1	POU class 2 homeobox 1	Pou2
POU2F2	POU class 2 homeobox 2	Pou2
POU2F3	POU class 2 homeobox 3	Pou2
POU3F1	POU class 3 homeobox 1	Pou3
POU3F2	POU class 3 homeobox 2	Pou3
POU3F3	POU class 3 homeobox 3	Pou3
POU3F4	POU class 3 homeobox 4	Pou3
POU4F1	POU class 4 homeobox 1	Pou4
POU4F2	POU class 4 homeobox 2	Pou4
POU4F3	POU class 4 homeobox 3	Pou4
POU5F1	POU class 5 homeobox 1	Pou5
POU5F1B	POU class 5 homeobox 1B	Pou5
POU6F1	POU class 6 homeobox 1	Pou6
SIX3	SIX homeobox 3	Six3/6
UNCX	UNC homeobox	Uncx
VENTX	VENT homeobox	Ventx

DB TFs are highly expressed in immune cells

Data derived from the RNA-Seq experiments of the FANTOM5 project was used to generate a cluster map of DB TF expression within the 100 tissues with the highest aggregate expression, presented as Figure 1A. Relative aggregate expression for individual tissues is presented as bars in Figure 1B and shows that within the 100 tissues, the DB TFs have the highest expression in immune cells/tissues. The immune cell types with highest aggregate DB TF expression are eosinophils, mast cells, $CD14^{+}/CD16^{-}$ monocytes, and $CD8^{+}$ T cells (Supplementary Figure 1).

DB TFs coexpress in neural and immune tissues

Tissues from several tissue types cluster together by their expression of DB TFs, with the largest clusters being neural (30 brain and eye/retinal tissues) and immune (29 immune cells/tissues). In the brain tissues cluster (Figure 1B) there are a corresponding 12 TF genes with high expression in brain tissues of the FANTOM5 dataset: SIX3, NR2E1, OTX2, ONECUT2, POU3F1, OTX1, POU3F2, POU3F4, OLIG2, NKX6-2, POU3F3, and SOX2. All of these genes, except OTX1, are identified in the ProteinAtlas as being enriched or enhanced in brain. GO analysis of the 12 genes by g:Profiler[51] webserver shows enrichment for biological process GO terms 'forebrain development', 'head development, 'brain development', 'central nervous system development', 'nervous system development', 'glial cell differentiation', 'ensheathment of neurons', 'axon ensheathment', and 'myelination'. Additionally, within the brain tissues cluster we observe sub-clusters of fetal and newborn tissues (Figure 1D). Overall, DB TFs show the highest expression in medial frontal gyrus (newborn), medial temporal gyrus (adult), parietal lobe (fetal), and occipital lobe (fetal) (Supplementary Figure 1).

In a similar fashion, the cluster of immune cells/tissues has corresponding high expression in a cluster of 27 TF genes: ARID3B, ARID3A, POU2F2, POU6F1, NR1H3, LIN54, MAFG, MAFK, PBX3, POU2F1, CEBPA, IRF9, STAT6, ARID5A, IRF2, ZNF384, MAFF, CUX1, MEF2A, MEF2D, JUN, RXRA, TCF7L2, IRF1, CEBPB, MEF2C, FOS. All of these genes, except NR1H3 and CEBPA, have above average expression (greater than ~18 TPM), in at least one immune cell type as cataloged in the Database of Immune Cell eQTLs (DICE) [52]. GO analysis of the 27 genes (g:profiler webserver) shows enrichment

for REACTOME pathways 'cytokine signaling in immune system', 'MAPK targets', and 'immune system'.

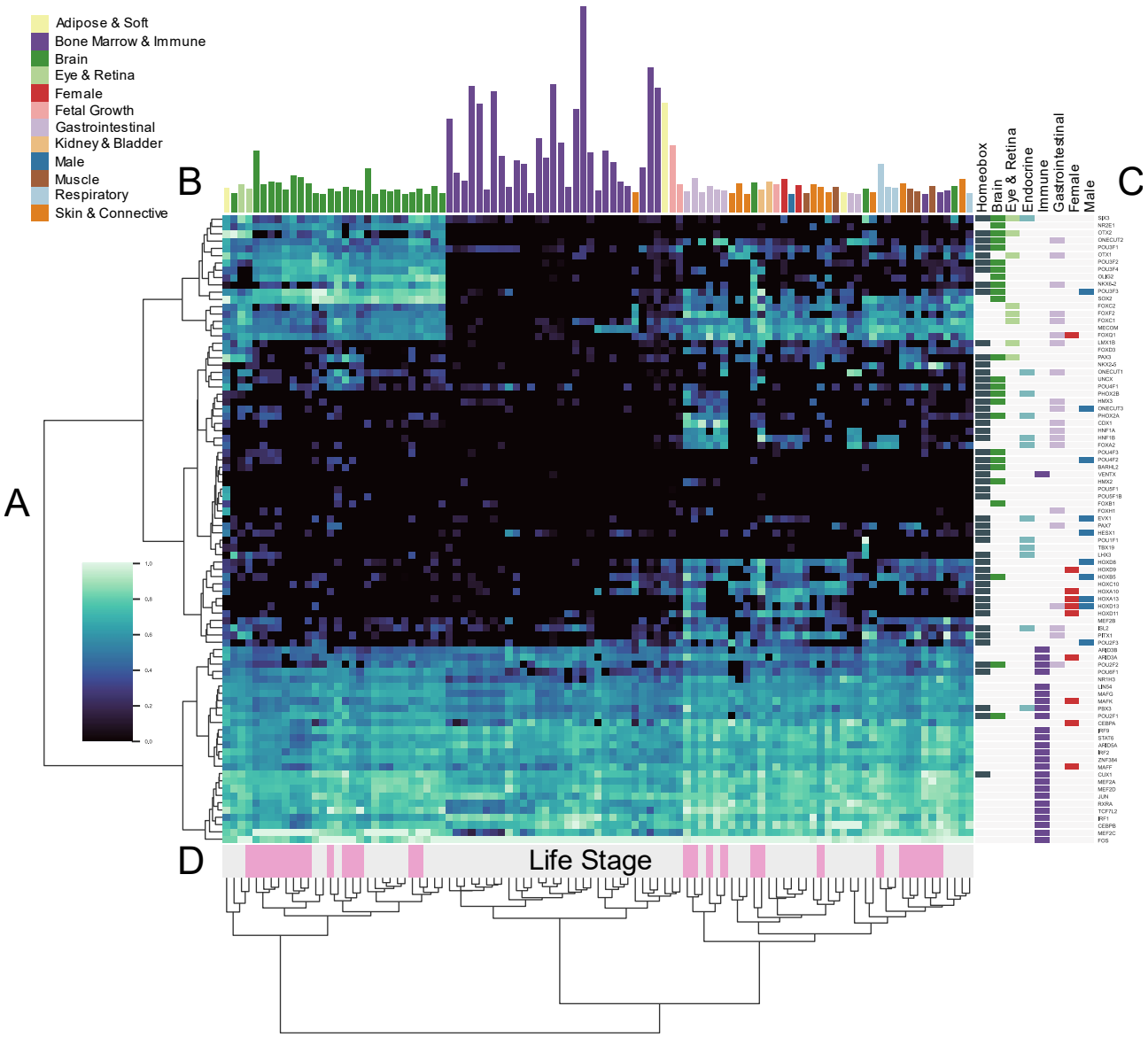


Figure 1. Cluster map of expression of transcription factors displaying variable binding in modern human vs. Neanderthal. (A) Expression data derived from the FANTOM 5 dataset is used to cluster 86 DB transcription factors in 100 tissues. (B) Aggregate expression for each tissue across all genes is presented as bars, and bars/tissues are colored by tissue family. Distinct clusters of brain and eye/retina tissues as well as immune cells/tissues are observed. (C) TF genes are labeled with colored bars corresponding to their identification as homeobox[50], above average expression in an immune cell type as defined by DICE database [52], or to have enriched or enhanced expression by tissue as defined by proteinatlas.org [53]. (D) Life stage for each tissue is described, where white and pink represents adult and fetal/newborn tissues, respectively. Expression values (TPM) have been normalized by log transformation and each tissue column has been adjusted on a zero to one scale.

DB TFs have developmental time-point related expression profiles in brain

Analysis of Allen Brain Atlas RNA-Seq expression data for DB TF genes revealed distinct clusters of time-point specific expression. Specifically, we observe clusters of expression for brain tissues at different grouped time points: pcw (8 to 37 weeks post conception), early (4 months to 4 years),

and late (8 years to 40 years). All of the pre-birth brain tissues cluster together, without admixture with the other grouped timepoints. The exception is a small cluster where 'cerebellar cortex' from all three groups is joined with 'pcw cerebellum' and 'pcw upper (rostral) rhombic lip'. All 15 pre-birth cortical brain tissues form a cluster. Similarly, a large cluster contains all of the other early and late tissues which primarily segregate into three age-related subclusters: D) late tissues, E) early tissues, and F) paired early/late tissues (amygdaloid, hippocampus, mediodorsal nucleus thalamus, and striatum).

