Predicting the impact of sequence motifs on gene regulation using single-cell data

Jacob Hepkema[1], Nicholas Keone Lee[1,2], Benjamin J. Stewart[1,3,4], Siwat Ruangroengkulrith[5], Varodom Charoensawan[5,6,7], Menna R. Clatworthy[1,3,4], Martin Hemberg[1,2,*]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK
[2]The Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QN UK
[3]Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK
[4]Cambridge University Hospitals NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 0QQ, UK
[5]Department of Biochemistry, Faculty of Science, Mahidol University, Bangkok 10400, Thailand
[6]Integrative Computational BioScience (ICBS) Center, Mahidol University, Nakhon Pathom 7310, Thailand.
[7]Systems Biology of Diseases Research Unit, Faculty of Science, Mahidol University, Bangkok 10400, Thailand.

[*]Corresponding author: mh26@sanger.ac.uk

Abstract

Binding of transcription factors (TFs) at proximal promoters and distal enhancers is central to gene regulation. Yet, identification of TF binding sites, also known as regulatory motifs, and quantification of their impact remains challenging. Here we present scover, a convolutional neural network model that can discover putative regulatory motifs along with their cell type-specific importance from single-cell data. Analysis of scRNA-seq data from human kidney shows that ETS, YY1 and NRF1 are the most important motif families for proximal promoters. Using multiple mouse tissues we obtain for the first time a model with cell type resolution which explains 34% of the variance in gene expression. Finally, by applying scover to distal enhancers identified using scATAC-seq from the mouse cerebral cortex we highlight the emergence of layer specific regulatory patterns during development.

Introduction

One of the central goals in genomics is to understand how different phenotypes are determined by the DNA sequence. In higher eukaryotes the vast majority of bases represent non-coding DNA and our ability to predict the function of these sequences is incomplete. An important role for non-coding DNA is to regulate the expression of protein coding genes, and a key mechanism is through binding of TFs to proximal promoters and distal enhancers.

Although the underlying principles and mechanisms of TF binding have been studied extensively, determining which motifs are functional and quantifying their importance remains a major challenge [1–3].

Identifying regulatory motifs from sequence alone is hard because most motifs are degenerate and the number of matches in the genome is typically much larger than the number of true bound sites in higher eukaryotes [4]. Using epigenetic information such as TF expression levels, TF binding, chromatin accessibility and histone modifications it is possible to obtain more accurate models of which motifs are relevant for expression [5,6]. Today, assays for quantifying the epigenome can be carried out for single cells [7–9], but there are no computational methods that can take full advantage of these data to infer regulatory motifs and their impact on molecular phenotypes, e.g. expression levels or open chromatin.

Although there are methods for inferring regulatory motifs from single-cell data, e.g. SCENIC [10], HOMER [11] and Basset [12], none of these methods is capable of both *de novo* motif identification and quantification of the contribution to gene expression levels. A variety of machine learning approaches have been applied to bulk RNA-seq and ChIP-seq data to identify sequence and epigenetic features that determine gene expression [6,13–16]. In recent years, deep learning methods have become very popular [17,18] and they have been used for the task of predicting gene expression from sequence [6,19,20].

Here, we present scover, a shallow convolutional neural network for *de novo* discovery of regulatory motifs and their cell type specific impact on gene expression from single cell data (**Fig 1a**). Scover takes as input a set of sequences, e.g. promoters or open chromatin regions, along with measurements of their activity, e.g. expression levels or peak sizes. The output is a set of convolutional filters that can be interpreted as regulatory motifs along with a vector of cell type specific influence scores for each motif. Scover was implemented using the PyTorch framework [21], and it is available under the MIT licence at https://github.com/jacobhepkema/scover.

Single-cell measurements from sequencing experiments typically have a large fraction of zeros which makes analyses challenging [22]. To overcome these issues scover reduces the sparsity by pooling q cells (default q = 100) from the same cell type. While a large q often allows for more accurate predictions, the drawback is that smaller cell types will be excluded. The effect of the pooling can be measured by the correlation between predicted and observed values (**Fig S1**). Reassuringly, we have found that results for the datasets considered in this manuscript are robust with respect to q, and it is up to the user to strike the balance between model accuracy and cell type resolution. Sequences are fed into a single convolutional layer where d (default d = 300) filters representing regulatory motifs are combined before being fed to a neural network layer resulting in a predicted expression value. The shallow architecture requires fewer parameters which allows smaller input datasets to be used and it makes it more likely that a whole motif representation will be found [23].

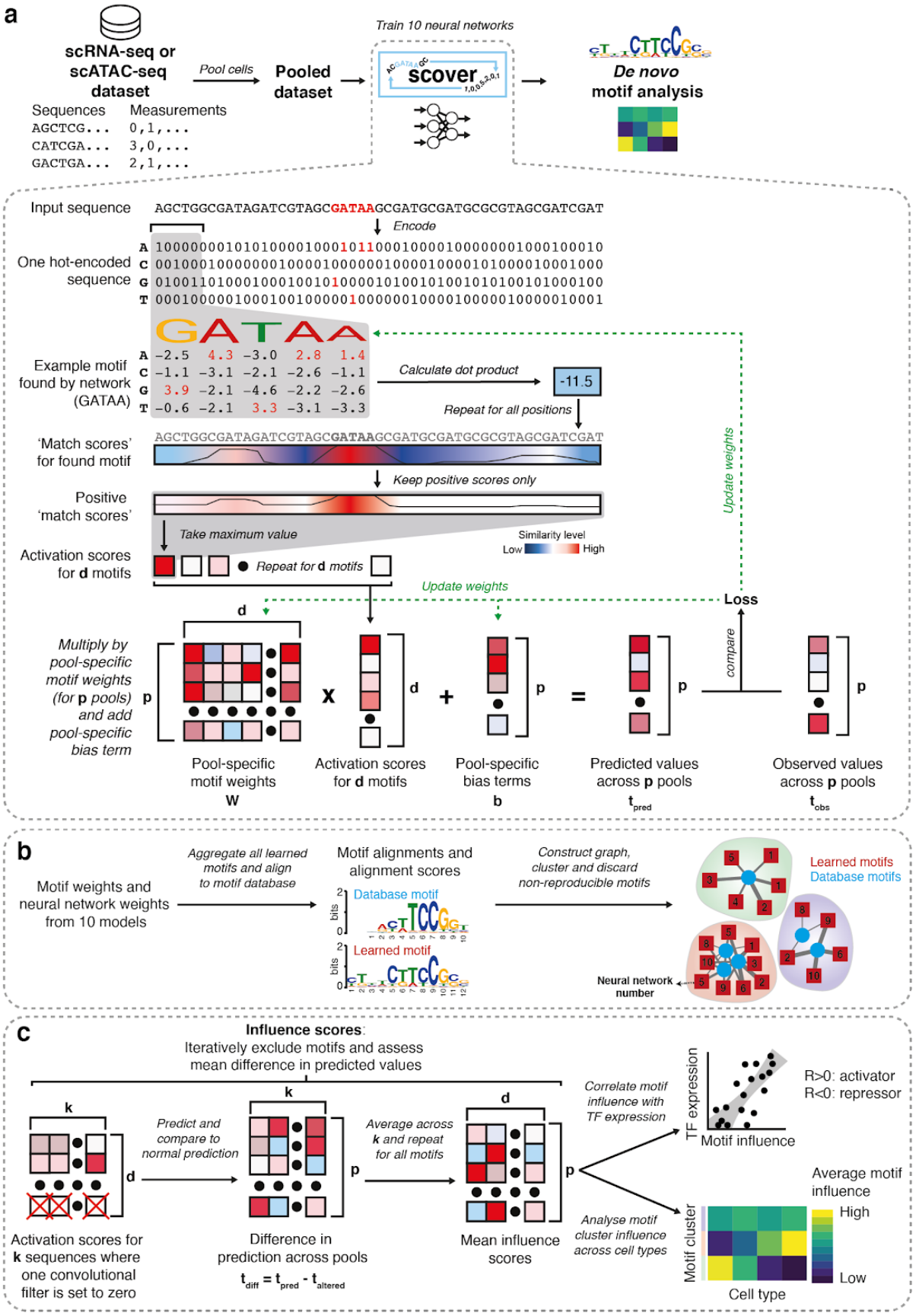

Figure 1: Overview of the scover workflow. (*a*) Cells are pooled to decrease the dataset sparsity and scover then infers regulatory motifs that are predictive of the signal associated

with a set of sequences using a single layer convolutional neural network. (*b*) The identified motifs are then merged and compared to annotated motifs if such a database is available. (*c*) To evaluate the contribution of each motif an influence score is calculated using a leave-one-out strategy.

Since the neural network is optimized using a stochastic approach, reproducibility can be achieved by running it r times (default r = 10) and excluding motifs that are found in fewer than 50% of models [24]. Since the randomly initiated convolutional filters may converge to the same point, many of the motifs identified by scover often have a high degree of similarity. To facilitate the overview of the outputs, scover automatically compares the motifs to those annotated in CIS-BP [25] (if available), and clusters the *rd* motifs into families (**Fig 1b**). Analysis of the convolutional filters and their associated weights can provide insights about the contribution of regulatory motifs to the observed measurements. Each motif i is assigned a match score s_{ij} for each input sequence j, and we consider a motif as active if $s_{ij}>s_i^m/2$, where s_i^m is the maximum activation score for i. The frequency score is defined as the fraction of sequences where a motif is active. To evaluate the importance of each motif, we employed a strategy inspired by Maslova et al [24] whereby an influence score n_{ip} is calculated for each motif i and pool p by excluding the motif from the analysis and comparing the prediction of the modified model to the original model (**Fig 1c**).

Since each motif family is typically bound by multiple different TFs, it is to be expected that clusters are heterogeneous. Thus, we recommend that motif families bound by many TFs are considered for further partitioning based on the influence scores through hierarchical clustering applied to the influence score matrix. We refer to a group of motifs with similar convolutional filters and influence scores as a motif cluster. Finally, influence scores in pools corresponding to the same cell type are averaged and influence scores for motifs belonging to the same cluster are summed. Since influence scores tend to be largely similar across cell types, we present the z-transformed values to highlight cell type specific patterns.

When more than one TF can bind to a motif cluster, additional information is required to determine their identities. Although transcriptome data alone is insufficient, we can rank the candidates by calculating the Spearman correlation r_{ct} between the summed influence scores for motifs in motif cluster c and the expression levels of TF t across all pools. We assume that if $|r_{ct}|$ is high, then the TF is more likely to bind to its cognate site, and if $r_{ct} > 0$ then c is an activator, otherwise it is a repressor.

Results

Scover identifies regulatory motifs in human kidney

We applied scover to a scRNA-seq dataset from human fetal and adult kidney containing a total of 63,480 cells [9]. The input sequences were 500 nts upstream and downstream of the transcription start site (TSS) of each expressed gene. We used d = 600, and of the 6,000 convolutional filters, 23% could be grouped into nine reproducible families corresponding to known motifs. The ETS family is heterogeneous, and sub-clustering yielded a total of ten motif clusters (**Fig. 2a,b, S2**). Closer inspection of the non-matching motifs revealed several

examples corresponding to poorly matching ETS motifs and motifs with high GC-content, suggesting that they are biologically meaningful nonetheless (**Fig. S3**). Based on the validation set, the scover model explains 15% of the variance in gene expression, and the performance drops to 12% if non-matching motifs are excluded.

To simplify the presentation of the results we grouped the 63 cell types into five categories: endothelial, immune, nephron epithelial, nephron progenitor and stromal (**Table S1**). Dimensionality reduction using principal component analysis (PCA) reveals that the biggest difference in influence is between the nephron progenitors and nephron epithelium cells. The second strongest trend separates immune cells from the remaining cell types (**Fig 2c**). Interestingly, the analysis also reveals a split between myeloid and lymphoid cells (**Fig S4a**).

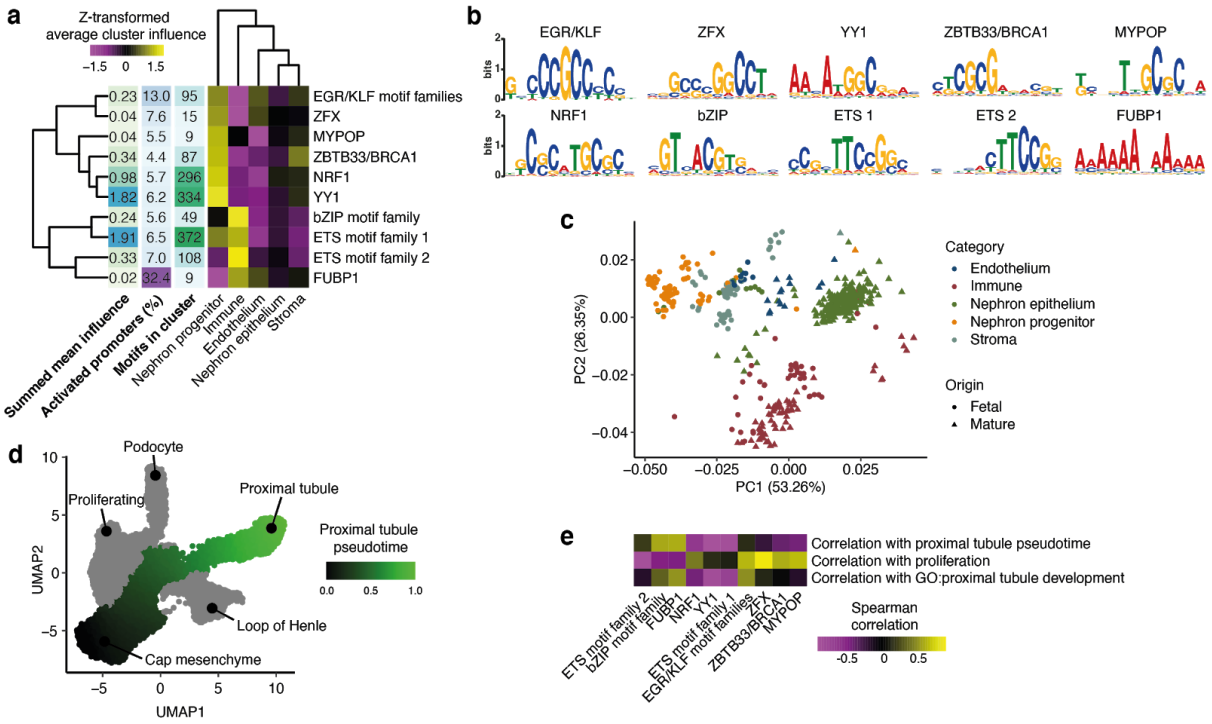

Figure 2: Analysis of proximal promoters for the fetal and adult human kidney. (*a*) Z-transformed influence scores for ten motif clusters and five cell categories. (*b*) Example motif logos from the ten motif clusters. (*c*) Projection of the pools onto the first two principal components of the influence score matrix reveals similar regulatory profiles. (*d*) Pseudotime trajectory for nephron progenitor development across single cells in fetal kidney represented as a UMAP plot. (*e*) Spearman correlations between motif influence scores and mean pseudotime values or expression of genes related to proliferation and proximal tubule development.

