nature portfolio

https://doi.org/10.1038/s41586-024-08391-z

Supplementary information

A foundation model of transcription across human cell types

In the format provided by the authors and unedited

Supplementary Information for: A foundation model of transcription across human cell types

Xi Fu[1,2+*], Shentong Mo[3,4+], Alejandro Buendia[1+], Anouchka Laurent[5^], Anqi Shao[6^], Maria del Mar Alvarez-Torres[1], Tianji Yu[1], Jimin Tan[7], Jiayu Su[1], Romella Sagatelian[1], Adolfo A. Ferrando[6,8], Alberto Ciccia[9], Yanyan Lan[10], David M. Owens[6,11], Teresa Palomero[5,11], Eric P. Xing[3,4*], Raul Rabadan[1,2*]

[1]Program of Mathematical Genomics, Department of Systems Biology, Columbia University, New York, NY, USA
[2]Department of Biomedical Informatics, Columbia University, New York, NY, USA
[3]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA
[4]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[5]Institute for Cancer Genetics, Columbia University, New York, NY, USA
[6]Department of Dermatology, Columbia University, New York, NY, USA
[7]Institute for Systems Genetics, New York University Grossman School of Medicine, New York, NY, USA
[8]Regeneron Genetics Center, Regeneron, Tarrytown, NY, USA
[9]Department of Genetics and Development, Columbia University, New York, NY, USA
[10]Institute for AI Industry Research, Tsinghua University, Beijing, China
[11]Department of Pathology & Cell Biology, Columbia University, New York, NY, USA

+These authors have contributed equally to this work.
^These authors have contributed equally to this work.
*Co-correspondence: xf2217@cumc.columbia.edu, epxing@cs.cmu.edu, rr2579@cumc.columbia.edu.

Table of Contents

Supplementary Information ***3***

Generalizability study of GET **3**

Computational cost comparison **4**

Overall distribution of GET-MPRA prediction **5**

Enrichment analysis of GET-MPRA prediction **5**

Cross-cell-type motif contribution analysis **6**

Cross-cell-type embedding analysis **8**

Analysis of the GET causal network **9**

Analysis of AlphaFold prediction on TF structure **11**

References **13**

Supplementary Figures

Supplementary Figure 1. Raw gel images. (Separate file)

Supplementary Figure 2. Readout distribution of GET-MPRA prediction for different types of elements 5

Supplementary Figure 3. Analysis of GET identified upstream regulators 7

Supplementary Figure 4. Example latent embedding space after various transformer layers 8

Supplementary Figure 5. Causal motif-motif interaction network inferred from all cell types 9

Supplementary Figure 6. Overall network statistics of causal motif-motif network. 10

Supplementary Tables

Supplementary Table 1. Cell types used in the training process. (Separate file)

Supplementary Table 2. Comparison of GET, Enformer, and HyenaDNA model 4

Supplementary Information

Generalizability study of GET

We investigate GET's generalizability to adult cell types when trained solely on fetal data. Our findings show an average R^2 of 0.53 across diverse adult cell types, surpassing the baseline (R^2 = 0.33) obtained using corresponding fetal cell types for prediction (e.g. utilizing fetal astrocyte expression to predict adult astrocyte expression) (Extended Data Fig. 3a). The only three cell types where GET cannot beat this baseline are cell types with low cell counts in either ATAC-seq (Blood Brain Barrier Endothelial Cell and Tuft Cell, n < 600) or RNA-seq (pancreatic acinar cell).

Diverse data generation platforms and processing methods often present a significant challenge for the universal generalizability of pretrained models. To assess whether GET can be transferred to new sequencing platforms in such scenarios, we benchmarked its performance on 10x multiome sequencing of the lymph node[74] (Extended Data Fig. 3b, Methods: ATAC-seq data processing and RNA-seq data processing) using the leave-cell-type-out evaluation approach. Notably, GET maintained consistent prediction outcomes for both finetuned and left-out cell types (Extended Data Fig. 3c).

We also analyzed the potential of the GET model to investigate non-physiological cell types, including patient tumor cells. We applied GET to 10x multiome data from 16 glioblastoma (GBM) IDH1 wild-type patients[26] (Extended Data Fig. 3d). The zero-shot performance on GBM tumor cells across held-out patients achieved a mean Pearson correlation of 0.68. When finetuning GET on a single patient sample, the Pearson correlation improved to above 0.90 (Extended Data Fig. 3e). Leave-one-chromosome-out finetuning on a single patient showed a mean Pearson of 0.73 across chromosomes (Extended Data Fig. 3f).

Given GET's demonstrated adaptability, we explored its applicability to other experimental assays. We applied GET to predict cap analysis of gene expression (CAGE) from the K562 cell line as a test of sequencing platform shift. When comparing finetuned GET to Enformer's predictions for K562 CAGE, GET achieved a Pearson correlation of 0.90 compared to Enformer's performance of 0.74 (Extended Data Fig. 4a).