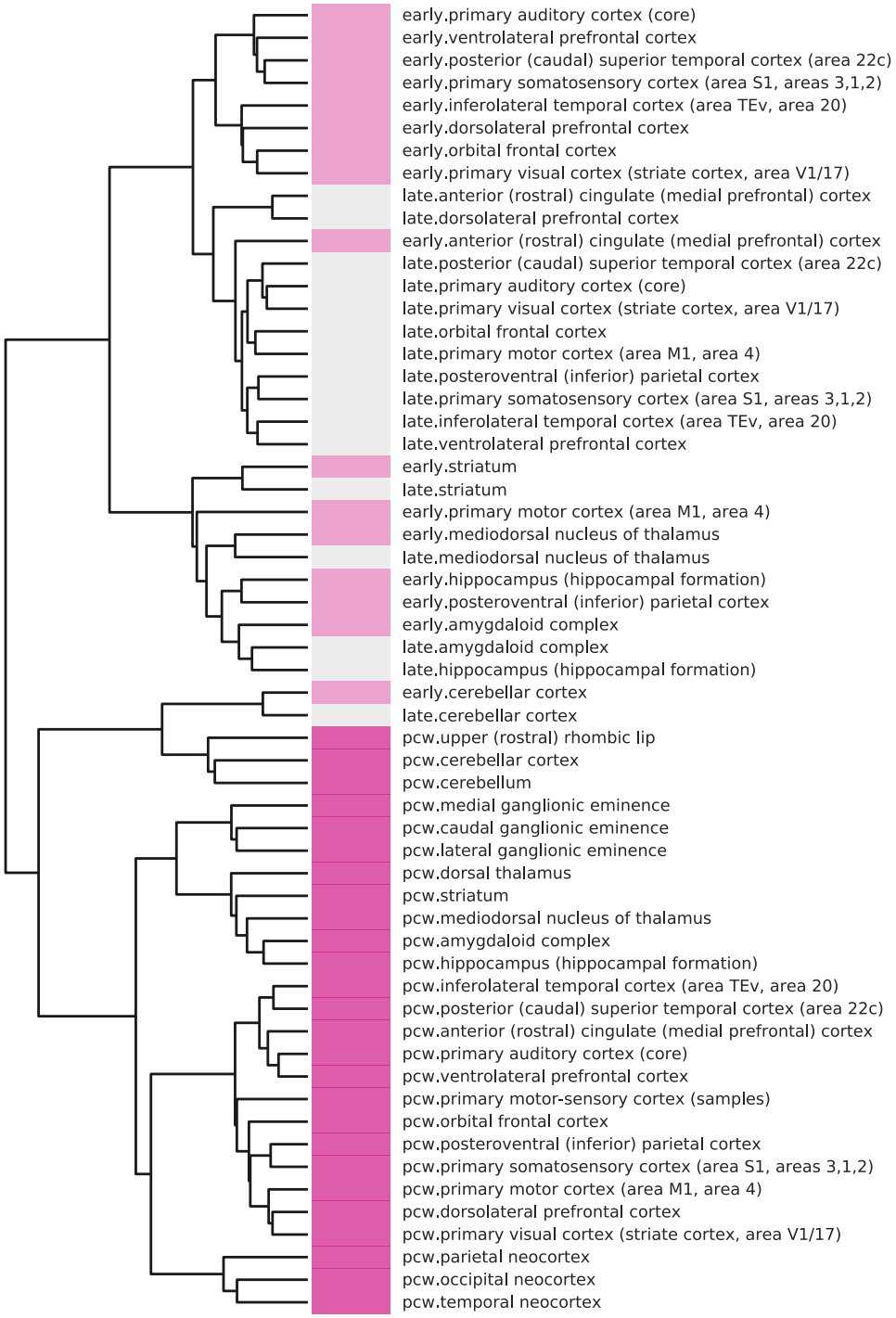


Figure 2. Cluster map of brain tissue expression of modern human vs Neanderthal DB TFs. Expression data from Allen Brain Atlas[54] was extracted as RPKM values for 26 unique tissues across 31 timepoints, converted to TPM and log transformed, and used to calculate the cluster map. Colors represent development stages: 8 to 37 weeks post conception (magenta), 4 months to 4 years (pink), and 8 years to 40 years (white).

Single-nucleus RNA-Seq of cortical cells reveals DB TF marker genes

Analysis of Allen Brain Atlas single-nucleus RNA-Seq data from 49,494 cortical brain cells was performed to identify which of the MH-Neanderthal DB TF genes may play a functional role in specific brain cell types. The clustering analysis identified 29 cell types which were stratified by brain region, cortical layer, and inhibitory/excitatory class (Figure 3A). Marker genes were identified for each of the cell type clusters and those clusters which had a MH-Neanderthal divergent binding TF as a marker gene were taken for further analysis. Clusters 9 and 24 each possessed 4 DB TFs as marker genes: CUX1, CUX2, POU6F2, and MEF2C. Analysis of these two clusters showed that the nuclei composing them primarily came from glutamatergic neurons of the L4c and L5 cortical layers of the primary visual cortex (V1C) (Figure 3A, 3G). Incorporation of cell type marker gene list from a classifier trained on mouse neurons [55] suggests that these two clusters contain cells of the types 'projecting neurons', granule neurons', and 'granule neuroblasts' (Figure 3B).

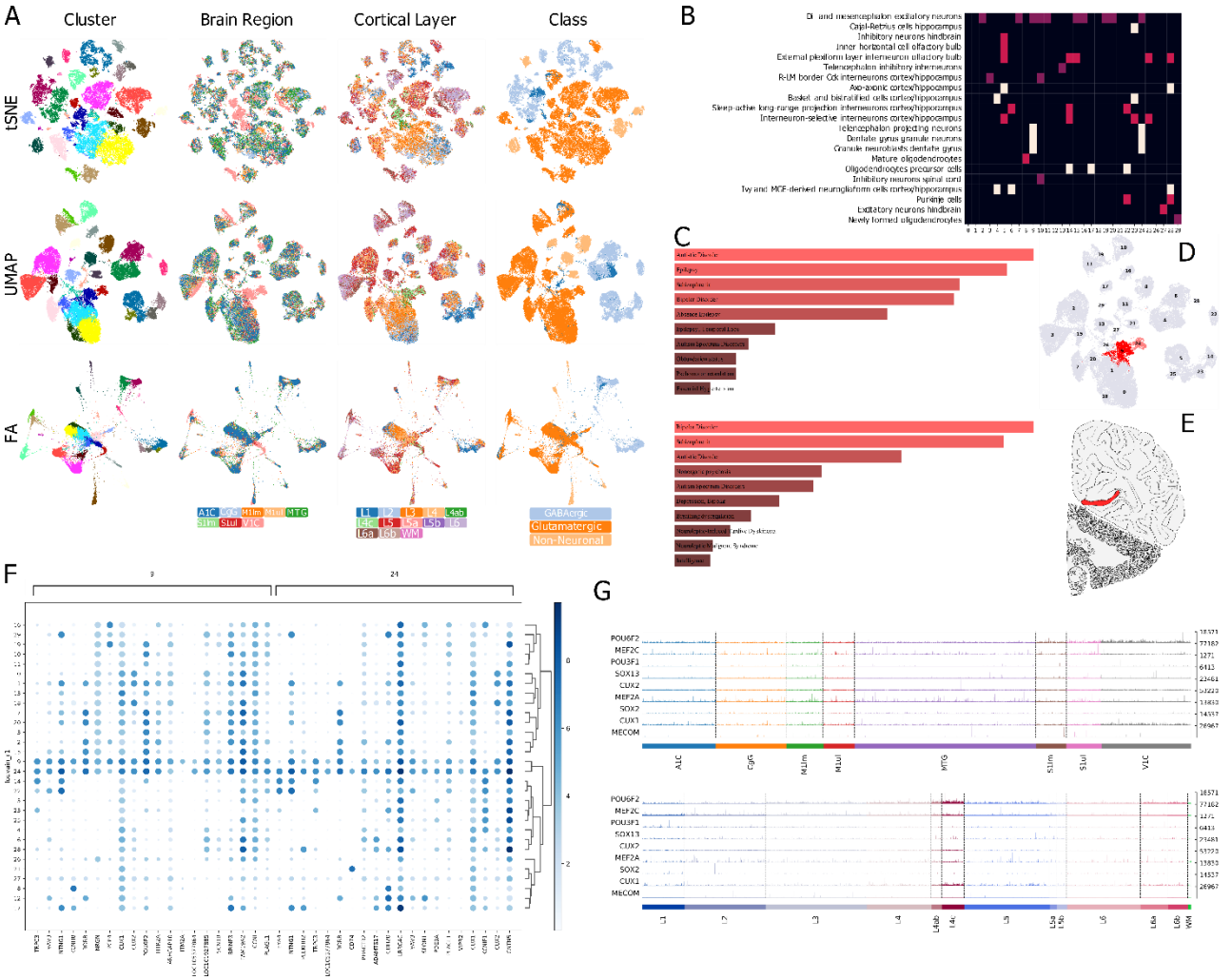


Figure 3. Analysis of single nucleus RNA-Seq data from brain cortical regions. A) tSNE, UMAP, and force-directed graph visualization of Allen Brain Atlas single nucleus RNA-Seq data from 49,494 brain cortex cells. Clustering was performed using Louvain community detection; brain region, cortical layer and excitatory/inhibitory status are mapped from annotations in the Allen Brain Atlas sample data. B) Presence of brain cell-type marker mouse gene orthologs, as classified previously, across the 29 Louvain-derived cell-type clusters. C) DisGeNET disease ontology terms enriched for the top 100 marker genes for cluster 9 (top) and cluster 24 (bottom). D) Location of clusters 9 and 24 in the UMAP visualization. These two clusters contain the most marker genes overlapping the TFs differentially binding in MH vs. Neanderthal. E) Visualization of the primary visual cortex, where the cells from clusters 9 and 24 predominantly originate. F) Dot matrix depicts expression of the top 20 marker genes for clusters 9 and 24, across all clusters. Dendrogram presents relationships of all cell-type clusters. G) Track plots of the six DB TFs which are marker genes in at least one cluster. Expression of these six genes is higher in primary visual cortex (V1C) and L4c and L5 cortical layers.