Since the largest differences in terms of influence scores are between nephron progenitors and nephron epithelium cells, we investigated the developmental trajectory by carrying out a pseudotime analysis. This revealed how the progenitors branch towards three distinct fates: podocytes, loop of Henle, and proximal tubules (**Fig 2d**). Focusing on the latter, we calculated the correlation between the expression levels of six markers associated with tubule development and the influence scores to reveal strong associations for bZIP and FUBP1 (**Fig 2e**). Similarly, we calculated the correlation between the pseudotime values and

the influence scores, revealing a similar profile, consistent with previous studies indicating a role for these motifs [26,27]. The same strategy can be used to relate motif clusters to other processes, and we find that ZFX, ZBTB33, MYPOP and KLF are positively correlated with 12 proliferation markers. This result is consistent with the known role of ZFX and KLF proteins in proliferation [28,29].

The frequency scores reveal that FUBP1 is the highest at ~30%, whereas most motifs only have good matches for 5-10% of promoters. Moreover, the aggregate influence scores show that ETS family 1, NRF1 and YY1 have the strongest impact on gene expression for all cell types. They have 2-100 fold higher influence than other motifs, suggesting that they are the dominant factors (**Fig 3a**). Both ETS and NRF1 are CpG-rich motifs (**Fig. 2b**), and their high influence scores are consistent with the observation that promoters with high CpG-content often are highly expressed [6]. It is noteworthy that KLF, ZFX and ZBTB33 have similar CpG-content, suggesting that the impact is specific to the motif and not to high CpG-levels in general [30]. Although YY1 has been classified both as a repressor and an activator [31], our results suggest that its main role is as an activator in the kidney. Inspection of influence scores across cell types reveals that KLF, ZBTB33, YY1 and ZFY have ~33% higher influence in the nephron progenitors compared to the nephron epithelium, whereas FUBP1 has ~33% higher influence in the nephron epithelium.

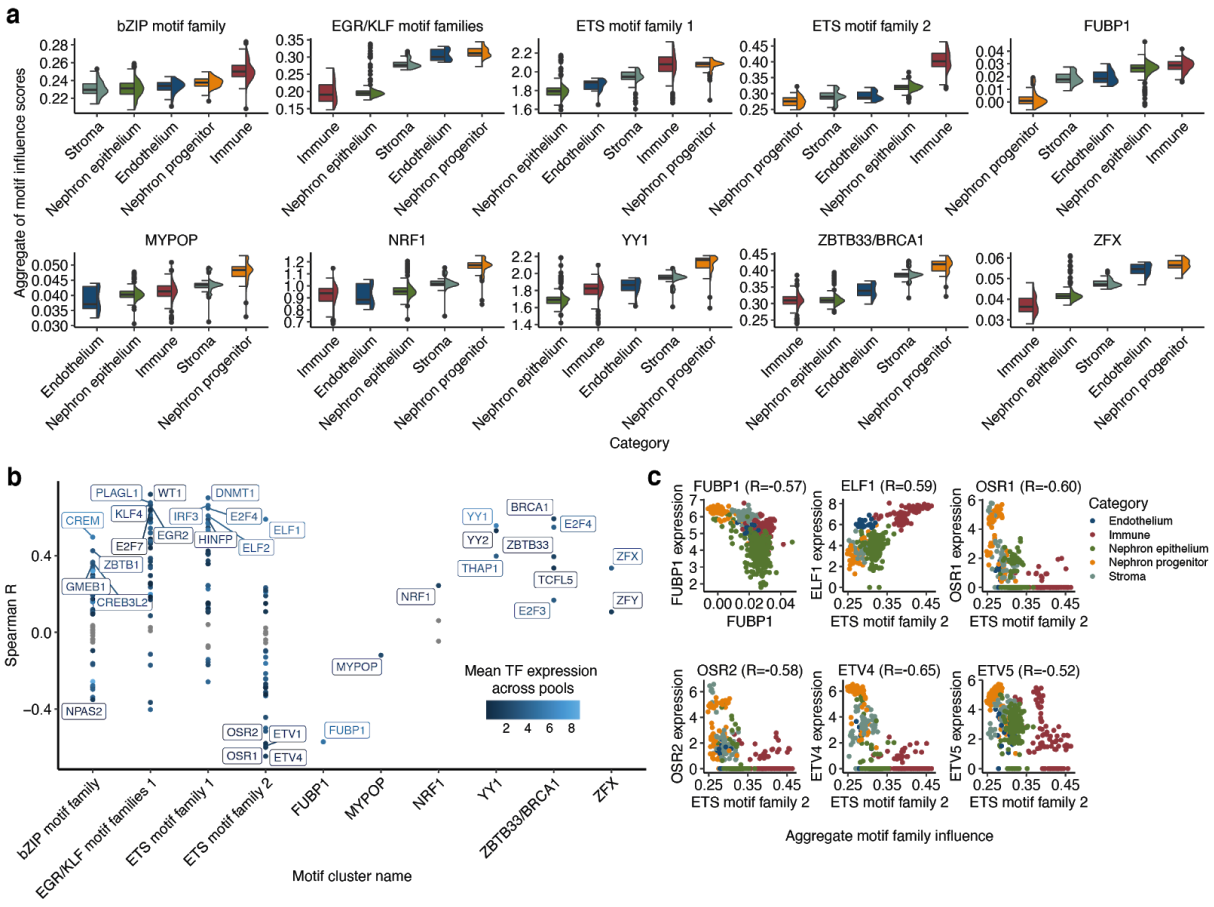

Figure 3: Motif cluster impact for the human kidney dataset. (*a*) Influence scores for ten motif clusters in five cell type categories. (*b*) Spearman correlation between motif influence scores and expression levels of TFs binding to the motif in question. TFs with *p*-value>=0.05

(after Benjamini-Hochberg correction) are in gray, the significant ones are colored by expression level, and selected top TFs for each cluster are named. (*c*) Correlation between motif influence and expression of selected TFs across pools. The Spearman correlation is reported at the top of each panel.

Comparison of TF expression levels and influence scores reveals that 39 of the 186 TFs associated with the nine motif families are strongly correlated with $|r|>0.5$ and a significant p-value (**Fig 3b**). For example, $r<0$ for *FUBP1* with influence scores close to zero for the nephron progenitors but positive elsewhere, suggests that *FUBP1* acts as a repressor in progenitors [32]. ETS family 2 shows a high positive correlation with *ELF1*, and our analysis indicates that it is most important in immune cells, consistent with previous studies [33]. For ETS family 2 we also find strong negative correlations for several TFs that are known to be important for kidney development: *OSR1*, *OSR2*, *ETV4* and *ETV5* (**Fig 3c**). All four factors have high expression and low influence scores in nephron progenitor cells, consistent with the reports that *OSR1* and *OSR2* act as repressors [34]. *ETV4* and *ETV5* are also known to be important for kidney development [35], but there are conflicting reports regarding their roles as activators or repressors. Taken together, this result supports the theory that ETS family TFs can bind similar sites and act in functionally redundant ways [36].

We can also use the expression information to further investigate interactions between regulatory elements. For example, the influence scores for ETS family 2 and KLF are negatively correlated (**Fig S4b**). Since the repressor and ETS family 2 member *OSR1* has $r<0$ (**Fig 3c**) [34], this suggests that TFs that are positively correlated with KLF could also serve as repressors. Indeed, it has been shown that *OSR1* and *WT1* interact synergistically during kidney organogenesis [37]. Similarly, the influence scores of KLF and bZIP are positively correlated, as well as their correlation with the expression with the TF *EGR1* and the DNA binding demethylase *TET1*. Previous studies have demonstrated synergistic binding between the two proteins [38].

Characterization of regulatory motifs in 18 mouse tissues

Next, we applied scover to 18 tissues from the Tabula Muris [8], containing a total of 67 cell types and 38,080 cells. Only 19.9% of the 6,000 convolutional filters matched seven different motif families, five of which were the same as for the kidney (**Fig 4a,b, S5**). Subclustering of the bHLH and ETS families resulted in nine different motif clusters. Again, we found that the non-matching motifs often resembled known motifs (**Fig S6**), and in their absence the fraction of variance explained is reduced from 34% to 23%.

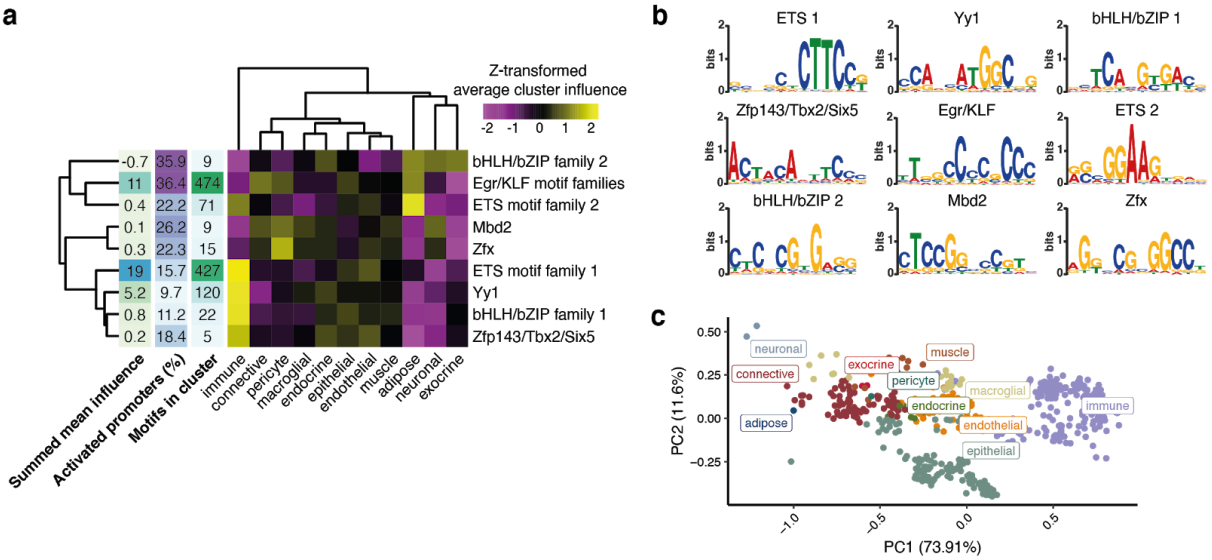

Figure 4: Analysis of proximal promoters from the Tabula Muris. (*a*) Z-transformed neural network influence for aggregated motif clusters. (*b*) Logos representing the nine motif clusters. (*c*) Pools are projected onto the space represented by the first two principal components of the motif influence matrix.

We categorized cell types as adipose, connective, endocrine, endothelial, epithelial, exocrine, immune, macroglia, muscle, neuron and pericyte (**Table S2**). Dimensionality reduction of the influence score matrix reveals three distinct groups: immune cells, neurons, and the remaining cells (**Fig 4c**), consistent with the expression analysis [39]. As with the kidney dataset we found that the immune cells separate well into known lineages (**Fig S7a**). The frequency scores show that bHLH family 2 and KLF are the most common as they are found at ~35% of promoters, with the other families found at <27% of promoters. The highest influence scores are found for ETS family 1, followed by Yy1 and KLF. Interestingly, bHLH family 2 has negative influence for all cell types (**Fig 5a)**. This finding is consistent with previous studies of the bHLH family highlighting the existences of a group acting as repressors [40].

Considering the pairwise correlations between the motif families, we note that they fall into two main groups (bHLH family 1, ETS family 1 and 2, Yy1, Zfp143 vs KLF, bHLH family 2, Mbd2, Zfx) (**Fig S7b**). The first group has the highest influence for immune cells, with the lowest values for neuronal and adipose cells. By contrast, the second group has high influence for neurons and pericytes, with the lowest values for exocrine and immune cells (**Fig 5a**). We hypothesized that in addition to being associated with specific cell types, the two groups of motifs would also be associated with distinct sets of cellular processes. We obtained 16,610 gene lists representing a diverse set of processes from the Gene Ontology (GO) database [41], and for each list we calculated the correlation between the average expression of the genes and influence scores across pools. We kept the terms with the top 1% of correlation scores, and as expected the motif clusters form the same two groups as before (**Fig 5b, Table S3**). Interestingly, the GO terms also fall into two groups. Reassuringly, the influence scores for the motifs that are highest in immune cells were positively correlated with genes associated with immune related terms. The motifs in this group have previously been reported to be important for immune cells [13,24,42,43–45]. The second

group corresponds to GO terms that are primarily related to differentiation and morphogenesis. This finding is also consistent with previous literature implicating the importance of KLF and bHLH proteins for differentiation [46–48].

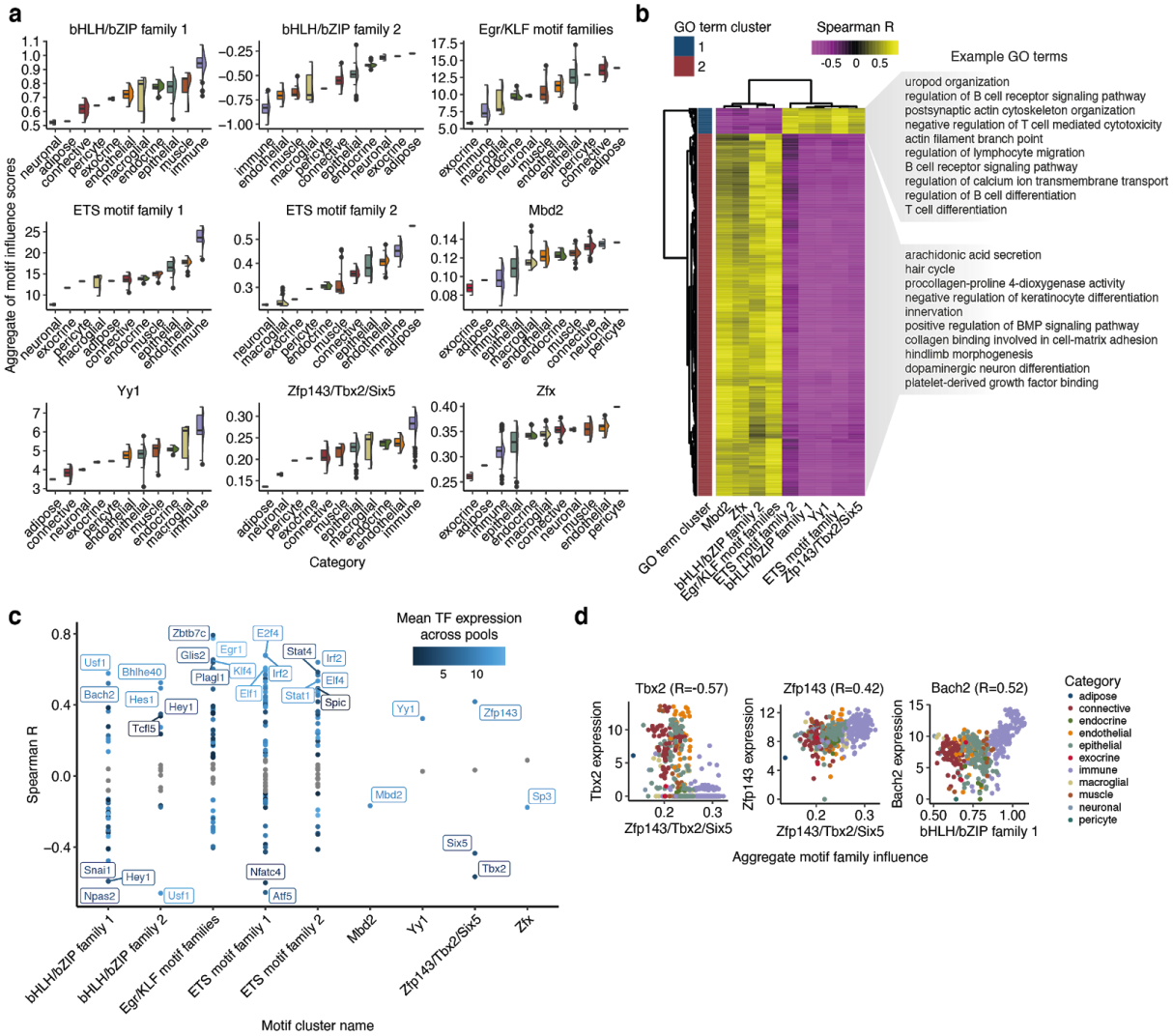

Figure 5: Motif cluster impact for the Tabula Muris dataset. (*a*) Aggregate motif influence scores for the different motif families. (*b*) Spearman correlations between expression of GO terms and aggregate motif influences in pools. Randomly selected GO terms are shown for each cluster. (*c*) Spearman correlation between motif influence and expression levels of TFs binding to the motif in question. TFs with *p*-value>=0.05 (after Benjamini-Hochberg correction) are in gray, the significant ones are colored by expression level, and selected top TFs for each cluster are named. (*d*) Correlations between aggregate motif family influence scores in pools and expression for selected TFs. The Spearman correlation is reported at the top of each panel.