We finetuned GET on scATAC-seq from fetal astrocytes, finding a Pearson correlation of 0.71 comparable to Enformer's prediction of DNase for the same cell type (Extended Data Fig. 4b). Using only motif inputs, we finetuned GET to predict quantitative K562 peak-level chromatin accessibility from bulk ENCODE OmniATAC (ENCSR483RKN) data, achieving a 0.81 Pearson correlation across different leave-one-chromosome-out experiments (Extended Data Fig. 4c, N=22), approaching Enformer's DNase prediction performance in K562. Variable peak sequence length was used in K562 OmniATAC data and partially encoded in the motif binding score, which may make the prediction easier for GET, while Enformer benefits from hundreds of other K562 epigenomic tracks used in its training. We also found that the finetuned ATAC-prediction models are relatively robust to leave-out-motif training when the number of omitted motifs is less than 10, indicating that the model has gained robustness to feature corruption from the masked prediction objective (Extended Data Fig. 4d, Method: Leave-out-motif evaluation).

Computational cost comparison

Here we provide training batch size, RAM per batch (GPU), inference time per 1000 samples, number of model parameters, and maximum sequence length for GET, Enformer, and HyenaDNA as a measure of computational cost. We trained or finetuned GET, Enformer, and HyenaDNA for different tasks with various dataset sizes and approaches (i.e. masked pretraining on chromatin accessibility and expression head finetuning for GET; linear probing on expression output tracks for Enformer; multiclass finetuning on chromatin profile prediction for HyenaDNA). Hence, we do not report a side-by-side comparison of training times for the methods.

Supplementary Table 2. Comparison of GET, Enformer, and HyenaDNA models

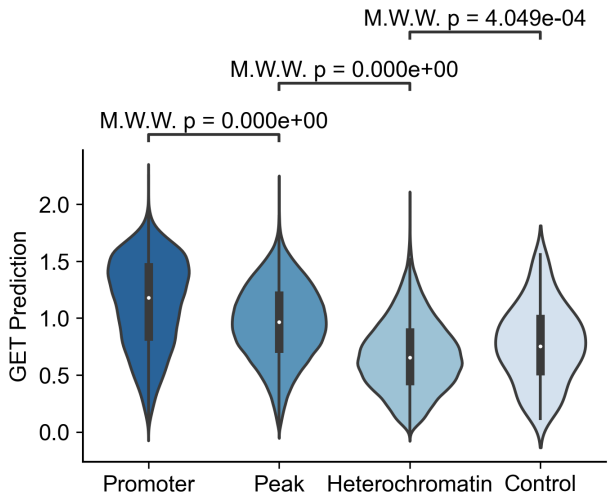
	GET	**Enformer**	**HyenaDNA**
Training batch size	64 examples per batch	12 examples per batch	1 example per batch
RAM per batch (GPU)	3 GB on 1 A6000 (200 region setting) 10 GB on 1 A6000 (900 region setting)	48 GB on 1 A6000	35 GB on 1 A6000
Inference time per 1000 samples	2.6 seconds on 1 RTX 3090 (200 region setting) 13.8 seconds on 1 RTX 3090 (900 region setting)	107 seconds on 1 RTX 3090	12,000 seconds on 1 A6000[†]
Number of model parameters	86M	251M	6.6M
Maximum sequence length	Dependent on density of peak calling (typical sequence length: 2.7M)[‡]	393,216	1M

[†] One RTX 3090 with 24 GB GPU memory was sufficient for running inference for GET and Enformer. Running inference for HyenaDNA required more RAM, in which we opted to use 1 A6000 with 48 GB GPU memory.

[‡] GET's effective sequence length depends on the density of peak calling for a particular region, i.e. a very sparsely sampled peak set achieves a theoretical upper bound of capturing the entire genome with one input sequence. We estimate a typical ATAC-seq peak set to contain 100,000 peaks. Assuming the peaks are evenly distributed across the genome, a 200-peak input sample into the model spans 6M base pairs, while a 900-peak input sample spans 27M base pairs. In our study, we primarily use 900-peak input samples in tandem with an atlas-level joint peak set (1M peaks), such that one input sequence encodes 2.7M base pairs on average. Capturing a receptive field at this scale aligns well with the biological prior that further genomic distances are rarely found to interact in three-dimensions, as measured by Hi-C data.

Overall distribution of GET-MPRA prediction

To compare the overall distribution of GET-MPRA predictions across all 226,243 elements with experiment measurements, we followed the categorization of the different types of elements originally provided by Agarwal et al.[6] “Promoters” are defined as protein-coding gene promoters, “Peaks” are defined as accessible peaks called from ENCODE K562 ATAC-seq data (which contain most of the enhancers), “Heterochromatin” are defined as genomic tiles of heterochromatin regions around selected genes (e.g. GATA1), and “Control Group” are defined as synthetic elements.