Gene ontology analysis of cell-type cluster marker genes

Gene ontology analysis of the top 100 marker genes from clusters 9 and 24 revealed associated diseases related to autism, bipolar disorder, and schizophrenia, among others (Figure 3C). The top three gene ontology analysis 'biological process' terms for cell type cluster 9 were "corticospinal neuron axon guidance (GO:0021966)" (>100-fold enrichment), "synaptic membrane adhesion (GO:0099560)" (~37-fold), and "glutamate receptor signaling pathway (GO:0007215)" (~19-fold). For 'cellular component' these terms were: "anchored component of presynaptic membrane (GO:0099026)" (~49-fold), "NMDA selective glutamate receptor complex (GO:0017146)" (~40-fold), and "intrinsic component of presynaptic active zone membrane (GO:0098945)" (~33-fold). For 'molecular function' these terms were: "transmembrane receptor protein tyrosine phosphatase activity (GO:0005001)" (52-fold), "calcium ion transmembrane transporter activity (GO:0015085)" (~9.5 fold), "calmodulin binding (GO:0005516)" (~8-fold). The full lists of ontology terms are included in Supplementary Table 2.

Likewise analysis of the top 100 marker genes from cluster 24 revealed 'biological process' terms: "synaptic transmission, glutamatergic (GO:0035249)" (~31-fold enrichment), "cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)" (~9-fold), and "calcium ion transmembrane transport (GO:0070588)" (~8-fold). For 'cellular component' these terms were: "anchored component of presynaptic membrane (GO:0099026)" (~52-fold), "intrinsic component of synaptic membrane (GO:0099240)" (~15-fold), and "postsynaptic density membrane (GO:0098839)" (~15-fold). For 'molecular function' these terms were: "calcium ion binding (GO:0005509)" (~4.5-fold). The full lists of ontology terms are included in Supplementary Table 2.

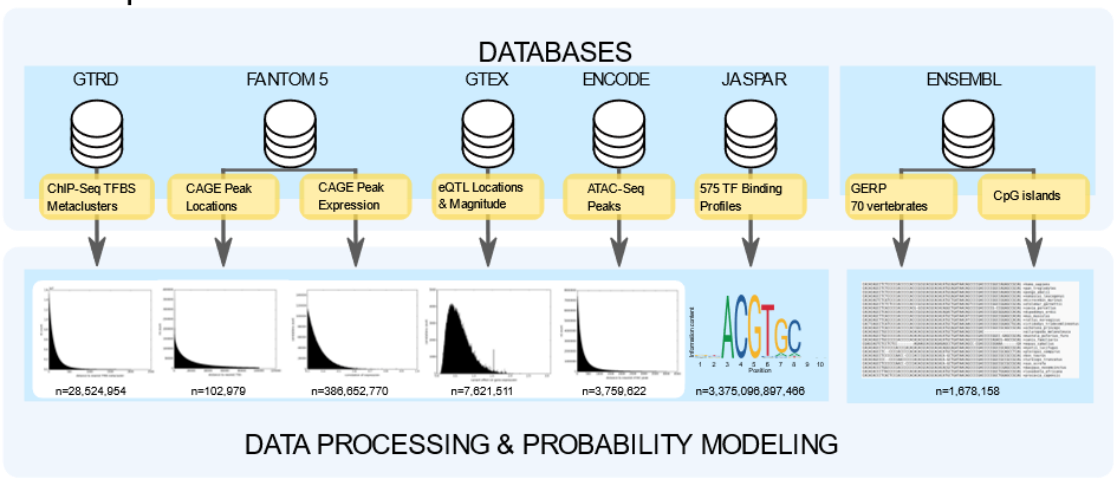
TFBSFootprinter availability

The tfbs_footprinter tool is available for use as a Python library (https://pypi.org/project/TFBS-footprinting/) and subsequently can be easily installed to a Linux system using a single command "pip install TFBS-footprinting". Due to size considerations, supporting experimental data for both human and non-human species is downloaded on demand on first usage. For Windows and Mac users (as well as those on Linux), a Docker image has been created (https://hub.docker.com/r/thirtysix/tfbs_footprinting) which includes all of the required dependencies and datasets for human, and can be installed with the docker command "docker pull thirtysix/tfbs_footprinting". Documentation on background, usage, and options is available both within the program and more extensively online (tfbs-footprinting.readthedocs.io).

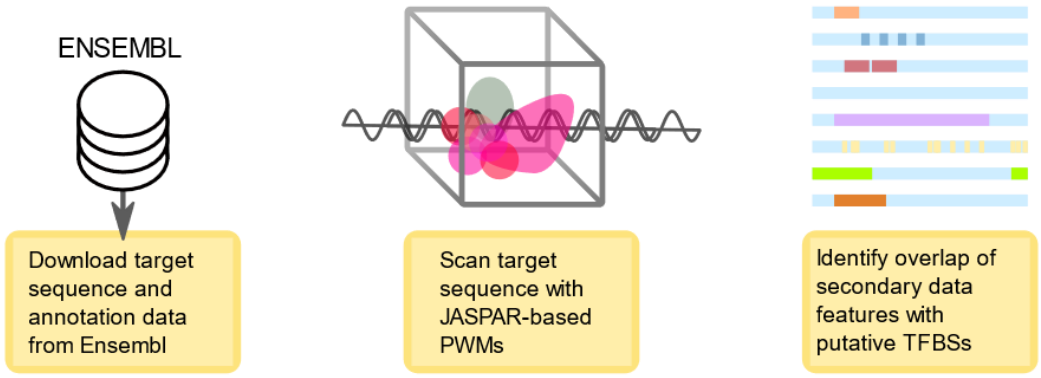
Experimental Datasets

Experimental data from a total of six databases were incorporated into the TFBSFootprinter algorithm (Figure 4A). Data from the relevant datasets were pre-processed to generate score distributions with which putative TFBS predictions could later be compared, as described in the Supplementary Methods. Each dataset allows for scoring of transcription-relevant markers in or near putative regulatory elements identified by PWM analysis: co-localization with ChIP-Seq metacluster, CAGE peak, ATAC-Seq peak, or CpG island; correlation of expression between predicted TF and gene of interest; co-localization of eQTL and effect on expression of target gene; measure of conservation in related vertebrate species (Figure 4C). The simplicity of this piece-wise approach allows for easy inclusion of additional TFBS relevant data in the future.

Precomputed



On Demand Analysis



Output

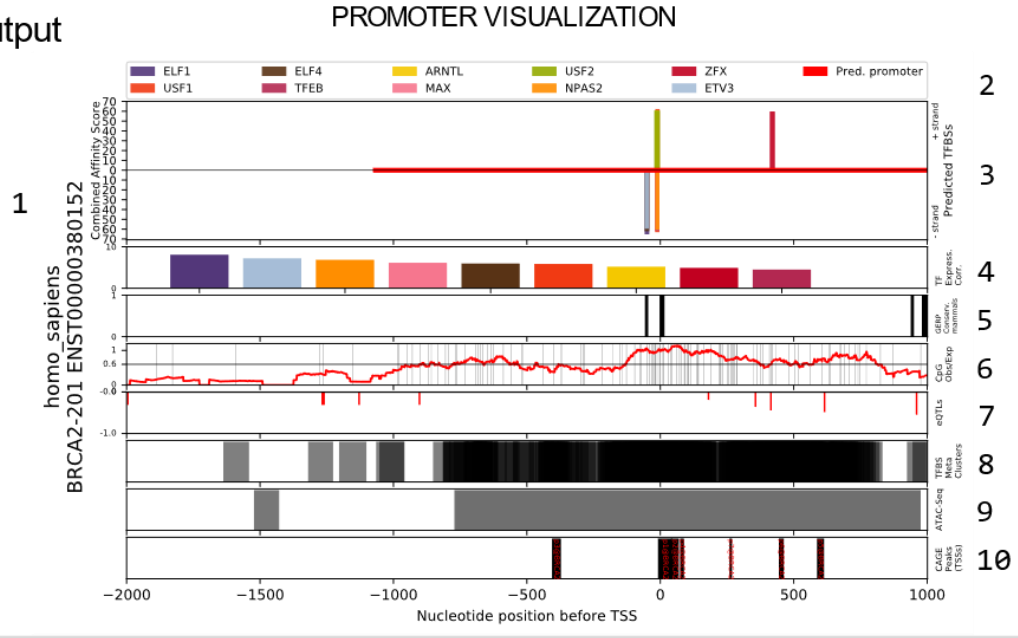


Figure 4. Outline of the datasets used in TFBS_footprinter. (**A**) A total of six empirical datasets are used to support computational prediction of TFBSs in the TFBSFootprinter tool. Experimental data is pre-processed to generate score distributions from which probability scores can be applied to putative TFBSs (n values indicate number of elements used to compute distributions). PWMs are created from JASPAR PFMs, and a segment of DNA is extracted from the Ensembl database for a region surrounding a target TSS of an Ensembl annotated transcript ID. Log-likelihood scoring is performed for six parameters on the target sequence. (**B**) A user-defined target promoter sequence and annotation data are downloaded, and JASPAR PWMs are created on-demand. PWM analysis of the promoter sequence generates putative TFBSs hits which are then compared with elements from pre-processed experimental datasets and scored using pre-generated probability scores relevant for the target genome. (**C**) The outputs of the TFBSFootprinter analysis are a publication-ready scalable vector graphics file (.svg), a table of results including predicted TFBSs names, locations, and scoring for each metric (not pictured), as well as individual files containing sequences and annotations. (C1) HUGO Gene Nomenclature Committee (HGNC)-based identifier + Ensembl transcript ID. (C2) Color coded legend of the top 10 transcription factors predicted to bind to this promoter. (C3) Graphical representation of promoter of transcript, predicted binding sites are indicated by bars. Bar height represents the combined affinity score (by default, a summation of log-likelihood scores from each transcription-relevant dataset, or alternatively a subset of datasets defined by user). Positive y-axis indicates binding on the positive (sense) strand and negative y-axis represents negative (anti-sense) strand. (C4) Score of the correlation of expression between each top predicted TF and the target gene. (C5) Highly conserved regions of 70-mammal alignment analyzed as determined by GERP analysis (black bars). (C6) Vertical lines represent CpG locations. The red line describes CpG ratio of human promoter sequence over a 100 nt window. (C7) Genetic variants identified in the GTEx database to have an effect on the target gene's expression (eQTLs). Green indicates positive impact on expression (positive y-axis) and red indicates negative (negative y-axis). (C8) TFBS metaclusters identified in the GTRD database (grey bars). (C9) Cell-type agnostic ATAC-Seq peaks (open chromatin) retrieved from the ENCODE database (grey bars). (C10) CAGE peaks indicating TSSs identified in the FANTOM database (black bars). Nucleotide positions at the bottom are relative to Ensembl defined transcription start site of the target transcript, and apply to C3 and C5-C10.