The highest correlations between influence scores and TF expression levels reveals several known regulatory principles (**Fig 5c**). For example, for the Zfp143/Tbx2 family we find a negative correlation with the expression of *Tbx2*, suggesting that it acts as a repressor (**Fig 5d**). Furthermore, we find that *Tbx2* has high expression in some of the cell types from the heart, consistent with reports in the literature suggesting that it can repress cardiac

development [49]. Another example is the bHLH family 1, where one of the most highly correlated TFs is *Bach2*, which is required for the maturation of B and T cells [50].

Identification of distal regulatory motifs in mouse cerebral cortex

We used scover to infer regulatory motifs from open chromatin by analyzing data from mouse cerebral cortex collected by Chen et al using SNARE-seq [7]. The cortex is organized in layers with distinct functions and we grouped cell types based on these categories. The dataset contains cells from both postnatal day 0 (P0) and adult, and we processed both time points after running the analyses separately. Since we were interested in distal enhancers, only 174,233 loci in P0 and 189,528 loci in adult that were outside the [-8 kb, 2kb] region relative to annotated TSSs were considered.

For P0 we found nine motif clusters and for adult we found ten, with six found for both time points (**Fig 6a, S8**). For P0 21.1% and for adult 15.7% of convolutional filters matched a known motif, and we again found that many of the non-matching filters correspond to partial matches of known motifs, e.g. Mef2c and Nfib (**Fig S9**). The fraction of variance explained by the model is smaller than for the scRNA-seq data with 13% for P0 and 8% for adult. As expected, the motifs identified had lower CpG-content than the ones found at promoters (**Fig S10**). Most motifs are infrequent, with the most common ones found in 15% of peaks and the plurality in <5%, suggesting a greater diversity than for promoters. As for the scRNA-seq data, there is one group of motif clusters with high influence scores (Ctcf, bHLH, KLF, Mef2c and Rfx), while the weaker clusters have 1-2 orders of magnitude lower scores (**Fig S11-14**).

Several of these findings, e.g. higher weights for Nfib in the lower layers and higher weights for Mef2c in the middle layers, are consistent with previous results based on bulk ATAC-seq from sorted populations [51]. The most notable finding is the Hlf motif (**Fig 6b**) which was not identified by Gray et al. Reports linking misregulation of Hlf in the dentate gyrus to epilepsy shows that the gene has important functions in the brain, suggesting that it could be important in the visual cortex as well [52,53]. Interestingly, like many other motifs, Hlf has non-uniform weights across layers (**Fig 6c**). *Hlf* expression has been reported to be increasing through development [54], consistent with the fact that we find the motif in adult but not in P0, and also with the lower expression at the earlier time point (**Fig 6d**).

During development the mouse cerebral cortex undergoes changes in gene expression [55,56]. We hypothesized that there would be concomitant changes in the distal regulatory profiles across layers. Analyzing open chromatin loci in different layers, we found that only ~25% of distal sites found in one layer are also found in all other layers, and there is little similarity between cells from the same cortical layer profiled at P0 and adult (**Fig S15**). When we consider the influence scores for the six motifs that were identified at both time points, we find greater variability in the adult compared to P0 (**Fig 6e**). These findings indicate substantial changes in the impact of regulatory motifs as the mouse cerebral cortex develops. Since the SNARE-seq protocol also profiles the transcriptome for the same cells, we can directly compare the expression levels of the relevant TFs. Consistent with our hypothesis that differences in the distal regulation between excitatory neurons from different

layers increase over time, we also find that the correlations between the motif influence scores and TF expression levels increase for *Mef2c*, *Rfx* and *Tcf4* (**Fig 6f**).

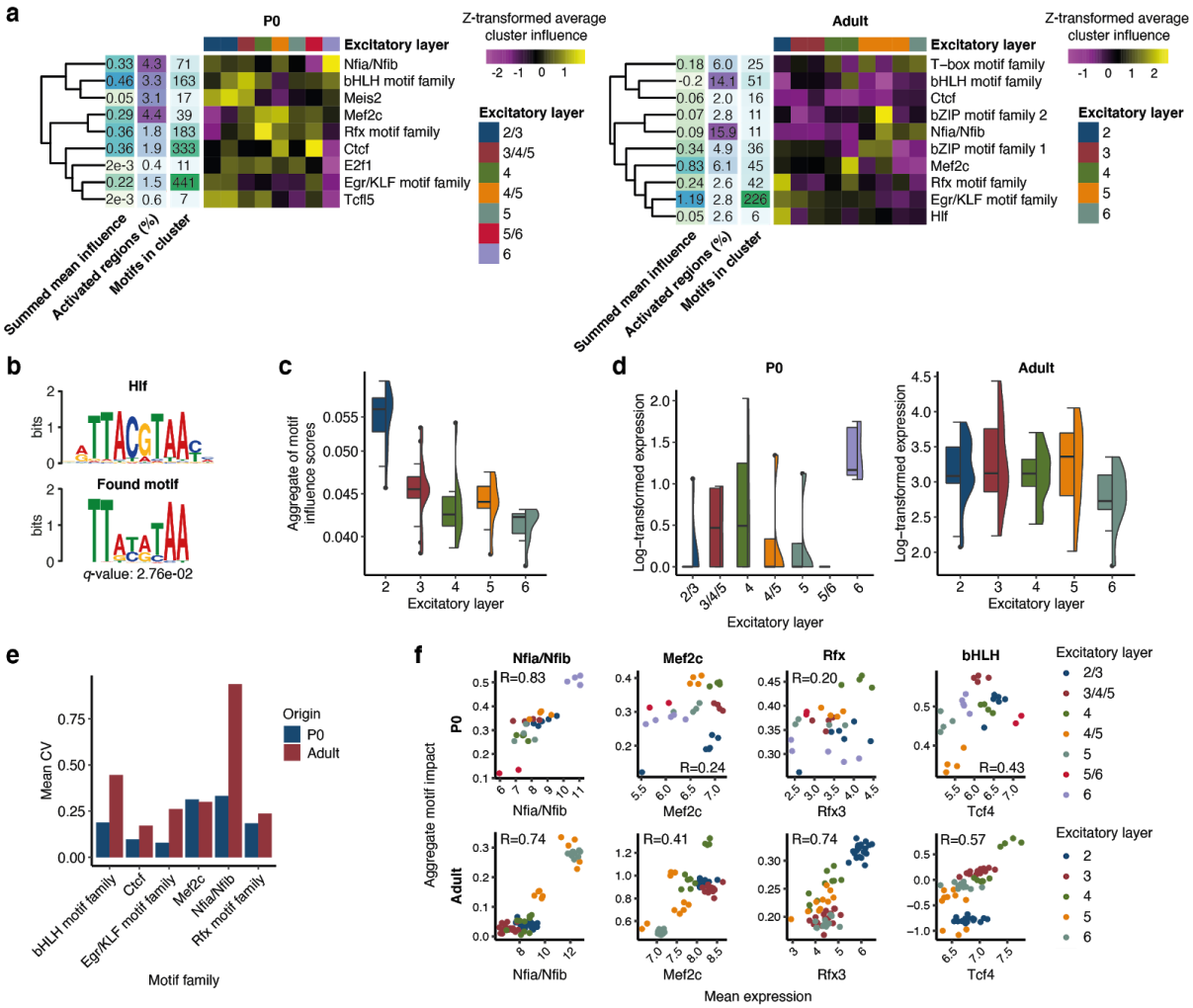

Figure 6: Analysis of distal open chromatin peaks from mouse cerebral cortex. (*a*) Z-transformed average neural network influence score per motif cluster in P0 and adult. (*b*) Aggregate motif influence scores for the Hlf motif cluster in the different layers for adult. (*c*) Expression values of Hlf across excitatory neuron pools for P0 and adult. (*d*) An example of a found motif aligned to the database Hlf motif. (*e*) Mean coefficient of variation calculated across pools for motif impacts across excitatory neurons for motifs found in both time points. (*f*) Correlations between aggregate motif family influence scores in pools and expression for selected TFs. The Spearman correlation is reported at the top of each panel.

Discussion

We have presented scover, a computational method based on a neural network with a single convolutional layer for inferring regulatory motifs from single cell data. Using a single GPU, scover can process tens of thousands of cells in <10 hours, making it suitable for currently available datasets. Previous attempts to develop quantitative models of regulatory motifs have been restricted to tissue level or cell lines [6,13], or to immune cells [24,42], and they frequently use histone marks rather than sequence to predict expression levels [57]. We have

for the first time shown that it is possible to simultaneously infer regulatory motifs and quantify their impact at the resolution of individual cell types.

Since scover uses a *de novo* strategy for finding motifs we expect that it will be especially helpful for researchers studying organisms where regulatory motifs are poorly annotated. For mouse and human non-matching motifs make a small contribution, and they are most likely a consequence of a failure to converge, although we cannot rule out the possibility that some of them are picking out CpG dinucleotides [6]. Our results are also consistent with other studies that have categorized TFs as either preferentially binding promoters or enhancers [42]. One such example is HNF4A, which is an important TF in the kidney [58], but because of its preference to distal enhancers it was not identified by scover [59].

Interestingly, there are several trends that are common to the kidney and the Tabula Muris datasets. Immune cells have regulatory profiles distinct from other cell types and they have the highest influence for YY1 and ETS motifs, while KLF and ZFX have low impact (**Fig 3a, 5a**). More generally, for the six motif clusters (bHLH, ETS 1 & 2, KLF, YY1, ZFX) that are found in both datasets, the influence scores are in good agreement with the highest impact for ETS motifs, followed by YY1 and KLF. Similarly, we find that KLF is the most widely used motif. The ETS family was heterogeneous in both datasets, consistent with previous categorizations [60]. Notable differences are that the Tabula Muris has a greater range of influence scores, for most motifs the extremes differ more than two-fold, while in the kidney the difference is less than 30%, and that the quality of the fit provided by the Tabula Muris is substantially better. Since the two datasets have similar number of cells and similar sparsity after pooling, we conjecture that the differences are due to the greater diversity of cell types in the Tabula Muris and the use of different experimental platforms.

Although larger and more deeply sequenced datasets could allow for the discovery of additional motif clusters and better model fits, we believe that the room for improvement is limited. First, studies of gene expression noise have estimated that some of the variability is inevitable due to random fluctuations [61]. Second, there are many other mechanisms of gene regulation that are not included in sequence data, e.g. TF concentrations, distal enhancers, histones, miRNAs, and RNA binding proteins. Nevertheless, our results are consistent with a recent study of mouse hematopoietic cells which used epigenetic data to estimate that ~50% of the variance in expression can be explained by the endogenous promoter [57].

Methods

Pooling

To reduce the sparsity, pools are created by grouping cells through sampling without replacement into *p* equally sized groups of *q* cells. Only cells belonging to the same cell type can be assigned the same pool and for each pool the expression values are the sum of the expression values of the single cells in that pool. Cell types that contain fewer than *q* cells are omitted.

One-hot encoding of nucleotide sequences

For the pooled transcriptomic datasets the sequences correspond to 500 bp upstream and downstream of gene TSS. For the Tabula Muris and SNARE-seq datasets, sequences were isolated from the mm10 reference genome using the Ensembl v.96 annotations [62]. For the human kidney dataset, sequences were isolated from the hg38 reference genome using the GENCODE v.33 gene annotations [63].

For the pooled accessibility datasets, each sequence corresponded to 250 bp around the centre of the peak region. Input nucleotide sequences are one-hot encoded into a 2D matrix such that the columns correspond to [A, C, G, T], respectively. For each position in the sequence, the corresponding row in the matrix will have a 1 at the column corresponding to the original nucleotide and a 0 otherwise.

Convolutional neural network architecture

Scover is a convolutional neural network composed of a convolutional layer, a rectified linear unit (ReLU) activation layer, a global maximum pooling layer, and a fully connected layer with multiple output channels. The convolutional layer takes as input the one-hot encoded DNA sequences, and the fully connected layer outputs predictions across the pools. Sequentially, the neural network carries out the following operations:

1. 2D convolution without bias values with *d* convolutional filters of size (*m*, 4)
2. A ReLU activation layer that sets negative values to 0
3. A global maximum pooling layer that takes the maximum value for each of *d* rectified convolutional operations
4. A fully connected layer with bias values with *d* inputs and *p* outputs

The number of output pools *p* depends on the number of cells per pool *q* and the sizes of the clusters. The default values are $d = 300$, $m = 12$ and $q = 100$. However, it is recommended that the user explore different combinations (please see GitHub tutorial for an example).

Training the neural network

The neural network is trained in two stages: a random search for hyperparameter optimization, and a training stage using the best hyperparameters.