Supplementary Figure 2. Readout distribution of GET-MPRA prediction for different types of elements (n = 226,243). Two sided Mann-Whitney U-test: Promoter vs. Peak: p < 1e-237; Peak vs. Heterochromatin: p < 1e-237; Heterochromatin vs. Control: p = 4.049e-04. ***indicates 1.00e-04 < p <= 1.00e-03; **** indicates p <= 1.00e-04.

Enrichment analysis of GET-MPRA prediction

Using all K562 (untreated) ChIP-seq peaks from the ChIP-Atlas, including 32 targeting histone marks and 562 targeting TFs (more than 500) and other targets (e.g. G-quadruplex, AGO1/2, Cas9), we performed enrichment analysis against the lentiMPRA promoter (approximately 57,000) and ATAC peak (approximately 169,000) library. This was further subdivided into four groups respectively based on GET prediction (P) (P+: >1; P-: <1) and observed readout (O) (O+: > 0.5; O-: < 0.5), visualized in Extended Data Fig. 5b.

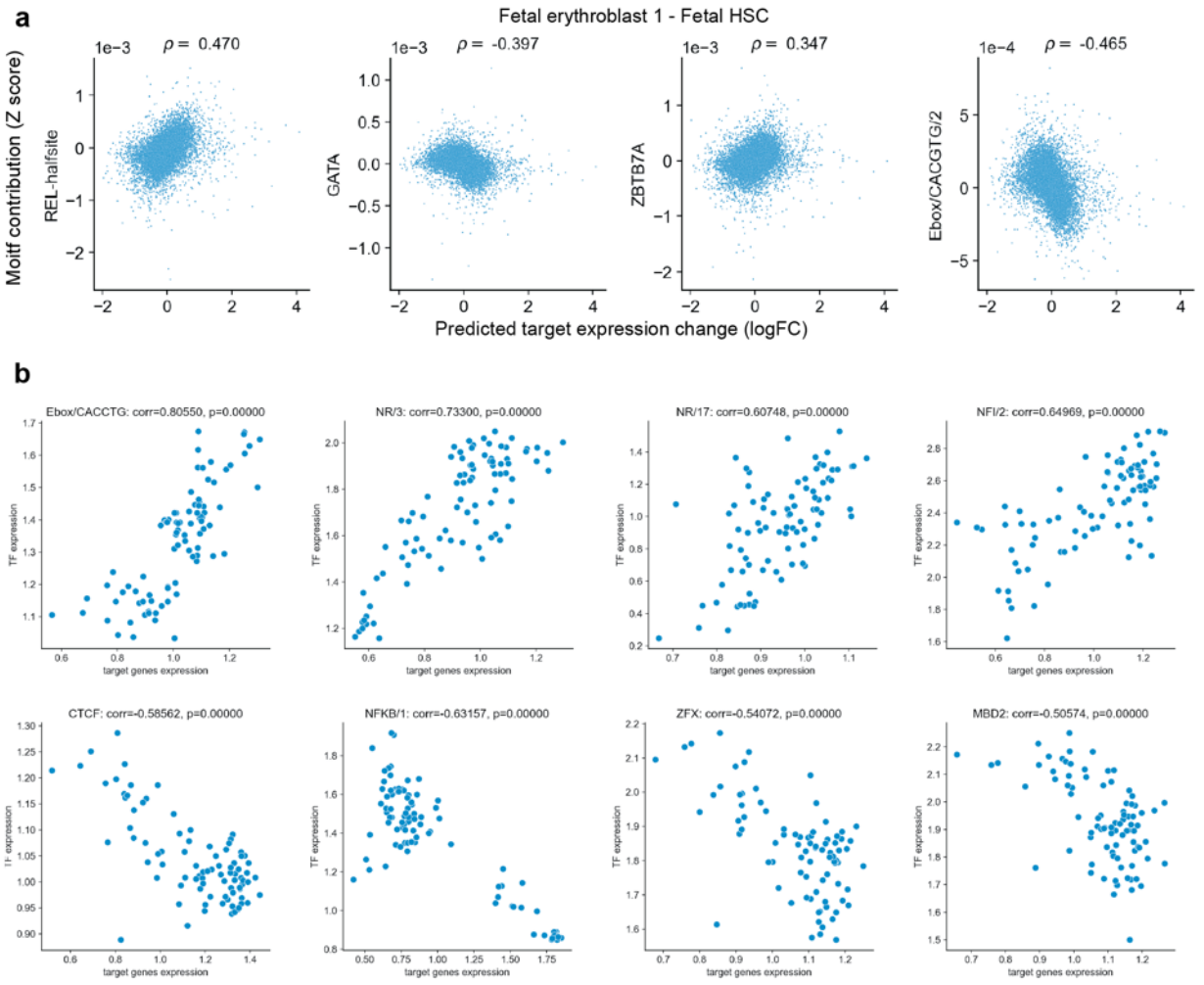
The enrichment Fisher’s exact test was performed using the bedtools fisher command. The overlapping ChIP-seq peaks for the same target were merged before testing. Log_2 fold enrichment for significant results (p < 0.05 in any group) were visualized and compared using a heatmap for all histone marks and the 100 most variable TFs (in terms of enrichment across the four expression-based groups; Extended Data Fig. 5c,d). We noticed that O+ promoters showed larger H3K4me3 enrichment, and P+ promoters showed larger H3K4me3 and H3K27ac enrichment. Interestingly, P+O- promoters have a distinct signature of H3K9me3, H2AK119Ub, and H3K79me2 compared to the P+O+ group, potentially relevant to promoters that were originally regulated by the Polycomb complex or in heterochromatin. Since lentiMPRA elements are