Inclusion of empirical datasets improves TFBSFootprinter accuracy

The performance of both individual datasets and combinations of datasets, in the identification of experimentally verified functional TFBSs and TFBS ChIP-Seq peaks, was tested by receiver operating characteristic (ROC) analysis (Figure 5, Figure 6; Supplementary Figures 2–3). Four different benchmarking approaches were used. The selection of true positives and true negatives is discussed in detail in the Supplementary Materials. Depending on the benchmarking approach used, different combinations of transcription-relevant features produced the best ROC scores. We observed that when using all available features TFBSFootprinter consistently outperformed the PWM, and that for the greater majority of TFs tested the best TFBSFootprinter model outperformed the best DeepBind model (benchmark 1, 60/65; benchmark 2, 28/40; benchmark 3, 13/14; and benchmark 4, 11/14). In the majority of cases a subset of the available transcription-relevant data produced the optimal model (Supplementary Figures 4–5). In addition, the features which were most frequently observed as components of the best models (Figure 5A, Figure 6A, Supplementary Figure 2A, Supplementary Figure 3A) varied by benchmark.

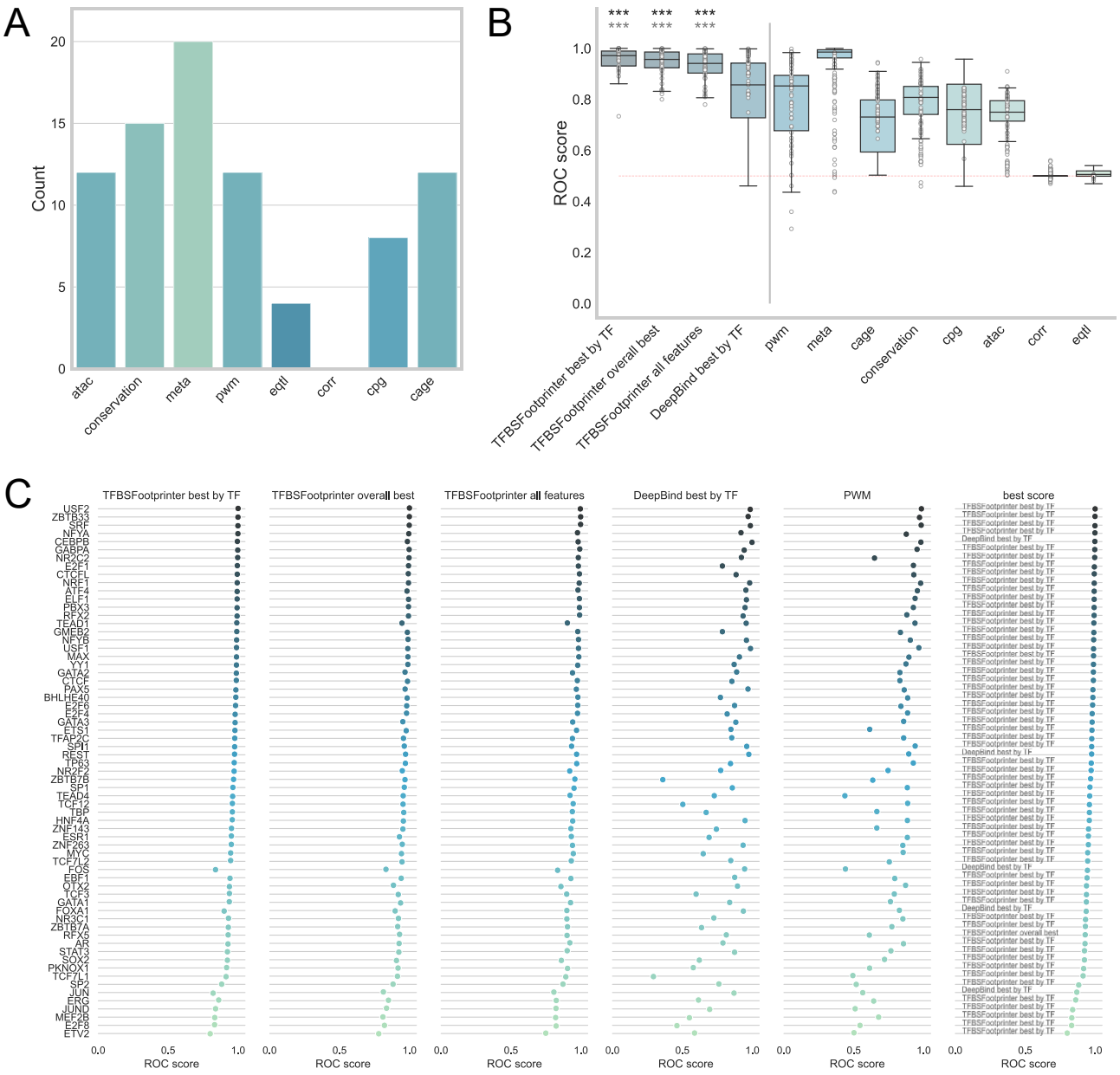


Figure 5. ROC Analysis model performance in the identification of TFBS ChIP-Seq peaks – strong and distal binding. ROC analysis was performed for TFs modeled by both JASPAR and DeepBind using associated GTRD ChIP-Seq peaks. Peaks with fold-enrichment ≥50x located within -900 and +200 bp relative to an Ensembl-defined protein-coding transcript as true positives, and paired sites 1,000 bp upstream as true negatives. ROC analysis was only performed on TFs which had at least 100 true positives and 100 true negatives. Each true positive/negative segment analyzed was 50 nucleotides long, with the highest TFBS score kept as representative. (A) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (B) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 40 TFs (left). ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (C) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by TF, based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models. Black asterisks denote significant difference of first four models with 'pwm' model and gray asterisks denote significant difference between TFBSFootprinter-based models and DeepBind model, as determined by related t-test; * p-value<0.05, ** p-value <0.005, and *** p-value <0.0005.