In the first stage, the hyperparameters are optimized through random search [64,65]. The dataset is divided into *K* folds (default: 10) and the neural network parameters are initialised

num_calibrations (default: 30) times. The neural network parameters are initialised as follows: the learning rate is randomly sampled from a log uniform distribution over the interval [*epsilon_min, epsilon_max*] (default: [5e-4, 5e-2]). The convolutional filter weights are initialized with values drawn from a normal distribution with mean 0 and standard deviation $sigma_motifs^2$, where *sigma_motifs* is randomly sampled from a log uniform distribution over the interval [*sigma_motifs_min, sigma_motifs_max*] (default: [1e-7, 1e-3]). Similarly, the fully connected layer weights are initialized with values drawn from a normal distribution with mean 0 and standard deviation $sigma_net^2$, where *sigma_net* is drawn from a log uniform distribution over the interval [*sigma_net_min, sigma_net_max*] (default: [1e-5, 1e-2]). The fully connected layer biases are initialized with the value 1e-5.

Each network is trained on the *K* different folds (default: 10) for a number of epochs (default: 24). For each fold the lowest validation set errors are calculated. The optimal network initialization is chosen as the network initialization with the lowest mean error for the validation set.

During the training stage, the network is initialized using the best initial parameters (the learning rate, *sigma_motifs*, and *sigma_net*). The training data is split into *num_candidates* new folds (default: 10). For each fold, the network is trained for the same amount of epochs as before, and network parameters for each fold are saved at the time point of lowest validation set error. The network uses a mean squared error loss function, and the network parameters are optimized using the stochastic gradient descent optimizer with a default batch size of 128.

To ensure that the motifs identified by scover capture biologically meaningful signals, we carry out a control experiment where we train scover using a dataset where the nucleotide order of the input sequences is permuted. Small differences in correlation between the predicted and observed values for the real and permuted data is a strong indication that the identified motifs are spurious matches and that scover has failed to identify an informative model. For all of the datasets considered in this manuscript, the difference in the squared Pearson correlation using the real and permuted data was 0.136 for the kidney dataset, 0.337 for the Tabula Muris dataset, 0.13 for the p0 SNARE-seq dataset, and 0.078 for the adult SNARE-seq dataset.

Alignment of motifs

To extract position weight matrices corresponding to the convolutional filters, rectified activations are calculated for each model by convolving the convolutional filters across the training set sequences and passing the output through a ReLU layer. Consequently, for each sequence the subsequence of length *m* that generates the highest positive activation is isolated for each convolutional filter. Sequences that do not generate a positive activation score for a given convolutional filter are not included. For each convolutional filter, the scores are converted to position frequency matrices (PFMs) of size 4x*m* by summing the occurrences of each nucleotide at the corresponding sequence position. The PFMs are converted to position probability matrices (PPMs) by dividing each entry in the PFM by the sum of the nucleotide frequencies for the corresponding sequence position. The concatenated PPMs of the 10 models are then stored in a single MEME-formatted file. The

motifs are aligned to a MEME-formatted motif database file using the Tomtom tool from the MEME suite [66] using the argument -thresh 0.05. Sequence logos of motifs that did not align were visualized using R package *ggseqlogo* [67].

Leave-one-out analysis and influence scores

After training *r* models, each model is reinitialized using the parameters from the optimal training point (the point at which the MSE on the validation set was the lowest). The prediction for motif *i* in pool *j* and sequence *k* is denoted b_{ijk}. We then set all parameters related to motif *i* to zero and calculate the perturbed score c_{ijk} for each pool. The influence score for motif *i* in pool *j* is calculated as $n_{ij} = \frac{1}{N}\sum_{k=1}^{N}(b_{ijk} - c_{ijk})$ where *N* is the number of sequences in the validation set. The influence scores from *d* motifs from *K* models are concatenated to get a matrix with *d*K* rows and *p* columns. For each motif, the mean influence per cell type is calculated as an average from the corresponding pools. For visualization purposes we also calculate a z-transformed version (based on cell types).

Using the R package *igraph* [68], the output of the Tomtom alignments is used to generate a bipartite graph, with the learned motif PPMs and corresponding database alignments as two disjoint sets. Motif clusters are defined using the *cluster_walktrap()* function from *igraph* with default parameters. To ensure reproducibility, motif clusters are excluded if they contain motifs from less than half of the *K* models. Next, global motif influence scores are calculated as $\sum|n_{ij}|$. Using the R package *mixtools* [69], a Gaussian mixture model with two components is fitted to the global motif influence scores. If all of the motifs can be assigned with a probability greater than 0.6 to one of the two resulting clusters, the motif cluster with lower mean global influence score is filtered out. Lastly, for most of the downstream analysis, motifs that did not align to known motifs are excluded. In some cases, larger motif clusters were subclustered based on hierarchical clustering of the euclidean distances between z-transformed mean motif influence scores across cell types.

Influence scores for motifs that belong to the same motif family were summed to generate an 'aggregate motif family influence' matrix with *num_motif_families* rows and *p* columns. Similar to before, this matrix is then converted to include mean aggregate motif family influence scores across cell types or cell type categories, and for visualization purposes, these scores are z-transformed.

Proliferation and proximal tubule marker analysis

We used the following genes as markers of cell proliferation: MKI67, PLK1, E2F1, FOXM1, MCM2, MCM7, BUB1, CCNE1, CCND1, CCNB1, and TOP2A [70]. The proliferation score of each pool was calculated as the average expression of this set of genes in that pool. Similarly, we used DLL1, ACAT1, PKD1, NOTCH2, AQP11 and HEYL for the proximal tubule signature genes.

Gene Ontology analysis

The GO terms for mouse were downloaded from Ensembl Biomart. We used all categories with fewer than 50 genes for the analysis. For each GO term, the correlations between the expression of the GO term genes (across pools) and the aggregate motif influences of each

motif family (across pools) were calculated. The top 500 GO terms in terms of their maximum correlation were hierarchically clustered based on euclidean distance, after which the tree was cut with $k = 2$ to form the clusters.

Pseudotime analysis

The developing nephron compartment of the fetal kidney dataset was subset to cells with a developmental relationship to cap mesenchyme (excluding ureteric bud). A two dimensional UMAP embedding was calculated in scanpy [71], and was clustered using k-means clustering (k=10). Pseudotime trajectories were computed with Slingshot [72] using the cluster corresponding to cap mesenchyme as a starting cluster. Proximal tubule pseudotime values were generated for cells along a continuous trajectory path from cap mesenchyme to proximal tubule.

The pseudotime values were averaged across the same cells that generated the pools in the combined human kidney dataset. In some cases, not all the cells in the pool had corresponding pseudotime values, as some cells fell outside of the k-means clusters along the principal curve corresponding to the proximal tubule pseudotime trajectory. The pseudotime values were only calculated for pools that had at least 100 cells (out of 120 cells per pool) with pseudotime values. The pooled (averaged) pseudotime values were subsequently normalized to range [0,1].

Code availability

The neural network is implemented in pytorch and is available at https://github.com/jacobhepkema/scover. The entire workflow is implemented as a Nextflow pipeline [73], and data generation and downstream analysis scripts are implemented in R. Documentation is available at https://scover.readthedocs.io/. Code that was used to generate the figures is available at https://github.com/jacobhepkema/scover_figures.

Datasets used

Human kidney dataset

We downloaded the data from https://www.kidneycellatlas.org/. The fetal and adult kidney datasets were concatenated in the cell dimension to create a matrix of 33660 features by 67,465 cells. After pooling the dataset with q=120 cells per pool, the data was filtered by excluding genes that were expressed in <2% of pools. This resulted in a dataset of 18,067 genes by 529 pools. Training the network with parameters $d = 600$, $m = 12$ and aligning the motifs to the database motifs took 6 hours and 20 minutes using a NVIDIA Tesla V100 GPU with 32GB memory.

Tabula Muris dataset

We downloaded 18 Tabula Muris FACS-sorted Smart-Seq2 datasets from https://doi.org/10.6084/m9.figshare.5829687.v8. The cells were concatenated in the cell dimension to create a matrix of 23,341 genes by 41,965 cells. After pooling the dataset with $q = 80$ cells per pool, the data was filtered by removing genes that were expressed in <2% of pools, resulting in a pooled dataset of 19180 genes by 476 pools. When training the model we used $d = 600$ and $m = 12$.

Batch correction
For both the combined human kidney and Tabula Muris datasets, many cell types from the individual datasets did not overlap. Moreover, it has been shown that merging datasets can distort expression profiles [74]. For these reasons, batch-correction methods were not used on the concatenated datasets, as this might remove biological signal in addition to batch effects.

SNARE-seq dataset
We downloaded the SNARE-seq data from the Gene Expression Omnibus database under accession number GSE126074. The P0 and adult datasets were both pooled with q = 100 cells per pool. Both datasets were filtered by excluding peaks that were present in <2% of pools. To ensure that only distal peaks were included we removed any peak found in the [-8kb, 2kb] region relative to any TSS found in the Ensembl annotations for mm10 (v.96). Furthermore, we excluded all non-neuronal cell types from the analysis, but we included the inhibitory neurons when fitting the model even though only the excitatory neurons were used for the downstream analyses (**Table S4**). Region overlap from the different time points and excitatory neuron layers was plotted using R package *UpSetR* [75]. When training scover we used d = 600 and m = 12 for P0, and d = 300 and m = 12 for adult.

Acknowledgements

We would like to thank Song Chen and Yingxi Lin for helpful discussions about the SNARE-seq data and Irina Abnizova, Ilias Georgakopolous-Soares, Leopold Parts and Nikos Patikas for comments and suggestions on the manuscript and the package. We are also grateful to Simon Murray for testing the repository and to Jimmy Lee for sharing code. JH was supported by a grant for “Search tools for scRNA-seq data” (2018-183503) from the Chan Zuckerberg Initiative. MH and NKL were funded by a core grant from the Wellcome Trust. MH, NKL, SR and VC would like to acknowledge funding from a Newton Mobility Grant (NI160206) from the Royal Society and the Thai Office of the Higher Education Commission.

Author contributions

The project was conceived by JH, NKL and MH. JH, NKL and SR wrote the code. JH and MH analyzed the data. BJS and MRC assisted with the analysis of the kidney data. VC, MRC and MH supervised the research. JH and MH wrote the manuscript with input from the other authors.

Conflicts of interest

There are no conflicts of interest

Supplementary figures

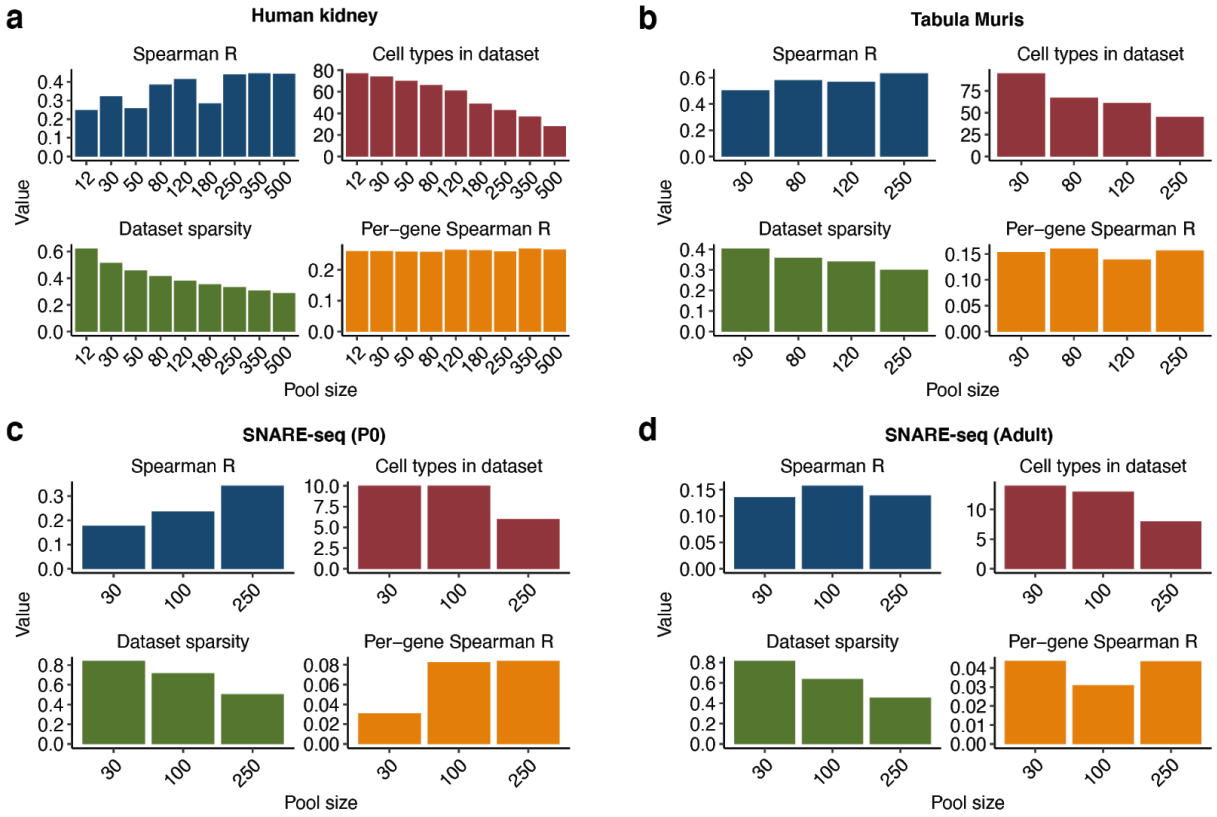

Figure S1: Correlations, sparsity (fraction of genes that are zero) and number of cell types for different pool sizes for the datasets considered in this manuscript.