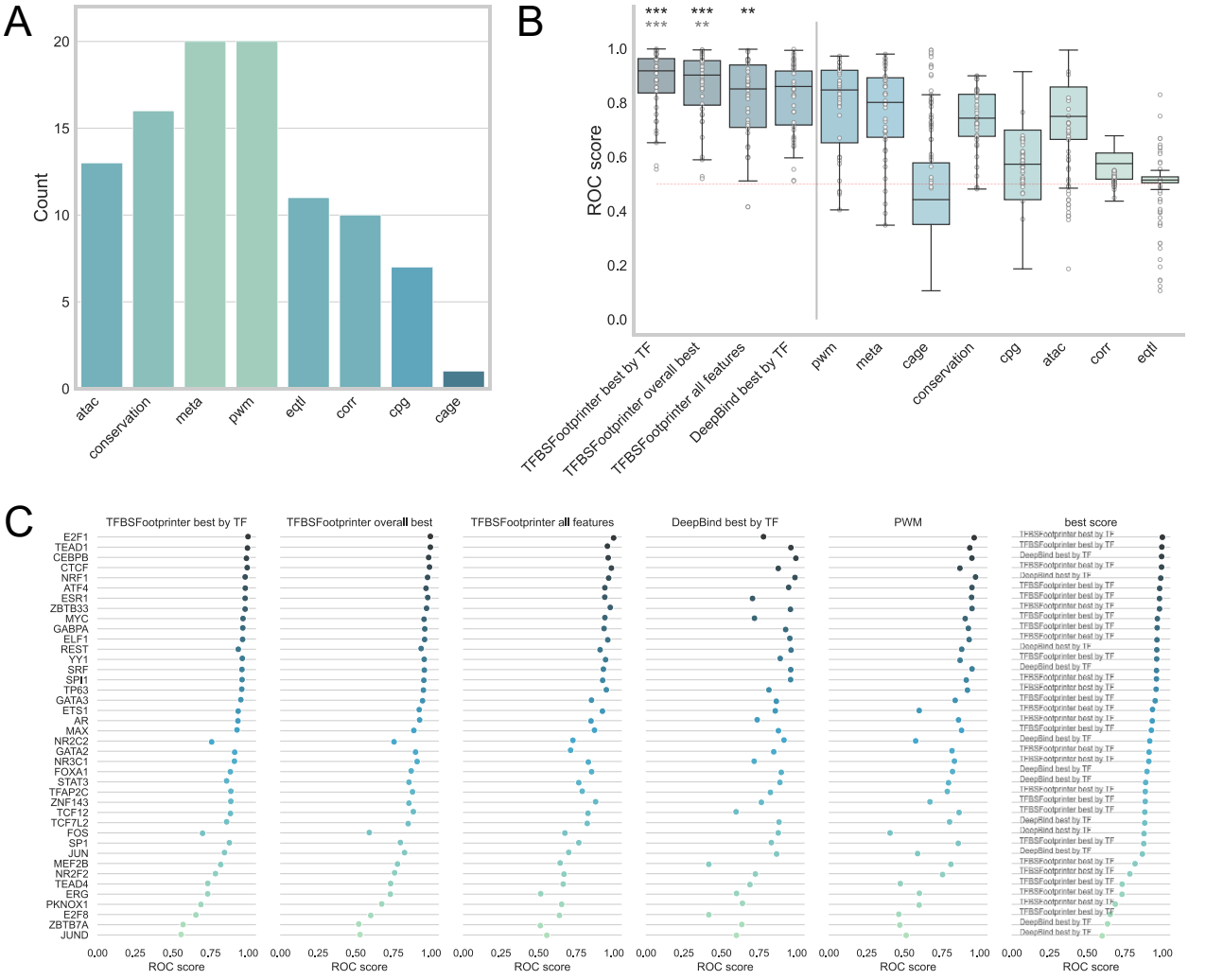


Figure 6. ROC Analysis model performance in the identification of TFBS ChIP-Seq peaks – strong and weak binding. ROC analysis was performed for TFs modeled by both JASPAR and DeepBind using associated GTRD ChIP-Seq peaks located within -900 and +200 bp relative to an Ensembl-defined protein-coding transcript, with fold-enrichment ≥50x as true positives, and those with fold-enrichment ≤2x as true negatives. ROC analysis was only performed on TFs which had at least 100 true positives and 100 true negatives. Each true positive/negative segment analyzed was 50 nucleotides long, with the highest TFBS score kept as representative. (A) Barplot of the frequency of experimental data type in the top 20 performing TFBSFootprinter models. (B) Boxplot of ROC scores for TFBSFootprinter and DeepBind for 40 TFs (left). ROC scores were also calculated based on using individual experimental metrics to show how well each contributes to accuracy of the combined model. (C) ROC scores for each individual TF tested, for each primary TFBS prediction model under study. The best scoring model among all is named for each TF (right). TFBSFootprinter best by TF, based on using the highest ROC score achieved by some combination of experimental data models; TFBSFootprinter overall best, based on using the combination of experimental data models which had the best average ROC score across all TFs analyzed; DeepBind best by TF, based on using the higher ROC score of the SELEX or ChIP-Seq DeepBind models. Black asterisks denote significant difference of first four models with 'pwm' model and gray asterisks denote significant difference between TFBSFootprinter-based models and DeepBind model, as determined by related t-test; * p-value<0.05, ** p-value <0.005, and *** p-value <0.0005.

Methods

Analysis of modern human vs. Neanderthal genetic variation

The Neanderthal Genome Project has cataloged total of 388,388 modern human vs. Neanderthal genomes SNPs [1]. These were reduced to 32,953 which occur in promoter regions (-900 to +100

nts relative to the TSS) of Ensembl-defined human transcripts. These were further reduced to 13,233 SNPs in the proximal promoters of those transcripts which are a defined by Ensembl as "protein-coding". Using the TFBSFootprinter tool, a 50bp region centered on each SNP was analyzed for binding of 575 TFs, for both for the modern human version and the Neanderthal version of the variant. TFBSFootprinter automatically retrieves the human sequences at a target region, and custom Python scripts were used to modify these sequences for analysis of the Neanderthal genome. All TFBS predictions which overlap the target SNP position were kept and the complete result set reduced using a Benjamini-Hochberg derived critical p-value corresponding to a false discovery rate cutoff of 0.01. For each putative TFBS meeting the cutoff in either subspecies, the corresponding matched pair of scores was kept. For each TF, using the compiled matched scores, the Wilcoxon rank statistical test was performed using the SciPy stats Python library[56] with a p-value cutoff of 0.01 used to identify statistically different scoring between subspecies.

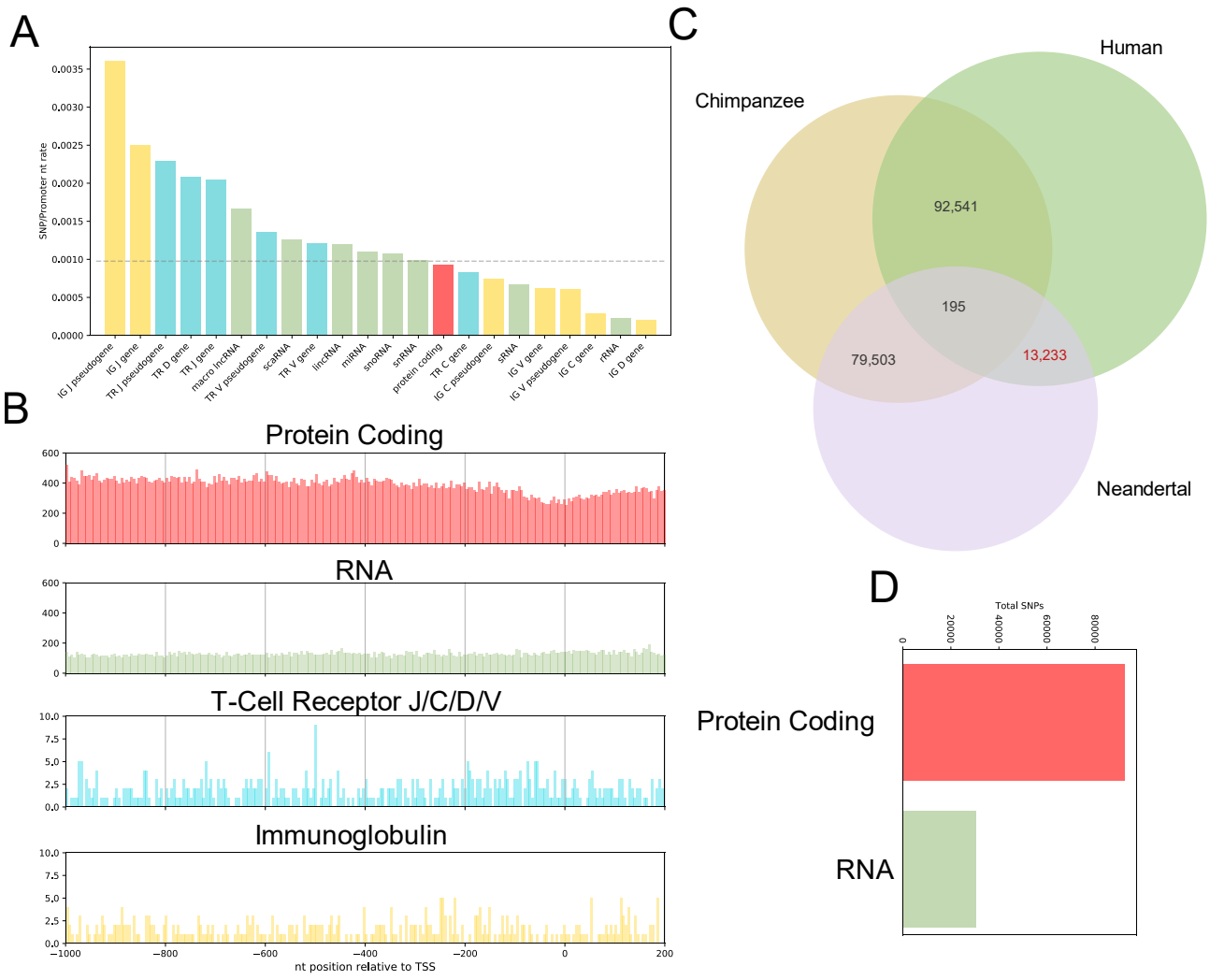


Figure 7. Modern human and Neanderthal variants used in analysis. A) Incidence rate of SNPs/nt in human promoters, of the various transcript classes, as defined by Ensembl. Bars are colored by transcript class, correspondent with panel B. B) Number of observed modern human/Neanderthal SNPs at each nt location in the promoters of various transcript types. A visible depletion of SNPs is observed for the promoters of protein-coding transcripts corresponding with increasing proximity to the TSS (nt position 0). C) Counts of total SNPs and indels in promoters of protein-coding transcripts, between the three species cataloged in the Neanderthal Genome Project. D) Comparison of counts of

human vs. chimp SNPs in promoters of protein-coding vs RNA transcripts. (IG J), immunoglobulin joining chain; (TR J), T-cell receptor joining chain; (TR D), T-cell receptor diversity chain; (TR V), T-cell receptor variable chain; (IG C), immunoglobulin constant chain; (IG V), immunoglobulin variable chain; (IG D), immunoglobulin diversity chain.

A total of 85 TF models were identified as scoring differently across human vs. Neanderthal SNP locations. For each of these we extracted RNA-Seq data from the FANTOM data set (across all CAGE peaks associated with that TF) and kept the data for the 100 tissues with the highest aggregate expression across all of the target TF genes. In the case of hetero-dimer JASPAR TF models (e.g., FOS::JUN, NR1H3::RXRA, and POU5F1::SOX2) the expression of each TF component gene was used. Expression was extracted as transcripts per million transcripts (TPM) and was normalized by $\log_2$ transformation. From the subsequent normalized expression data values, hierarchical clustering was performed and visualized using SciPy, Matplotlib, and Seaborn Python libraries[56-58].

The results of hierarchical clustering revealed a cluster of TF genes which showed unique expression in neural and immune tissues. These gene sets were used to perform PANTHER-based gene ontology enrichment analysis using the www.geneontology.com webserver[59, 60], with the default statistical settings using the Fisher's Exact test method and a FDR threshold of $p<0.05$.

Analysis of brain expression of differentially binding TFs

Expression data in the form of reads per kilobase per million reads (RPKM) were extracted for 26 unique tissues across 31 timepoints (from 8 weeks post-conception to 40 years after birth) from the Allen Brain Atlas (brainspan.org). RPKM values were then converted to TPM and $\log_2$ transformed. Ages were grouped into three phases of growth for simplicity of analysis and interpretation: pcw (8 to 37 weeks post conception), early (4 months to 4 years), and late (8 years to 40 years). Correspondingly, log-transformed TPM values were grouped and averaged and used to perform clustermap analysis to identify groupings tissues at time phases with similar expression profiles.