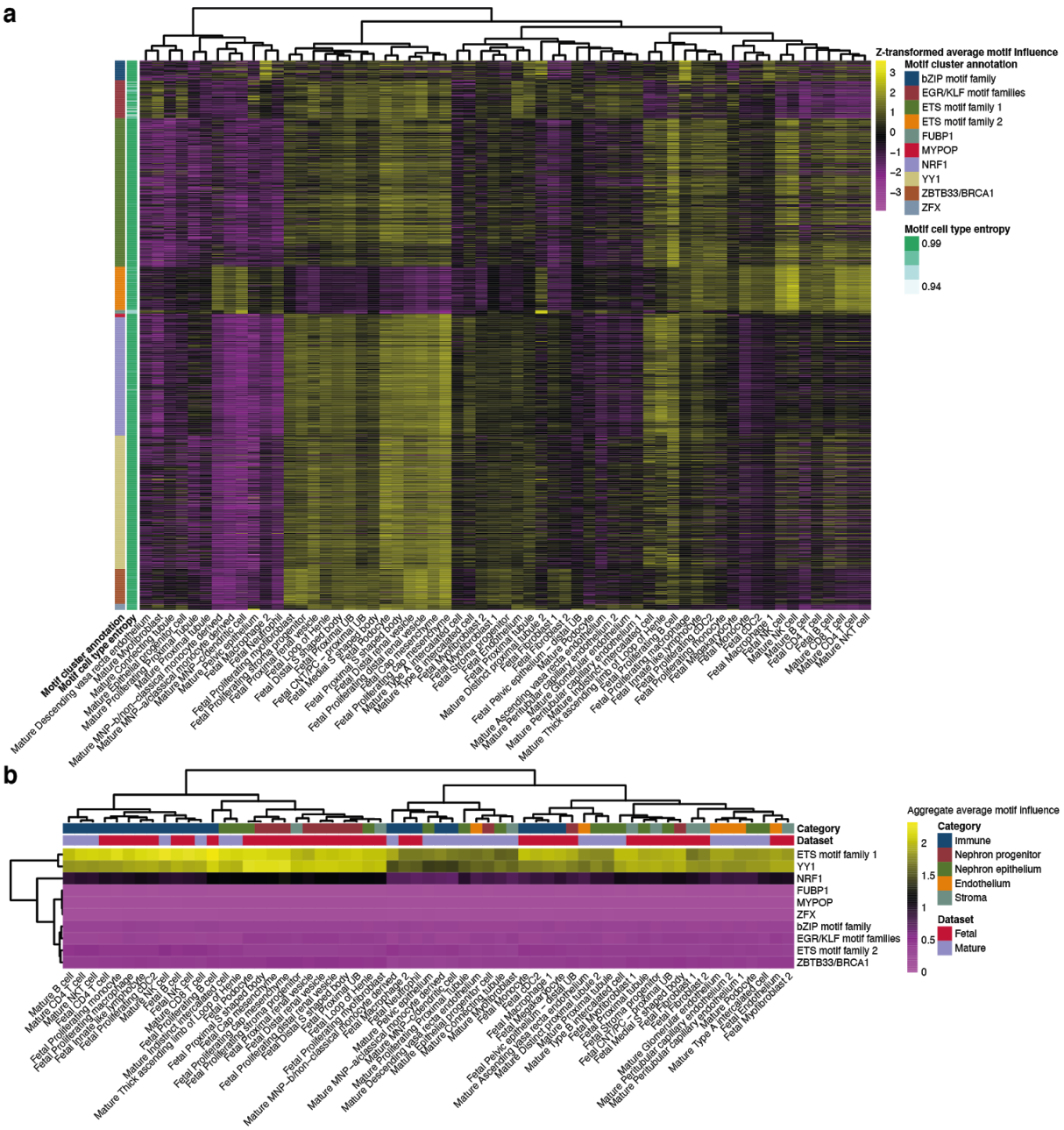

Figure S2: (*a*) *Z*-transformed average influence scores of individual motifs across human kidney cell types. (*b*) Aggregate average motif influence scores of motif clusters across human kidney cell types.

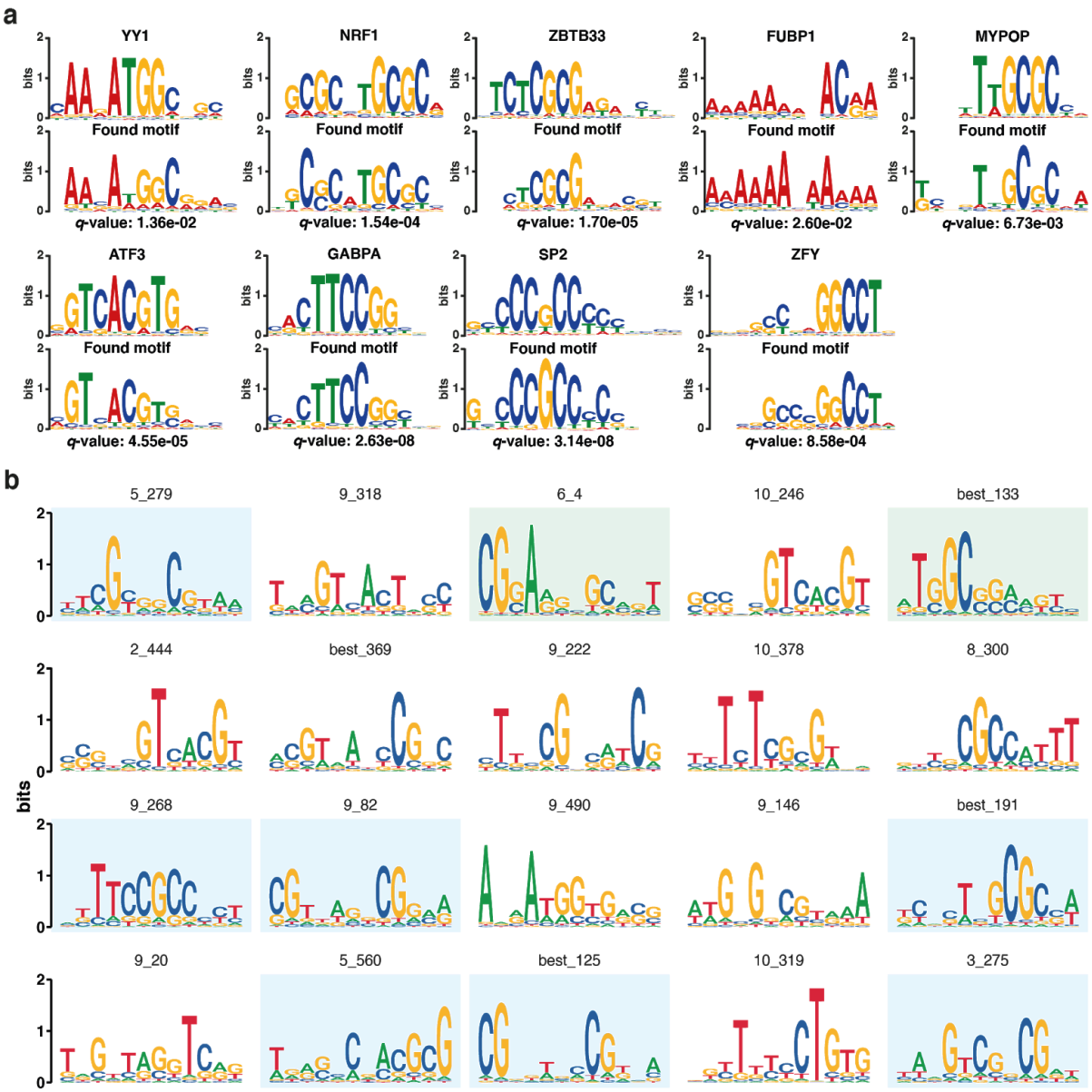

Figure S3: (*a*) Example motif alignments for the human kidney dataset. (*b*) Random selection of 20 motifs with high impact scores that do not significantly align to the CIS-BP database. Several of the motifs have high GC-content (marked blue) or include a GGAAG pattern reminiscent of ETS (marked green).

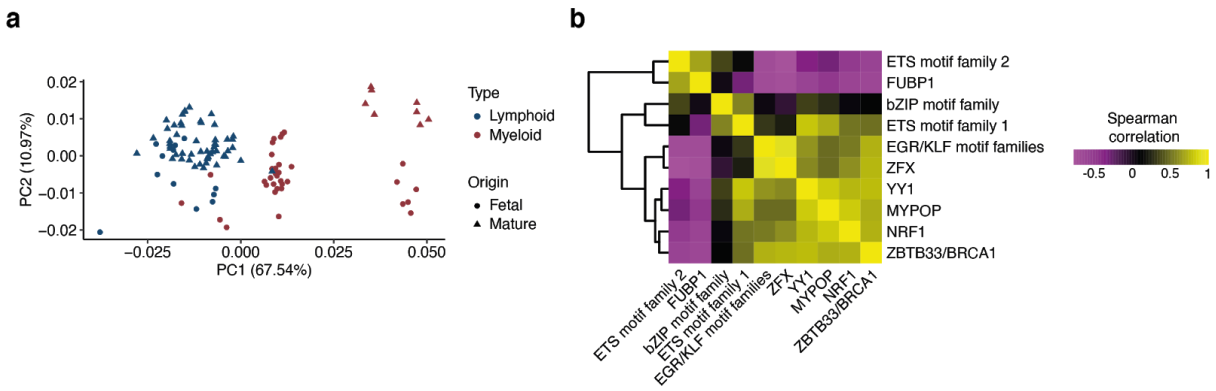

Figure S4: (*a*) Projection of the immune cell type pools onto the space represented by the first two principal components of the influence score matrix for the human kidney dataset. (*b*) Pairwise spearman correlations between motif cluster influence scores for the human kidney dataset.

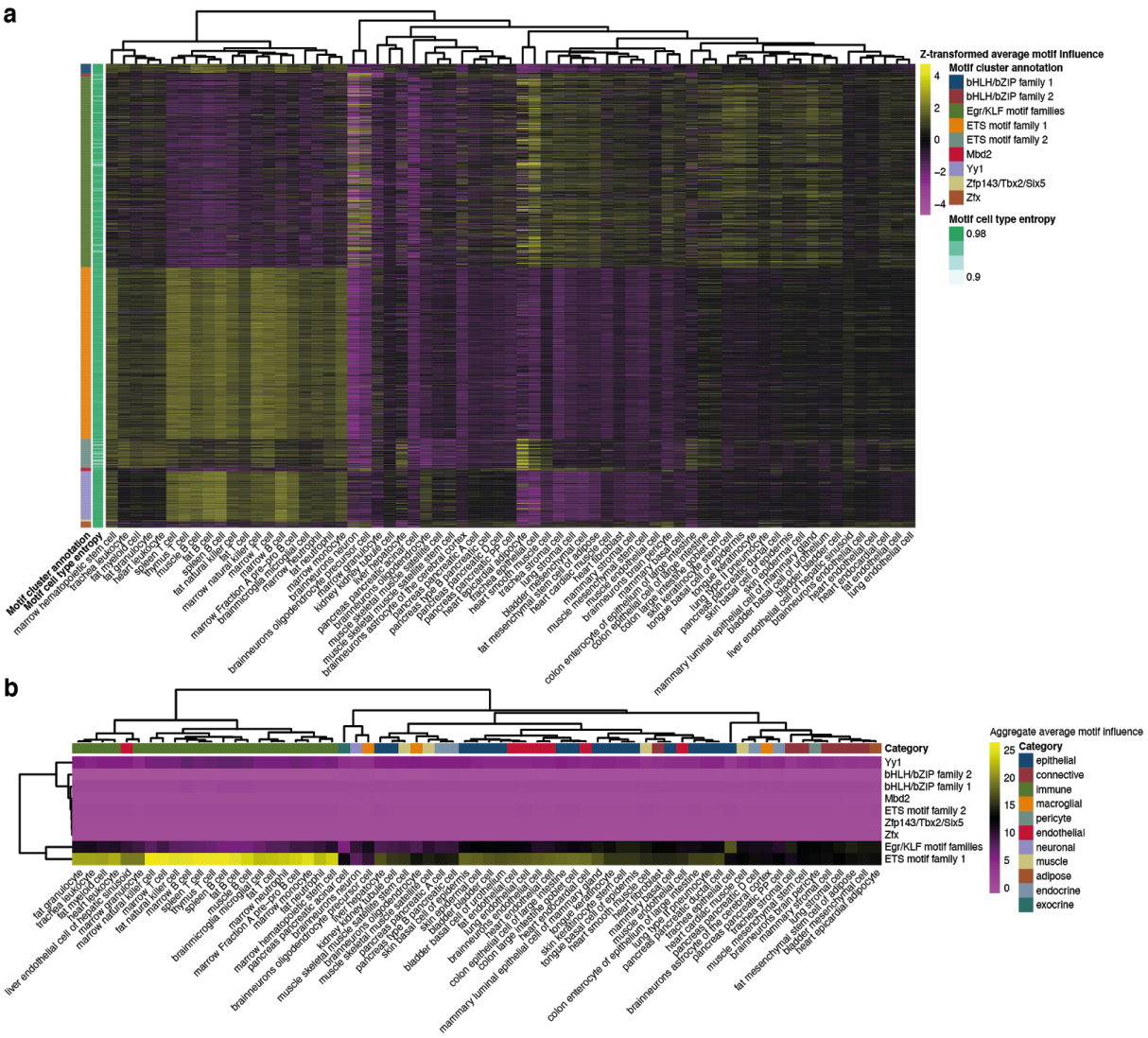

Figure S5: (*a*) Z-transformed average influence scores of individual motifs across Tabula Muris cell types. (*b*) Aggregate average motif weights of motif clusters across Tabula Muris cell types.

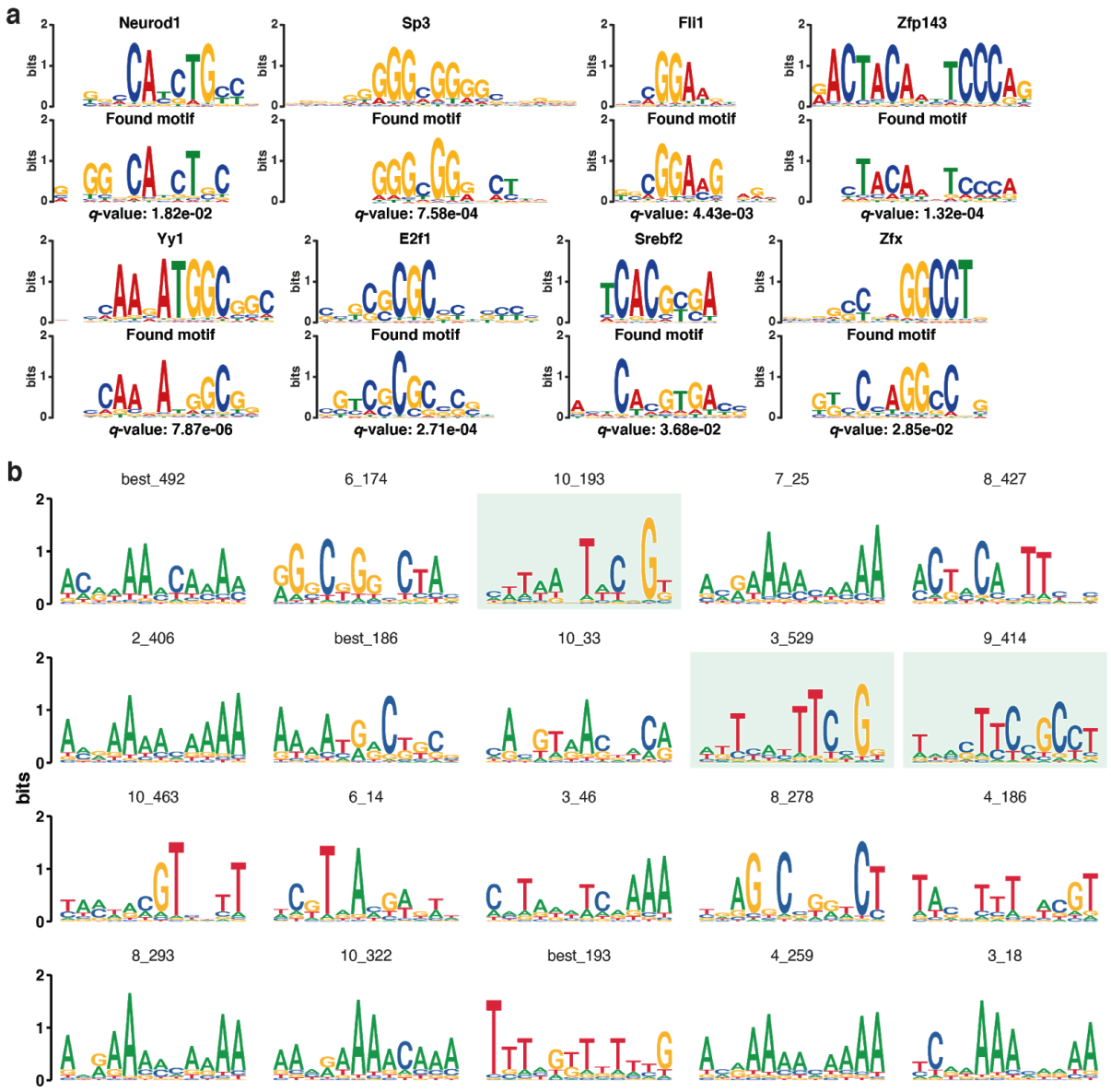

Figure S6: (*a*) Example motif alignments for the Tabula Muris dataset. (*b*) Examples of randomly selected motifs with high influence scores that did not align to CIS-BP. Some of the motifs resemble the ETS family with a TTCNG pattern (marked in green).