Analysis of brain sample scRNA-Seq

The Allen Brain Atlas has performed single-nucleus RNA-Seq analysis of 49,494 nuclei derived from 8 brain cortex regions within the middle temporal gyrus (MTG), anterior cingulate gyrus (CgGr), primary visual cortex (V1C), primary motor cortex (M1C), primary somatosensory cortex (S1C) and primary auditory cortex (A1C) [61]. These data were downloaded in the proprietary Allen Brain Atlas ".tome" format from the data portal (https://portal.brain-map.org/atlases-and-data/rnaseq). Exon and intron read counts were extracted using R library 'scrattch.io' (https://github.com/AllenInstitute/scrattch/) and loaded into Python library SCANPY [62]. Using a modified workflow described previously in [63], samples were filtered by Gaussian fit of read count ($75{,}000<x<1{,}600{,}000$), expressed gene count ($2{,}250<x$), and number of cells in which a gene is expressed (>50), resulting in a final count of 48,814 cells and 24,615 genes for further analysis. Counts were normalized by cell, log transformed, principle component analysis performed with 15 components, and k-nearest neighbors computed using SCANPY, and then the full data set normalized with R package 'scran' [64]. Batch correction by individual and sample region was performed with SCANPY. The top 4,000 genes with highly differential expression were identified for cluster analysis which was performed with three models for comparison: T-distributed stochastic neighbor embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and force directed graph models. The top 100 marker genes were identified as those with higher expression

unique to each cluster by Welch t-test in SCANPY. Expression of the DB TF list genes which were identified as a marker gene in a cluster was mapped onto cluster figures. Previously identified mouse brain cell type marker genes were extracted from a pre-trained classifier based on scRNA-Seq [55] to aid in identification of cell type clusters. Annotation data regarding brain region, cortical layer, and GABAergic/glutamatergic/non-neuronal cell type features were extracted from Allen Brain Atlas sample data for mapping onto derived cell type clusters. Imaging related to visualization of primary visual cortex (Figure 3E) were downloaded using the Allen Brain Atlas API [54].

Gene ontology analysis of target cluster marker genes was performed using the Protein Analysis Through Evolutionary Relationships (PANTHER) tool at the geneontology.org webserver [60]. Ontological term overabundance among cluster marker gene lists were established by Fisher's exact test and results were filtered by FDR<0.05; analyses were performed for 'biological process', 'molecular function', and 'cellular component' terms. Disease term gene ontology analysis was performed using Enrichr [65, 66] based on ontology compiled by DisGeNET [67].

TFBSFootprinter Methodology and Scoring

A computational pipeline was created in Python to allow for automated vertebrate promoter sequence retrieval from the Ensembl database (Ensembl version 94 was used in this analysis). The user-defined target sequence is then analyzed using 575 different transcription factor position weight matrices (PWMs), or a user-defined subset, derived from PFMs taken from the JASPAR database. Each TFBS prediction results in a log-likelihood score indicating the likelihood of a particular TF binding the DNA at that location. After this initial step, seven additional gene transcription related features are assessed for each TFBS prediction, each of which generate their own log-likelihood score based on proximity or overlap with these features. The features which may be considered for each TFBS prediction are: vertebrate sequence conservation (GERP), proximity to CAGE peaks (FANTOM5), correlation of expression between target gene and TF predicted to bind promoter (FANTOM5), overlap with ChIP-Seq TF metaclusters (GTRD), overlap with ATAC-Seq peaks (ENCODE), eQTLs (GTEx), and observed/expected CpG ratio (Ensembl). A summation of these scores, for each putative TFBS, then equals a value which we describe as the 'combined affinity score'. In this way the model's parameters are significantly more empirically flexible and therefore robust, and ultimately generate a more complete picture of a binding site instead of just computational prediction of binding affinity to a static set of nucleotides. In the Supplementary Materials are described the derivation of the log-likelihood scores for each component of the TFBS footprinting tool. A flowchart describing the steps involved in TFBS footprinter analysis is included as Supplementary figure 6.

Analyzing effect of feature combinations on TFBSFootprinter accuracy

Subsequently, for each method, and for each TF, the correlating true positive and true negative scores were used to generate receiver operator characteristic (ROC) curves using the 'roc_curve' module of the scikit-learn Python library [68].

Using the TFBSFootprinter tool, all 128 possible combinations of transcription-relevant features (PWM, CAGE, eQTL, metaclusters, ATAC-Seq, CpG, sequence conservation, expression correlation) which include PWM as a component were used in scoring of the true positives and true negatives. This allowed identification of the best possible feature-combination TFBSFootprinter model for each TF, which is described as 'TFBSFootprinter best by TF', as well as the TFBSFootprinter model which performed best on average across all TFs, named as 'TFBSFootprinter best overall'. In assessment of

the DeepBind tool both available models, based on SELEX or ChIP-Seq data, were used. Using related T-test, comparisons of ROC scores were made between PWM and each of TFBSFootprinter best by TF, TFBSFootprinter overall best, TFBSFootprinter all features, and DeepBind best by TF. Similarly, comparisons were made between DeepBind and each of TFBSFootprinter best by TF, TFBSFootprinter overall best, TFBSFootprinter.

Discussion

Regulatory changes drive evolution

Current knowledge points to the fact that, across the tree of life, prior to and during major radiations of new species, there comes a commensurate increase in the number of regulatory genes. Comparative genomic analyses suggest that at the time of eurkaryogenesis, or the origin of the last eukaryotic ancestor (LECA), a significant increase in novel TF classes occurred[69]. Eukaryogenesis is one of the major transitions of life on Earth, with an explosion of diversity owing to the endosymbiotic synthesis of an energy producing α-proteobacteria mitochondrion-progenitor within an archaeon[70]. If the new inhouse energy source could be described as the engine driving the diversity arising from LECA, the complementary argument could be made that TFs did the steering. Later, prior to the colonization of land by plants, there was an increase in the TF families of the ancestral aquatic streptophytes which were then already present in the first land plants[71]. Again, prior to the radiation of bilaterian (multicellular) metazoans an increase in transcription factor families occurred [72], roughly quadrupling in ratio [73]. Arguably, these increases in TF family numbers immediately prior to major transitions in life could be pointed to as evidence of the critical role of TFs in the rise and adaptation of eukaryotes and the complexity arising from their progenitors.

Far more recently, a transposon-mediated shift in regulatory signaling in mammalian pregnancy has led to the endometrial stromal cell type[74] and rewiring of a stress response producing the decidual stromal cell type[75]. In humans, the development of our cognitive skills is possibly due to a delay of synaptic-function gene expression and corresponding synaptogenesis in the pre-frontal cortex, owing to transcription factors such as myocyte enhancer factor 2A (MEF2A) [76-79]. At the same time, evidence is accumulating that maturation occurred earlier in the closely-related Neanderthal, as it does in chimpanzee [80], which is supported by samples of jaw [81], tooth [82], and, most importantly, cranium [83]. A number of SNPs have been identified in the promoter of MEF2A that have become fixed since separation of human and Neanderthal lineages [76], and its inclusion as a DB TF in our results, and as a marker gene in various cortical cell types, is exciting and warrants deeper investigation. Furthermore, development of globular brain structure was not present in modern humans at the time of divergence with Neanderthal and Denisovan lineages but has been a unique product of modern human development within the last 35,000–100,000 years[84] making it also particularly attractive for further analyses.

Modern human brain and immune regulatory changes

Hierarchical clustering of TF expression revealed that TF genes associated with neural and immune function formed distinct groups. Within the tissue clustering axis, in the neural cluster, adult and fetal/newborn neural tissues cluster separately. From this discovery in a broader set of data it

became important to focus more closely on the expression of these DB TFs in neural tissues specifically. Analysis of RNA-Seq data from whole brain tissues further revealed differences in expression of these genes in a chrono, and likely developmentally, dependent manner. Specifically, expression of DB TFs in all tissues for the two groups occurring within 8 weeks post conception to 4 years post birth segregated from all tissues for the group defined by 8 years to 40 years of age.

Recent proliferation of single-cell transcriptomics has begun to clarify the direct role of transcription factors in cell-type determination and identity [85]. To determine in which cell types DB TFs are expressed in adult brain, single-nucleus data for 49,494 cortical cells was examined. A subset of the DB TFs were annotated as marker genes in some of the 29 cell types identified by Louvain clustering: POU Class 6 Homeobox 2 (POU6F2), cut-like homeobox 2 (CUX2), cut-like homeobox 1 (CUX1), Myocyte Enhancer Factor 2A (MEF2A), MDS1 and EVI1 complex locus (MECOM), Myocyte Enhancer Factor 2C (MEF2C), POU class 3 homeobox 1 (POU3F1), SRY-box 2 (SOX2), and SRY-box 13 (SOX13).

Human-accelerated regions (HARs) are locations in the genome where the rate of evolutionary change has accelerated since divergence with chimpanzee. A study of HARs has shown that they are enriched for TFBSs generally, and for TFs associated with neural development specifically. The DB TFs identified in our study show significant overlap with TFs which have been revealed to be active in HAR regions and whose dysregulation is associated with mental disorder. POU6F2 possesses an intronic HAR where a mutant allele (GRCH38:chr7:39,033,595) is associated with ASD [86]. The promoter of the CUX1 gene has been determined by ChIA-Pet analysis of chromatin to interact with HAR426, located ~200kb away. Unrelated individuals with intellectual disability (ID) (IQ<40) and autism spectrum disorder (ASD) have been identified with a homozygous mutation (GRCh38:chr7:101,606,361) in this HAR. Luciferase reporter assay indicates that when the mutant version of this HAR interacts with the promoter of the CUX1 gene expression is increased by three-fold, while cultured differentiated neurons with enhanced CUX1 expression produced an increased synaptic spine density. Similarly, another HAR is shown to interact with the promoter of MEF2C, and a mutation in this HAR (GRCh38:chr5: 88,480,873) creates a putative MEF2A binding site, reducing expression by ~50%. Mutations in the MEF2C gene are associated with autism[87, 88], mental retardation[87-89], schizophrenia[90, 91], epilepsy[87, 88], and speech abnormalities[87].

It has been shown that for the early postnatal period of mouse, FOXP2 is a negative regulator of MEF2C, and that the likely result is promotion of synaptogenesis of cortical striatum[92]. Interestingly, the FOXP2 gene, perhaps most noted as a "speech" gene, does not show a significant difference in binding between modern human and Neanderthal in our results. However, in addition to the interaction with MEF2C mentioned, there is a POU3F2 (which our results show is a DB TF) binding site within the FOXP2 gene which has been shown to affect FOXP2 expression and which is associated with a selective sweep occurring since the divergence of humans and Neanderthal [93]. Further analysis is warranted to discover if a difference in FOXP2 binding is observed when analyzing other relevant areas, such as introns, 3' UTR, enhancers, etc.

Lots of work left to do

These and related questions led us to wonder what are the broader regulatory disparities between modern humans and Neanderthal. Using our novel tool TFBSFootprinter, which utilizes several transcription-relevant datasets to augment classical PWM scoring, we analyzed 13,233 promoterome SNPs occurring between the two species. What we have found is that developmental

homeobox and forkhead box genes dominate among those TFs which bind differentially among modern humans vs. Neanderthal, that for most TFs the strength of binding is more often increased in MH, that differentially-binding TFs are most strongly expressed in immunological cell types, and that DB TFs appear to coexpress in immune and neural tissues.

These observations are interesting as they appear to align with what is known already, or assumed, about the differences between modern human and Neanderthal, but also provide novel information for extended research.

Limitations

We have not included analyses of 3' UTR, introns, or enhancers. The SNP dataset we used for analysis is based on multiple Neanderthal individuals and has a lower sequencing coverage than newer datasets. The inclusion of multiple individuals is useful for comparing modern humans to Neanderthals as a group, but can only provide a more general comparison. Using a higher coverage dataset would allow greater assurance that SNPs under inquiry are legitimate. We chose to perform the TFBSFootprinter analysis using all transcription-relevant features available, in the future we plan to expand testing and assessment of empirical datasets and incorporate an option to use the combination of features which is proven best for each individual TF.

TFBSFootprinter

The TFBSFootprinter tool incorporates 7 different transcription-relevant empirical data features in the prediction of TFBSs. It can take as input any Ensembl transcript ID from any of 125 vertebrate species available in the Ensembl database. Starting with a list of Ensembl transcript ids for a target species (e.g., *Homo sapiens*) TFBSFootprinter will retrieve from the Ensembl REST server, a user-defined region of DNA sequence surrounding each transcription start site (TSS). The sequence is then scored using up to 575 JASPAR TFBS profiles. User-defined p-values may be used to filter results; and the corresponding score thresholds have been determined by scoring each JASPAR TFBS profile on the complete human genome. Each putative TFBS is then additionally scored based on proximity/overlap with TSS, TFBS metaclusters, open chromatin, and eQTLs which affect expression levels of the proximal gene, as well as conservation of sequence, correlation of expression with proximal (target) gene, and CpG content. Additional transcription-relevant data can be added easily in the future as is appropriate.

We believe that TFBSFootprinter provides an excellent way to predict TFBSs, thus easily supplementing current investigations into gene function or providing a means to perform larger scale analyses of groups of related target genes. The ability to identify conserved binding sites in a large number of species particularly widens its applicability to researchers studying various vertebrates. After completion of analysis a publication ready figure depicting the top scoring TFBS candidates is produced. Additionally, a number of tables (.csv) and JavaScript Object Notation (.json) files presenting various aspects of the results are output. Primary among these is a list of computational predictions in the target species which are supported by empirical data, sorted by a sum of the combined log likelihood scores (the combined affinity score). Importantly, scoring of non-human species becomes limited by the availability of external data for that species; at this time the only data available for non-human species are sequence conservation, CpG, and JASPAR motifs.

Funding

This work was supported by the Finnish Cultural Foundation and FimLab to HB, Academy of Finland to SP, and Jane & Aatos Erkko Foundation to SP.

Acknowledgements

Heini Huhtala is acknowledged for assistance in statistical techniques and professor Matti Nykter and Payam Emami Khoonsari PhD are gratefully thanked for discussions on practical and theoretical concerns.

Conflicts of Interest

Authors declare they have no conflicts of interest.

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH *et al*: **A draft sequence of the Neandertal genome**. *Science* 2010, **328**(5979):710-722.
2. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, Sawyer S, Fu Q, Heinze A, Nickel B *et al*: **Patterns of coding variation in the complete exomes of three Neandertals**. *Proc Natl Acad Sci U S A* 2014, **111**(18):6666-6671.
3. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C *et al*: **The complete genome sequence of a Neanderthal from the Altai Mountains**. *Nature* 2014, **505**(7481):43-49.
4. Prufer K, de Filippo C, Grote S, Mafessoni F, Korlevic P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyregne S *et al*: **A high-coverage Neandertal genome from Vindija Cave in Croatia**. *Science* 2017, **358**(6363):655-658.
5. Hajdinjak M, Fu Q, Hubner A, Petr M, Mafessoni F, Grote S, Skoglund P, Narasimham V, Rougier H, Crevecoeur I *et al*: **Reconstructing the genetic history of late Neanderthals**. *Nature* 2018, **555**(7698):652-656.
6. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL *et al*: **Genetic history of an archaic hominin group from Denisova Cave in Siberia**. *Nature* 2010, **468**(7327):1053-1060.
7. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C *et al*: **A high-coverage genome sequence from an archaic Denisovan individual**. *Science* 2012, **338**(6104):222-226.
8. Zuckerkandl E: **Perspectives in Molecular Anthropology**. In: *Classification and Human Evolution.* Edited by Washburn S. New York: Routledge; 1964: 265.
9. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science* 1975, **188**(4184):107-116.
10. Chen H, Li C, Zhou Z, Liang H: **Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases**. *Cell Syst* 2018, **6**(5):604-611 e604.
11. Liang H, Lin YS, Li WH: **Fast evolution of core promoters in primate genomes**. *Mol Biol Evol* 2008, **25**(6):1239-1244.
12. Castellanos M, Mothi N, Munoz V: **Eukaryotic transcription factors can track and control their target genes using DNA antennas**. *Nat Commun* 2020, **11**(1):540.
13. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: **Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli**. *Nucleic Acids Res* 1982, **10**(9):2997-3011.
14. Cheneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B: **ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments**. *Nucleic Acids Res* 2018, **46**(D1):D267-D275.
15. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G *et al*: **DNA-binding specificities of human transcription factors**. *Cell* 2013, **152**(1-2):327-339.
16. Ogawa N, Biggin MD: **High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro**. *Methods Mol Biol* 2012, **786**:51-63.
17. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nat Biotechnol* 2006, **24**(11):1429-1435.
18. Siddharthan R: **Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix**. *PLoS One* 2010, **5**(3):e9722.

19. Zhao Y, Ruan S, Pandey M, Stormo GD: **Improved models for transcription factor binding site identification using nonindependent interactions**. *Genetics* 2012, **191**(3):781-790.
20. Mathelier A, Wasserman WW: **The next generation of transcription factor binding site prediction**. *PLoS Comput Biol* 2013, **9**(9):e1003214.
21. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G *et al*: **JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework**. *Nucleic Acids Res* 2018, **46**(D1):D260-D266.
22. Ruan S, Stormo GD: **Comparison of discriminative motif optimization using matrix and DNA shape-based models**. *BMC Bioinformatics* 2018, **19**(1):86.
23. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X *et al*: **Diversity and complexity in DNA recognition by transcription factors**. *Science* 2009, **324**(5935):1720-1723.
24. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity**. *Nat Biotechnol* 2011, **29**(6):480-483.
25. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S *et al*: **Evaluation of methods for modeling transcription factor sequence specificity**. *Nat Biotechnol* 2013, **31**(2):126-134.
26. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y: **The functional consequences of variation in transcription factor binding**. *PLoS Genet* 2014, **10**(3):e1004226.
27. Chiu TP, Rao S, Mann RS, Honig B, Rohs R: **Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding**. *Nucleic Acids Res* 2017, **45**(21):12565-12576.
28. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS: **Origins of specificity in protein-DNA recognition**. *Annu Rev Biochem* 2010, **79**:233-269.
29. Zhou T, Yang L, Lu Y, Dror I, Dantas Machado AC, Ghane T, Di Felice R, Rohs R: **DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W56-62.
30. Chiu TP, Yang L, Zhou T, Main BJ, Parker SC, Nuzhdin SV, Tullius TD, Rohs R: **GBshape: a genome browser database for DNA shape annotations**. *Nucleic Acids Res* 2015, **43**(Database issue):D103-109.
31. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R: **Quantitative modeling of transcription factor binding specificities using DNA shape**. *Proc Natl Acad Sci U S A* 2015, **112**(15):4654-4659.
32. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data**. *Genome Res* 2011, **21**(3):447-455.
33. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK: **Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape**. *Nat Biotechnol* 2014, **32**(2):171-178.
34. Qian Z, Lu L, Liu X, Cai YD, Li Y: **An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization**. *Bioinformatics* 2007, **23**(18):2449-2454.
35. Khamis AM, Motwalli O, Oliva R, Jankovic BR, Medvedeva YA, Ashoor H, Essack M, Gao X, Bajic VB: **A novel method for improved accuracy of transcription factor binding site prediction**. *Nucleic Acids Res* 2018, **46**(12):e72.