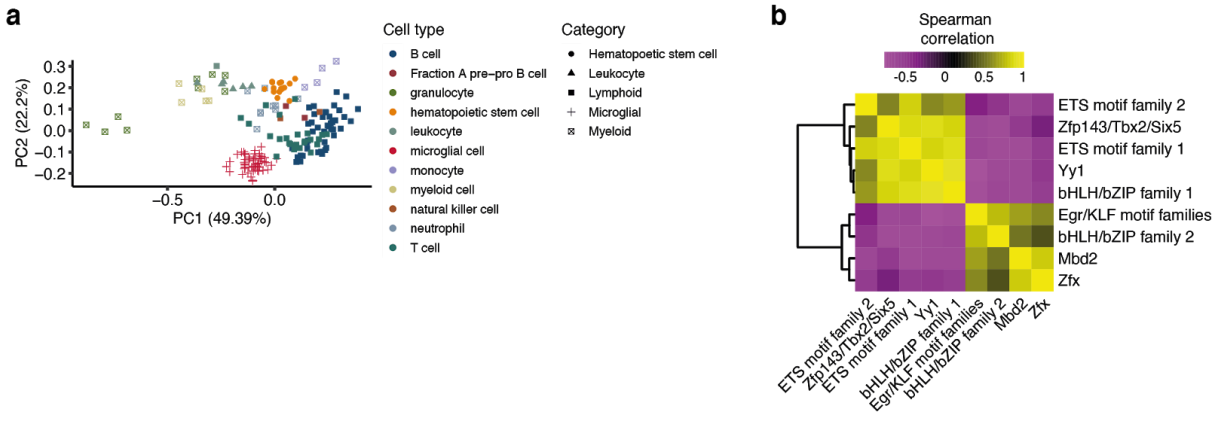

Figure S7: (*a*) Projection of the immune cell type pools onto the space represented by the first two principal components of the influence score matrix for Tabula Muris. (*b*) Spearman correlations between motif family influence scores for Tabula Muris.

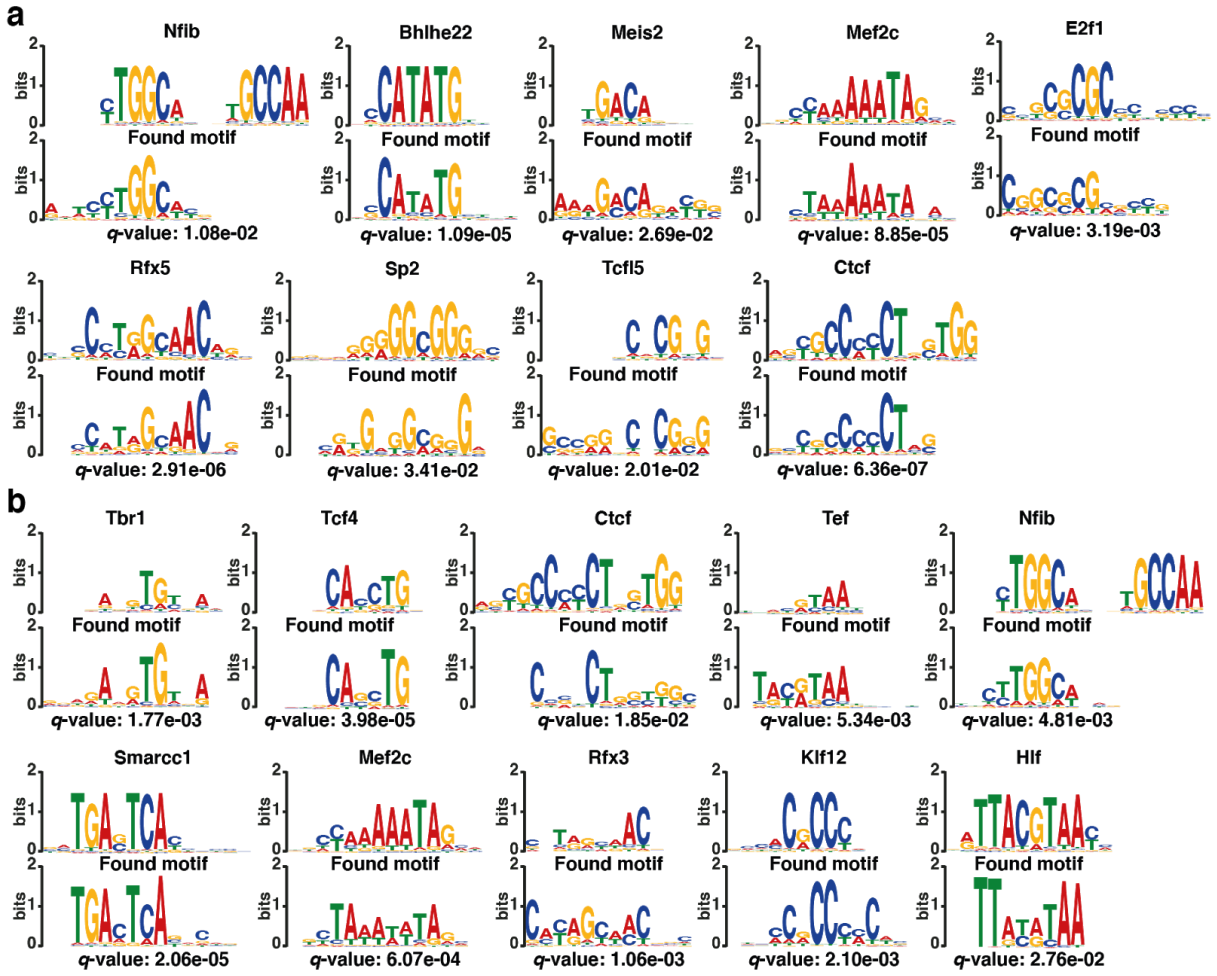

Figure S8: Example motif alignments for the SNARE-seq (*a*) P0 and (*b*) adult datasets.

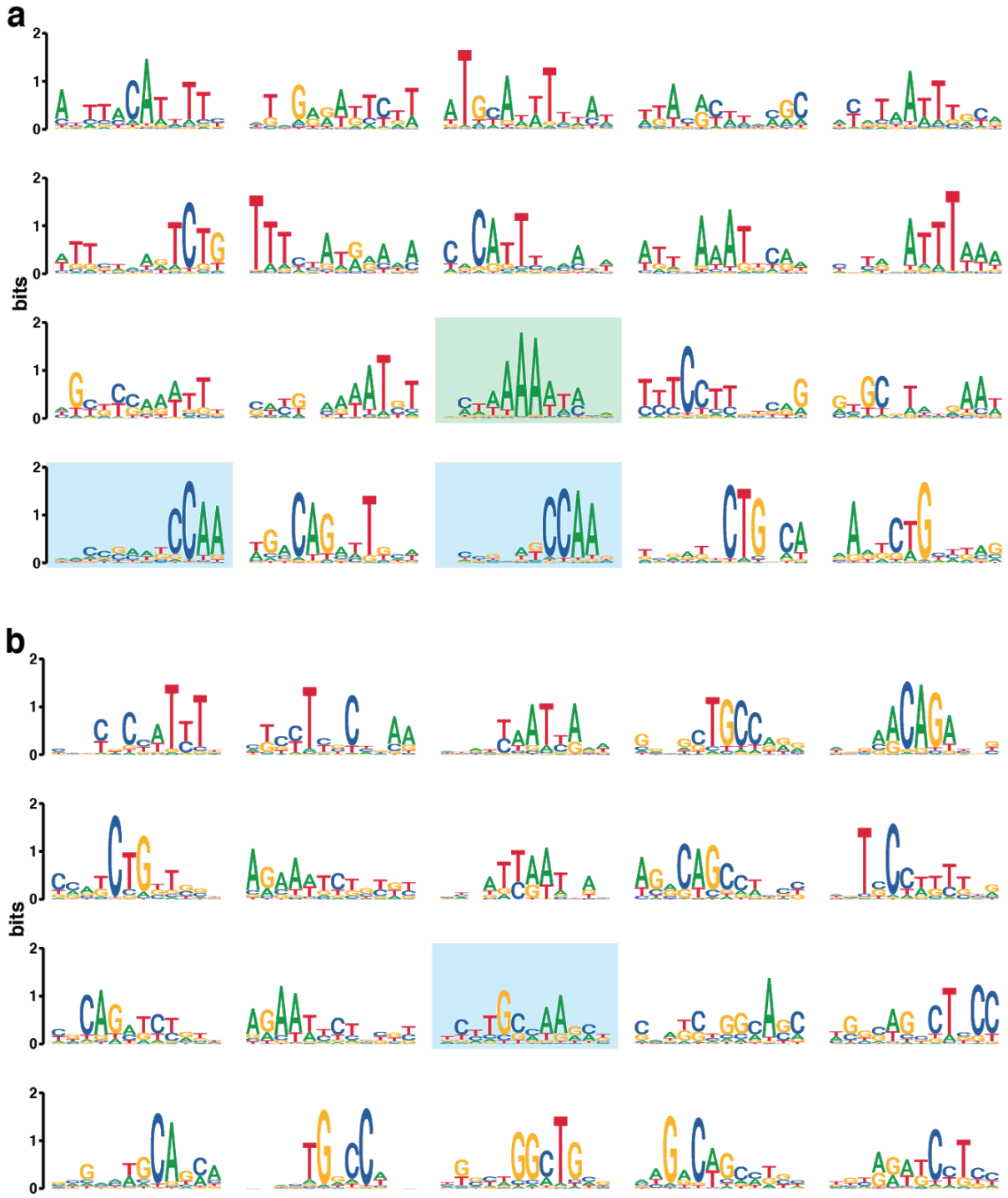

Figure S9: (*a*) Examples of randomly selected motifs with high influence scores that did not align to CIS-BP for the P0 dataset from SNARE-seq. Some motifs are similar to Mef2c (marked green) and Nfib (marked blue). (*b*) Same for the adult dataset. One motif is similar to Nfib (blue shade).

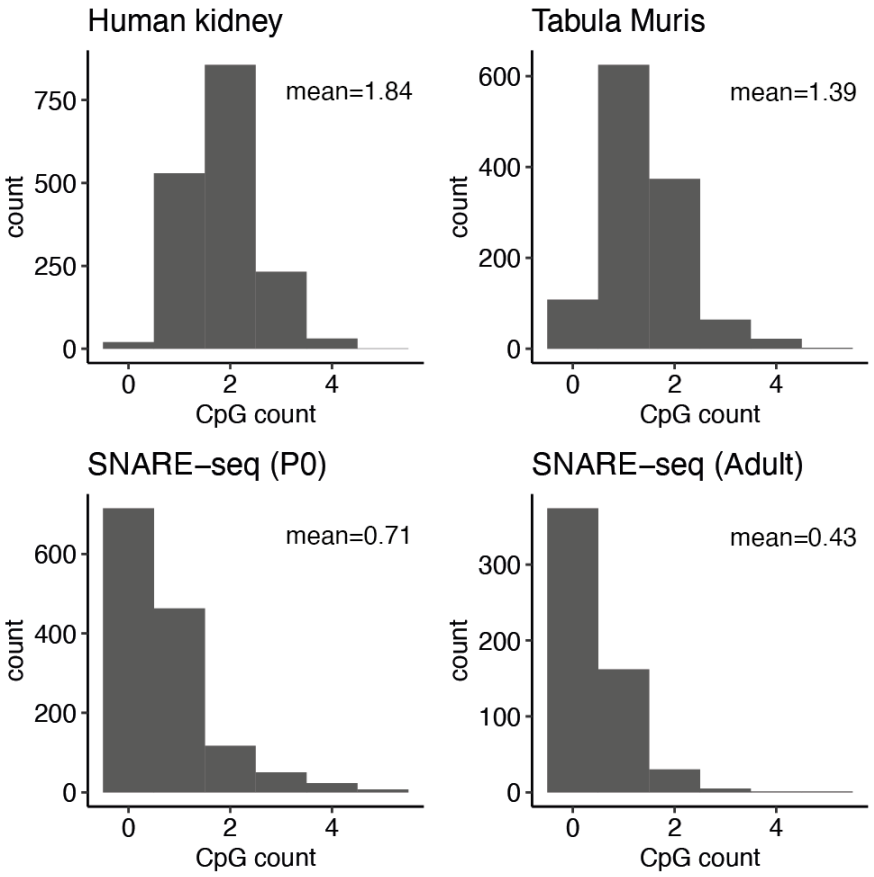

Figure S10: CpG count for the motifs found using the different datasets.

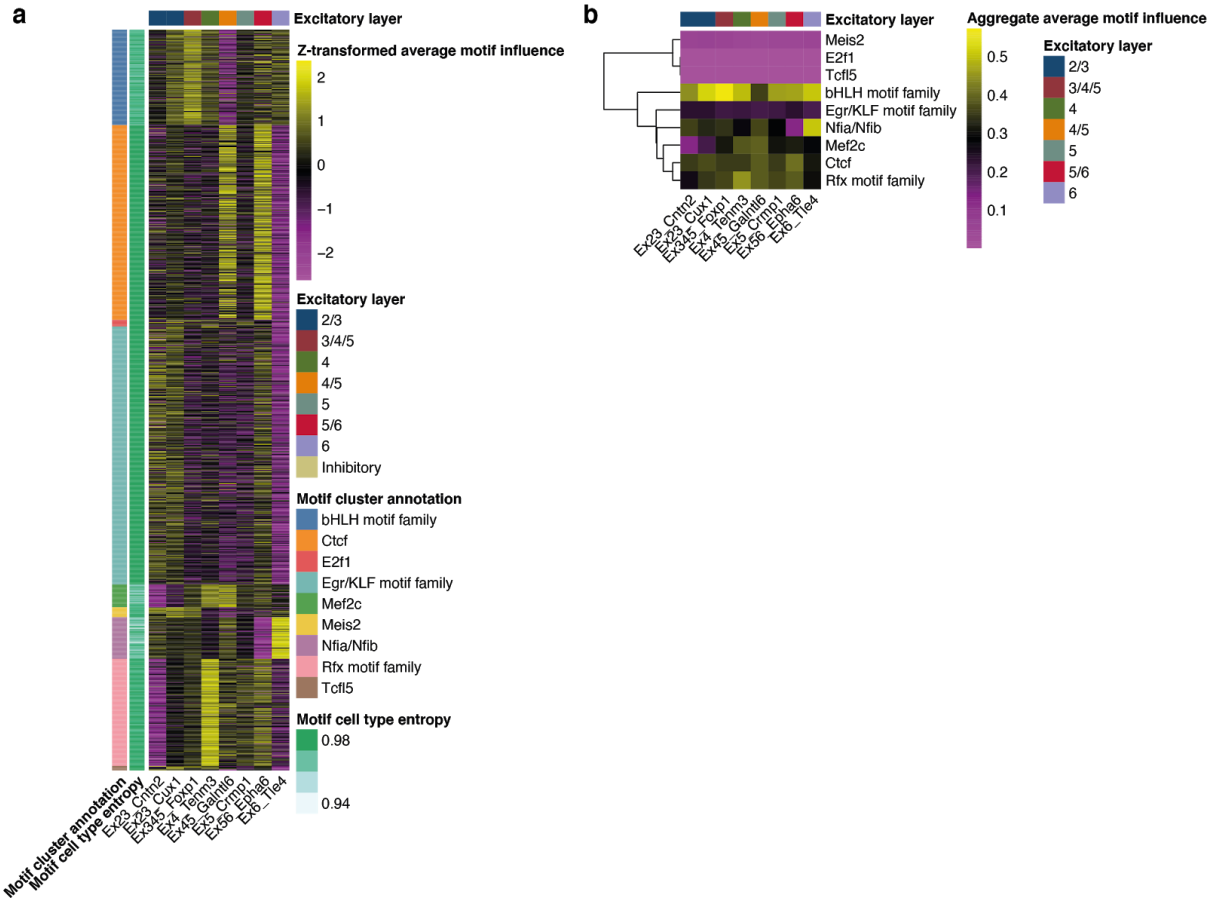

Figure S11: (*a*) Z-transformed average influence scores of individual motifs across excitatory cell types for the SNARE-seq P0 dataset. (*b*) Aggregate average motif weights of motif clusters across excitatory cell types for the SNARE-seq P0 dataset.

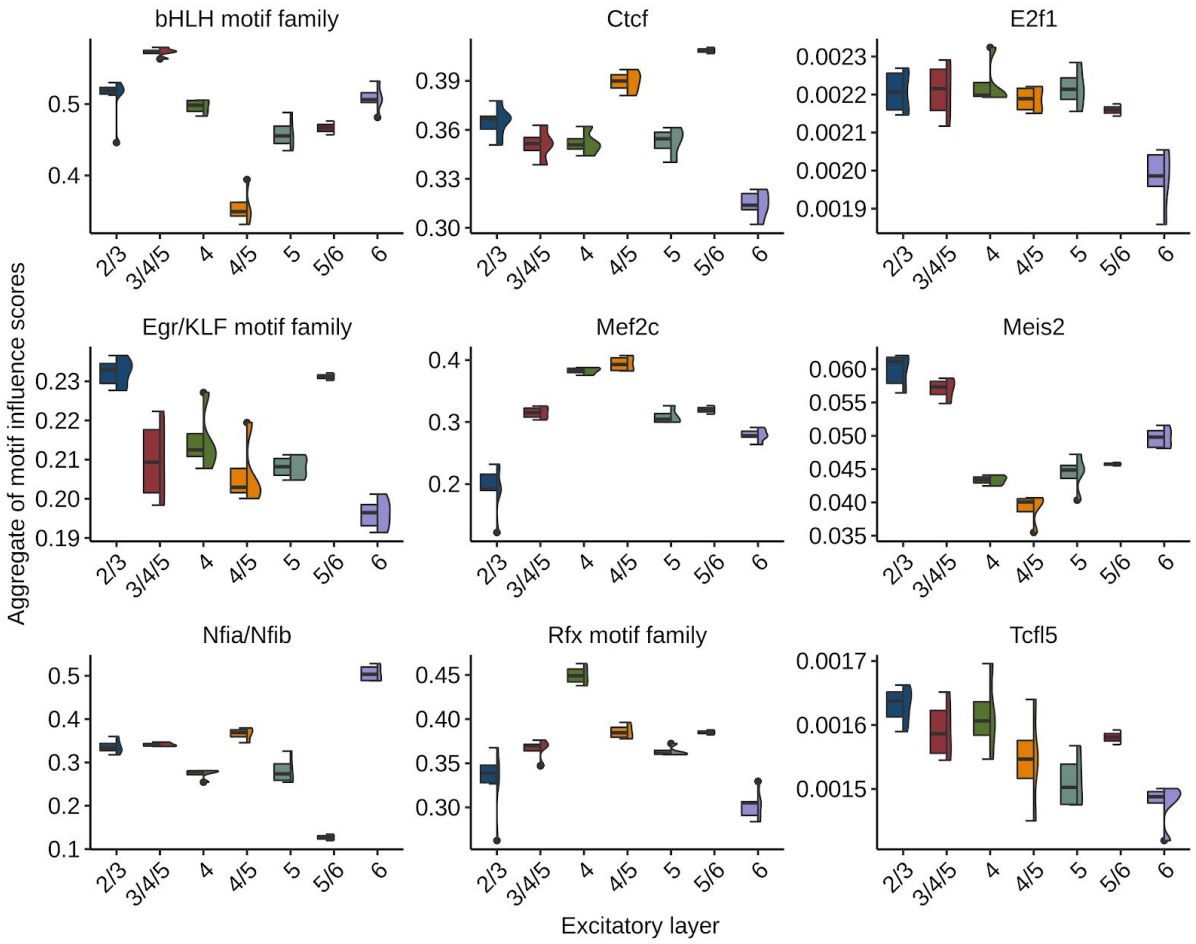

Figure S12: Aggregate motif weights in nine clusters across the excitatory layers from the SNARE-seq P0 dataset.

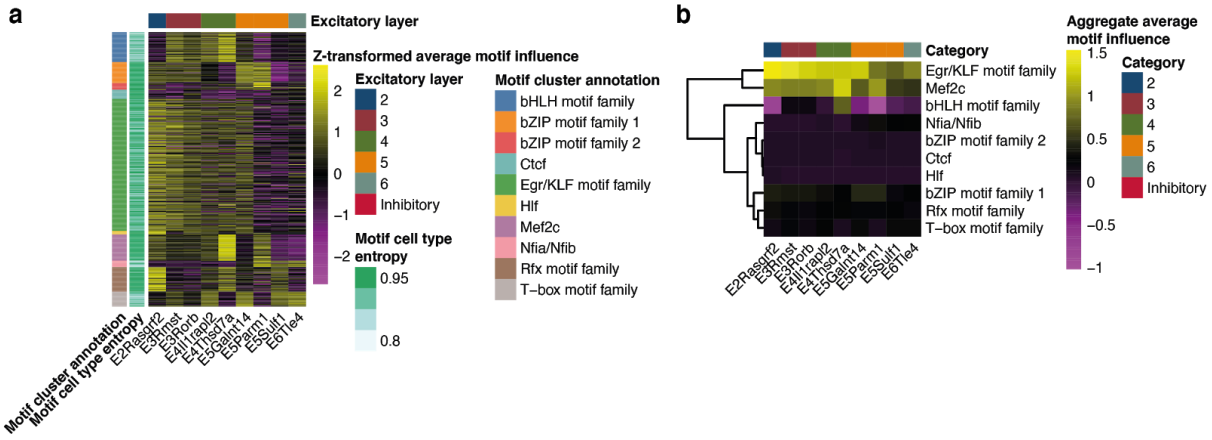

Figure S13: (*a*) *Z*-transformed average influence scores of individual motifs across excitatory cell types for the SNARE-seq adult dataset. (*b*) Aggregate average motif weights of motif clusters across excitatory cell types for the SNARE-seq adult dataset.

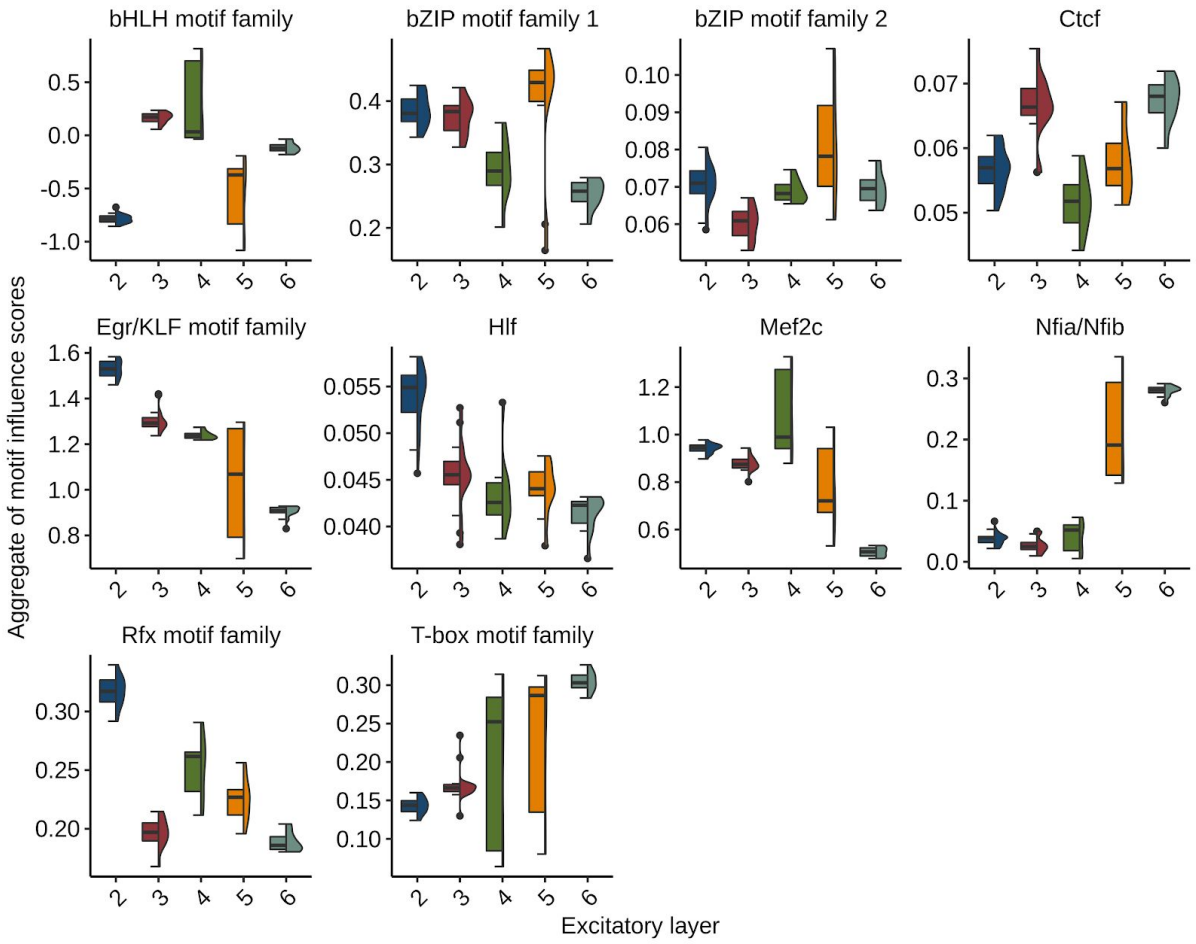

Figure S14: Aggregate motif weights in ten clusters across the excitatory layers from the SNARE-seq adult dataset.

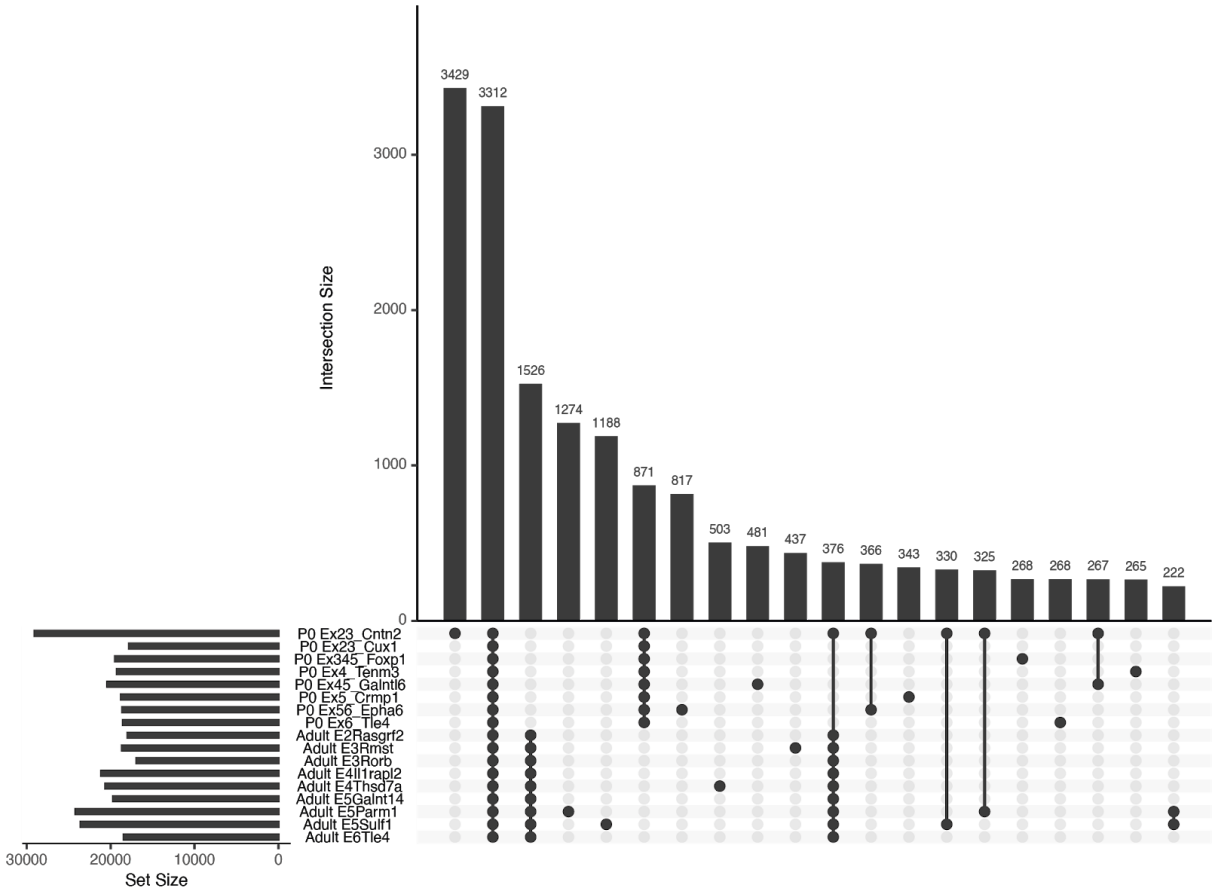

Figure S15: Overlap of open chromatin regions from different layers and time points.