36. Duren Z, Chen X, Jiang R, Wang Y, Wong WH: **Modeling gene regulation from paired expression and chromatin accessibility data**. *Proc Natl Acad Sci U S A* 2017, **114**(25):E4914-E4923.
37. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG: **Identification of transcription factor binding sites using ATAC-seq**. *Genome Biol* 2019, **20**(1):45.
38. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome**. *Proc Natl Acad Sci U S A* 2002, **99**(2):757-762.
39. Hemberg M, Kreiman G: **Conservation of transcription factor binding events predicts gene expression across species**. *Nucleic Acids Res* 2011, **39**(16):7092-7102.
40. Wenger AM, Clarke SL, Guturu H, Chen J, Schaar BT, McLean CY, Bejerano G: **PRISM offers a comprehensive genomic approach to transcription factor function prediction**. *Genome Res* 2013, **23**(5):889-904.
41. Koudritsky M, Domany E: **Positional distribution of human transcription factor binding sites**. *Nucleic Acids Res* 2008, **36**(21):6795-6805.
42. Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR: **Mapping functional transcription factor networks from gene expression data**. *Genome Res* 2013, **23**(8):1319-1328.
43. Ma S, Snyder M, Dinesh-Kumar SP: **Discovery of Novel Human Gene Regulatory Modules from Gene Co-expression and Promoter Motif Analysis**. *Sci Rep* 2017, **7**(1):5557.
44. Consortium GT: **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* 2013, **45**(6):580-585.
45. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M: **A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals**. *Nat Commun* 2016, **7**:11101.
46. Shi W, Fornes O, Mathelier A, Wasserman WW: **Evaluating the impact of single nucleotide variants on transcription factor binding**. *Nucleic Acids Res* 2016, **44**(21):10106-10116.
47. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J *et al*: **Systematic localization of common disease-associated variation in regulatory DNA**. *Science* 2012, **337**(6099):1190-1195.
48. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, Aerts S: **Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling**. *PLoS Genet* 2015, **11**(2):e1004994.
49. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D *et al*: **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium**. *Nucleic Acids Res* 2013, **41**(Database issue):D171-176.
50. Holland PW, Booth HA, Bruford EA: **Classification and nomenclature of all human homeobox genes**. *BMC Biol* 2007, **5**:47.
51. Reimand J, Kull M, Peterson H, Hansen J, Vilo J: **g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W193-200.
52. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G *et al*: **Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression**. *Cell* 2018, **175**(6):1701-1715 e1716.

53. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A *et al*: **Proteomics. Tissue-based map of the human proteome**. *Science* 2015, **347**(6220):1260419.
54. Ding SL, Royall JJ, Sunkin SM, Ng L, Facer BA, Lesnar P, Guillozet-Bongaarts A, McMurray B, Szafer A, Dolbeare TA *et al*: **Comprehensive cellular-resolution atlas of the adult human brain**. *J Comp Neurol* 2016, **524**(16):3127-3481.
55. Pliner HA, Shendure J, Trapnell C: **Supervised classification enables rapid annotation of cell atlases**. *Nat Methods* 2019, **16**(10):983-986.
56. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J *et al*: **SciPy 1.0: fundamental algorithms for scientific computing in Python**. *Nat Methods* 2020, **17**(3):261-272.
57. Hunter JD: **Matplotlib: A 2D graphics environment**. *Computing in science & engineering* 2007, **9**(3):90-95.
58. Michael Waskom OB, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, Adel Qalieh: **mwaskom/seaborn: v0.8.1 (September 2017)**. In., v0.8.1 edn: Zenodo; 2017.
59. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, Ami GOH, Web Presence Working G: **AmiGO: online access to ontology and annotation data**. *Bioinformatics* 2009, **25**(2):288-289.
60. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD: **PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools**. *Nucleic Acids Res* 2019, **47**(D1):D419-D426.
61. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, Close JL, Long B, Johansen N, Penn O *et al*: **Conserved cell types with divergent features in human versus mouse cortex**. *Nature* 2019, **573**(7772):61-68.
62. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis**. *Genome Biol* 2018, **19**(1):15.
63. Luecken MD, Theis FJ: **Current best practices in single-cell RNA-seq analysis: a tutorial**. *Mol Syst Biol* 2019, **15**(6):e8746.
64. Lun AT, Bach K, Marioni JC: **Pooling across cells to normalize single-cell RNA sequencing data with many zero counts**. *Genome Biol* 2016, **17**:75.
65. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**. *BMC Bioinformatics* 2013, **14**:128.
66. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A *et al*: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update**. *Nucleic Acids Res* 2016, **44**(W1):W90-97.
67. Pinero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, Garcia-Garcia J, Sanz F, Furlong LI: **DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants**. *Nucleic Acids Res* 2017, **45**(D1):D833-D839.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al*: **Scikit-learn: Machine Learning in Python**. *J Mach Learn Res* 2011, **12**:2825-2830.

69. de Mendoza A, Sebe-Pedros A: **Origin and evolution of eukaryotic transcription factors**. *Curr Opin Genet Dev* 2019, **58-59**:25-32.
70. Koonin EV: **Viruses and mobile elements as drivers of evolutionary transitions**. *Philos Trans R Soc Lond B Biol Sci* 2016, **371**(1701).
71. Catarino B, Hetherington AJ, Emms DM, Kelly S, Dolan L: **The Stepwise Increase in the Number of Transcription Factor Families in the Precambrian Predated the Diversification of Plants On Land**. *Mol Biol Evol* 2016, **33**(11):2815-2819.
72. Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM: **Genesis and expansion of metazoan transcription factor gene classes**. *Mol Biol Evol* 2008, **25**(5):980-996.
73. Paps J, Holland PWH: **Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty**. *Nat Commun* 2018, **9**(1):1730.
74. Lynch VJ, Leclerc RD, May G, Wagner GP: **Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals**. *Nat Genet* 2011, **43**(11):1154-1159.
75. Erkenbrack EM, Maziarz JD, Griffith OW, Liang C, Chavan AR, Nnamani MC, Wagner GP: **The mammalian decidual cell evolved from a cellular stress response**. *PLoS Biol* 2018, **16**(8):e2005594.
76. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y *et al*: **Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques**. *Genome Res* 2012, **22**(4):611-622.
77. Sousa AMM, Meyer KA, Santpere G, Gulden FO, Sestan N: **Evolution of the Human Nervous System Function, Structure, and Development**. *Cell* 2017, **170**(2):226-247.
78. Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, Kelso J, Nickel B, Dannemann M, Bahn S *et al*: **Transcriptional neoteny in the human brain**. *Proc Natl Acad Sci U S A* 2009, **106**(14):5743-5748.
79. Somel M, Liu X, Khaitovich P: **Human brain evolution: transcripts, metabolites and their regulators**. *Nat Rev Neurosci* 2013, **14**(2):112-127.
80. Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, Fobbs AJ, Sousa AM, Sestan N, Wildman DE *et al*: **Prolonged myelination in human neocortical evolution**. *Proc Natl Acad Sci U S A* 2012, **109**(41):16480-16485.
81. Lacruz RS, Bromage TG, O'Higgins P, Arsuaga JL, Stringer C, Godinho RM, Warshaw J, Martinez I, Gracia-Tellez A, de Castro JM *et al*: **Ontogeny of the maxilla in Neanderthals and their ancestors**. *Nat Commun* 2015, **6**:8996.
82. Smith TM, Tafforeau P, Reid DJ, Pouech J, Lazzari V, Zermeno JP, Guatelli-Steinberg D, Olejniczak AJ, Hoffman A, Radovcic J *et al*: **Dental evidence for ontogenetic differences between modern humans and Neanderthals**. *Proc Natl Acad Sci U S A* 2010, **107**(49):20923-20928.
83. Gunz P, Neubauer S, Maureille B, Hublin JJ: **Brain development after birth differs between Neanderthals and modern humans**. *Curr Biol* 2010, **20**(21):R921-922.
84. Neubauer S, Hublin JJ, Gunz P: **The evolution of modern human brain shape**. *Sci Adv* 2018, **4**(1):eaao5961.
85. Arendt D, Bertucci PY, Achim K, Musser JM: **Evolution of neuronal types and families**. *Curr Opin Neurobiol* 2019, **56**:144-152.
86. Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S *et al*: **Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior**. *Cell* 2016, **167**(2):341-354 e312.

87. Paciorkowski AR, Traylor RN, Rosenfeld JA, Hoover JM, Harris CJ, Winter S, Lacassie Y, Bialer M, Lamb AN, Schultz RA *et al*: **MEF2C Haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways**. *Neurogenetics* 2013, **14**(2):99-111.
88. Zhou WZ, Zhang J, Li Z, Lin X, Li J, Wang S, Yang C, Wu Q, Ye AY, Wang M *et al*: **Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype-phenotype correlations**. *Hum Mutat* 2019, **40**(6):801-815.
89. Nowakowska BA, Obersztyn E, Szymanska K, Bekiesinska-Figatowska M, Xia Z, Ricks CB, Bocian E, Stockton DW, Szczaluba K, Nawara M *et al*: **Severe mental retardation, seizures, and hypotonia due to deletions of MEF2C**. *Am J Med Genet B Neuropsychiatr Genet* 2010, **153B**(5):1042-1051.
90. Schizophrenia Working Group of the Psychiatric Genomics C: **Biological insights from 108 schizophrenia-associated genetic loci**. *Nature* 2014, **511**(7510):421-427.
91. Mitchell AC, Javidfar B, Pothula V, Ibi D, Shen EY, Peter CJ, Bicks LK, Fehr T, Jiang Y, Brennand KJ *et al*: **MEF2C transcription factor is associated with the genetic and epigenetic risk architecture of schizophrenia and improves cognition in mice**. *Mol Psychiatry* 2018, **23**(1):123-132.
92. Chen YC, Kuo HY, Bornschein U, Takahashi H, Chen SY, Lu KM, Yang HY, Chen GM, Lin JR, Lee YH *et al*: **Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c**. *Nat Neurosci* 2016, **19**(11):1513-1522.
93. Maricic T, Gunther V, Georgiev O, Gehre S, Curlin M, Schreiweis C, Naumann R, Burbano HA, Meyer M, Lalueza-Fox C *et al*: **A recent evolutionary change affects a regulatory element in the human FOXP2 gene**. *Mol Biol Evol* 2013, **30**(4):844-852.