Supplementary tables

Table S1: Assignment of cell types to categories for the human kidney dataset

Cell type	Category
Fetal B cell	Immune
Fetal Cap mesenchyme	Nephron progenitor
Fetal CD4 T cell	Immune
Fetal cDC2	Immune
Fetal CNT/PC - proximal UB	Nephron epithelium
Fetal Distal renal vesicle	Nephron progenitor
Fetal Distal S shaped body	Nephron progenitor
Fetal Endothelium	Endothelium
Fetal Fibroblast 1	Stroma
Fetal Fibroblast 2	Stroma
Fetal Innate like lymphocyte	Immune
Fetal Loop of Henle	Nephron epithelium
Fetal Macrophage 1	Immune
Fetal Macrophage 2	Immune
Fetal Medial S shaped body	Nephron progenitor

Fetal Megakaryocyte	Immune
Fetal Monocyte	Immune
Fetal Myofibroblast 1	Stroma
Fetal Myofibroblast 2	Stroma
Fetal Neutrophil	Immune
Fetal NK cell	Immune
Fetal Pelvic epithelium - distal UB	Nephron progenitor
Fetal Podocyte	Nephron epithelium
Fetal Proliferating B cell	Immune
Fetal Proliferating cap mesenchyme	Nephron progenitor
Fetal Proliferating cDC2	Immune
Fetal Proliferating distal renal vesicle	Nephron progenitor
Fetal Proliferating macrophage	Immune
Fetal Proliferating monocyte	Immune
Fetal Proliferating myofibroblast	Stroma
Fetal Proliferating stroma progenitor	Stroma
Fetal Proximal renal vesicle	Nephron progenitor
Fetal Proximal S shaped body	Nephron progenitor
Fetal Proximal tubule	Nephron epithelium
Fetal Proximal UB	Nephron progenitor
Fetal Stroma progenitor	Stroma
Mature Ascending vasa recta endothelium	Endothelium
Mature B cell	Immune
Mature CD4 T cell	Immune
Mature CD8 T cell	Immune
Mature Connecting tubule	Nephron epithelium
Mature Descending vasa recta endothelium	Endothelium
Mature Distinct proximal tubule 2	Nephron epithelium
Mature Epithelial progenitor cell	Nephron progenitor
Mature Glomerular endothelium	Endothelium
Mature Indistinct intercalated cell	Nephron epithelium
Mature MNP-a/classical monocyte derived	Immune
Mature MNP-b/non-classical monocyte derived	Immune
Mature MNP-c/dendritic cell	Immune
Mature Myofibroblast	Stroma
Mature NK cell	Immune
Mature NKT cell	Immune
Mature Pelvic epithelium	Nephron epithelium
Mature Peritubular capillary endothelium 1	Endothelium
Mature Peritubular capillary endothelium 2	Endothelium
Mature Podocyte	Nephron epithelium
Mature Proliferating Proximal Tubule	Nephron epithelium
Mature Proximal tubule	Nephron epithelium
Mature Thick ascending limb of Loop of Henle	Nephron epithelium
Mature Type A intercalated cell	Nephron epithelium
Mature Type B intercalated cell	Nephron epithelium

Table S2: Assignment of cell types to categories for the Tabula Muris dataset

Cell type	Category
bladder basal cell of urothelium	epithelial
bladder bladder cell	epithelial
bladder mesenchymal cell	connective
brainmicroglia microglial cell	immune
brainneurons astrocyte of the cerebral cortex	macroglial
brainneurons brain pericyte	pericyte
brainneurons endothelial cell	endothelial
brainneurons neuron	neuronal
brainneurons oligodendrocyte	macroglial
brainneurons oligodendrocyte precursor cell	macroglial
colon enterocyte of epithelium of large intestine	epithelial
colon epithelial cell of large intestine	epithelial
colon large intestine goblet cell	epithelial
fat B cell	immune
fat endothelial cell	endothelial
fat granulocyte	immune
fat mesenchymal stem cell of adipose	connective
fat myeloid cell	immune
fat natural killer cell	immune
fat neutrophil	immune
fat T cell	immune
heart cardiac muscle cell	muscle
heart endocardial cell	endothelial
heart endothelial cell	endothelial
heart epicardial adipocyte	adipose
heart fibroblast	connective
heart leukocyte	immune
heart smooth muscle cell	muscle
kidney kidney tubule cell	epithelial
liver endothelial cell of hepatic sinusoid	endothelial
liver hepatocyte	epithelial
lung endothelial cell	endothelial
lung stromal cell	connective
lung type II pneumocyte	epithelial
mammary basal cell	epithelial
mammary luminal epithelial cell of mammary gland	epithelial
mammary stromal cell	connective
marrow B cell	immune
marrow Fraction A pre-pro B cell	immune
marrow granulocyte	immune
marrow hematopoietic stem cell	immune
marrow monocyte	immune
marrow natural killer cell	immune
marrow neutrophil	immune
marrow T cell	immune
muscle B cell	immune
muscle endothelial cell	endothelial
muscle mesenchymal stem cell	connective

muscle skeletal muscle satellite cell	muscle
muscle skeletal muscle satellite stem cell	muscle
pancreas pancreatic A cell	endocrine
pancreas pancreatic acinar cell	exocrine
pancreas pancreatic D cell	endocrine
pancreas pancreatic ductal cell	epithelial
pancreas pancreatic PP cell	endocrine
pancreas type B pancreatic cell	endocrine
skin basal cell of epidermis	epithelial
skin epidermal cell	epithelial
skin keratinocyte stem cell	epithelial
spleen B cell	immune
spleen T cell	immune
thymus T cell	immune
tongue basal cell of epidermis	epithelial
tongue keratinocyte	epithelial
trachea epithelial cell	epithelial
trachea leukocyte	immune
trachea stromal cell	connective

Table S3: Assignment of excitatory cell types to categories for the SNARE-seq postnatal mouse dataset

Cell type	**Category**
Ex23_Cntn2	2/3
Ex23_Cux1	2/3
Ex345_Foxp1	3/4/5
Ex4_Tenm3	4
Ex45_Galntl6	4/5
Ex5_Crmp1	5
Ex56_Epha6	5/6
Ex6_Tle4	6

Table S4: Assignment of excitatory cell types to categories for the SNARE-seq adult mouse dataset

Cell type	**Category**
E2Rasgrf2	2
E3Rmst	3
E3Rorb	3
E4Il1rapl2	4
E4Thsd7a	4
E5Galnt14	5
E5Parm1	5
E5Sulf1	5
E6Tle4	6

Bibliography

1. Lambert, S. A. *et al.* The human transcription factors. *Cell* **172,** 650–665 (2018).
2. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39,** 381–399 (2014).
3. Jayaram, N., Usvyat, D. & R Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* **17,** 547 (2016).
4. Wunderlich, Z. & Mirny, L. A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* **25,** 434–440 (2009).
5. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117,** 185–198 (2004).
6. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31,** 107663 (2020).
7. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37,** 1452–1457 (2019).
8. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
9. Stewart, B. J. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365,** 1461–1466 (2019).
10. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14,** 1083–1086 (2017).
11. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
12. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26,** 990–999 (2016).
13. Smith, A. D., Sumazin, P., Xuan, Z. & Zhang, M. Q. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci USA* **103,** 6275–6280 (2006).
14. Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.* **22,** 1711–1722 (2012).
15. Madsen, J. G. S. *et al.* Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome Res.* **28,** 243–255 (2018).
16. Balwierz, P. J. *et al.* ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24,** 869–884 (2014).
17. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20,** 389–403 (2019).
18. Marx, V. Machine learning, practically speaking. *Nat. Methods* **16,** 463–467 (2019).
19. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep learning library for sequence data. *Nat. Methods* **16,** 315–318 (2019).
20. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28,** 739–750 (2018).
21. Paszke, A. *et al.* Automatic differentiation in PyTorch. in (NIPS 2017, 2017).
22. Andrews, T. S. & Hemberg, M. M3Drop: dropout-based feature selection for scRNASeq.

Bioinformatics **35,** 2865–2867 (2019).
23. Koo, P. K. & Eddy, S. R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.* **15,** e1007560 (2019).
24. Maslova, A. *et al.* Deep learning of immune cell differentiation. *Proc Natl Acad Sci USA* **117,** 25655–25666 (2020).
25. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443 (2014).
26. Kreidberg, J. A. WT1 and kidney progenitor cells. *Organogenesis* **6,** 61–70 (2010).
27. Wang, P. *et al.* Dissecting the Global Dynamic Molecular Profiles of Human Fetal Kidney Development by Single-Cell RNA Sequencing. *Cell Rep.* **24,** 3554-3567.e3 (2018).
28. Fang, Q. *et al.* Knockdown of ZFX suppresses renal carcinoma cell growth and induces apoptosis. *Cancer Genet.* **207,** 461–466 (2014).
29. Kaverina, N. V. *et al.* WT1 is necessary for the proliferation and migration of cells of renin lineage following kidney podocyte depletion. *Stem Cell Reports* **9,** 1152–1166 (2017).
30. Nguyen, T. A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26,** 1023–1033 (2016).
31. Gordon, S., Akopyan, G., Garban, H. & Bonavida, B. Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene* **25,** 1125–1142 (2006).
32. Debaize, L. & Troadec, M.-B. The master regulator FUBP1: its emerging role in normal cell function and malignant development. *Cell. Mol. Life Sci.* **76,** 259–281 (2019).
33. Seifert, L. L. *et al.* The ETS transcription factor ELF1 regulates a broadly antiviral program distinct from the type I interferon response. *PLoS Pathog.* **15,** e1007634 (2019).
34. Tena, J. J. *et al.* Odd-skipped genes encode repressors that control kidney development. *Dev. Biol.* **301,** 518–531 (2007).
35. Lu, B. C. *et al.* Etv4 and Etv5 are required downstream of GDNF and Ret for kidney branching morphogenesis. *Nat. Genet.* **41,** 1295–1302 (2009).
36. Boros, J. *et al.* Overlapping promoter targeting by Elk-1 and other divergent ETS-domain transcription factor family members. *Nucleic Acids Res.* **37,** 7368–7380 (2009).
37. Xu, J., Liu, H., Chai, O. H., Lan, Y. & Jiang, R. Osr1 Interacts Synergistically with Wt1 to Regulate Kidney Organogenesis. *PLoS ONE* **11,** e0159597 (2016).
38. Sun, Z. *et al.* EGR1 recruits TET1 to shape the brain methylome during development and upon neuronal activity. *Nat. Commun.* **10,** 3892 (2019).
39. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583,** 760–767 (2020).
40. Skinner, M. K., Rawls, A., Wilson-Rawls, J. & Roalson, E. H. Basic helix-loop-helix transcription factor gene family phylogenetics and nomenclature. *Differentiation.* **80,** 1–8 (2010).
41. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
42. Yoshida, H. *et al.* The cis-Regulatory Atlas of the Mouse Immune System. *Cell* **176,** 897-912.e20 (2019).

43. Murre, C. Helix-loop-helix proteins and lymphocyte development. *Nat. Immunol.* **6,** 1079–1086 (2005).
44. Kleiman, E., Jia, H., Loguercio, S., Su, A. I. & Feeney, A. J. YY1 plays an essential role at all stages of B-cell differentiation. *Proc Natl Acad Sci USA* **113,** E3911-20 (2016).
45. Hwang, S. S. *et al.* YY1 inhibits differentiation and function of regulatory T cells by blocking Foxp3 expression and activity. *Nat. Commun.* **7,** 10789 (2016).
46. McConnell, B. B. & Yang, V. W. Mammalian Krüppel-like factors in health and diseases. *Physiol. Rev.* **90,** 1337–1381 (2010).
47. Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell. Biol.* **20,** 429–440 (2000).
48. Wood, K. H. & Zhou, Z. Emerging molecular and biological functions of MBD2, a reader of DNA methylation. *Front. Genet.* **7,** 93 (2016).
49. Christoffels, V. M. *et al.* T-box transcription factor Tbx2 represses differentiation and formation of the cardiac chambers. *Dev. Dyn.* **229,** 763–770 (2004).
50. Sidwell, T. & Kallies, A. Bach2 is required for B cell and T cell memory differentiation. *Nat. Immunol.* **17,** 744–745 (2016).
51. Gray, L. T. *et al.* Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *elife* **6,** (2017).
52. Rambousek, L. *et al.* Aberrant expression of PAR bZIP transcription factors is associated with epileptogenesis, focus on hepatic leukemia factor. *Sci. Rep.* **10,** 3760 (2020).
53. Hawkins, N. A. & Kearney, J. A. Hlf is a genetic modifier of epilepsy caused by voltage-gated sodium channel mutations. *Epilepsy Res.* **119,** 20–23 (2016).
54. Hitzler, J. K. *et al.* Expression patterns of the hepatic leukemia factor gene in the nervous system of developing and adult mice. *Brain Res.* **820,** 1–11 (1999).
55. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571,** 505–509 (2019).
56. Tropea, D. *et al.* Gene expression changes and molecular pathways mediating activity-dependent plasticity in visual cortex. *Nat. Neurosci.* **9,** 660–668 (2006).
57. Xiang, G. *et al.* An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. *Genome Res.* **30,** 472–484 (2020).
58. Lindgren, D. *et al.* Cell-Type-Specific Gene Programs of the Normal Human Nephron Define Kidney Cancer Subtypes. *Cell Rep.* **20,** 1476–1489 (2017).
59. Weltmeier, F. & Borlak, J. A high resolution genome-wide scan of HNF4α recognition sites infers a regulatory gene network in colon cancer. *PLoS ONE* **6,** e21667 (2011).
60. Wei, G.-H. *et al.* Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J.* **29,** 2147–2160 (2010).
61. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **20,** 536–548 (2019).
62. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48,** D682–D688 (2020).
63. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47,** D766–D773 (2019).
64. Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* (2012).
65. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,**

831–838 (2015).
66. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).
67. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33,** 3645–3647 (2017).
68. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695,** 1–9 (2006).
69. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools : an*r* package for analyzing finite mixture models. *J. Stat. Softw.* **32,** (2009).
70. Whitfield, M. L., George, L. K., Grant, G. D. & Perou, C. M. Common markers of proliferation. *Nat. Rev. Cancer* **6,** 99–106 (2006).
71. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).
72. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19,** 477 (2018).
73. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35,** 316–319 (2017).
74. Chazarra-Gil, R., Hemberg, M., Kiselev, V. Y. & van Dongen, S. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *BioRxiv* (2020). doi:10.1101/2020.05.22.111211
75. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).