relatively newly inserted to the genome, these repressive mechanisms may have not yet been set up, leading to a discrepancy between RNA-seq based prediction and lentiMPRA readout. Similarly, accessible peaks are enriched in H3K27ac and H3K4me1/2/3 in general, while P+O+ peaks are particularly enriched in H3K122ac, a signature of active enhancers and transcription. Comparing P-O+ peaks to P+O+ peaks, H3K27me2, H3K14cr, and H4K20me1 are more enriched. Pioneer factors like GABPA, NFYA, and NEUROD1 are enriched in P+ and particularly P+O+ promoter and enhancers. Other signature transcription factors are CEBPZ (P+ vs. P- promoters and P-O+ vs. P-O- peaks) and WDR5 (P+O+ vs. P-O+ peaks).

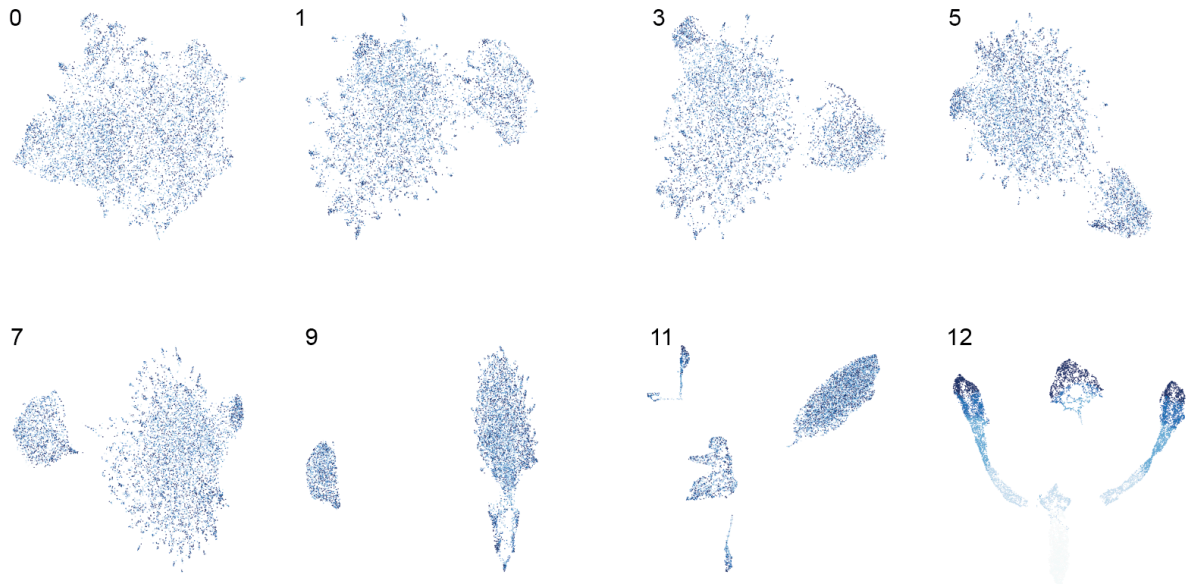
Cross-cell-type motif contribution analysis

To assess GET's capability to detect significant regulatory alterations across different cell states, we focused on the differential expression between fetal erythroblast and fetal hematopoietic stem cells (HSC). Our expectation is that the genes that mark the lineage differentiation should show larger gradients from lineage specific transcription factors than those that are unchanged across lineages. Our findings confirmed this, as we noted substantial Spearman correlation between the motif contributions in erythroblasts and the differential expression log fold changes for several erythroblast-related transcription factors, including GATA, ZBTB7A, and MZF1 (Supplementary Fig. 3a).

Extending our analysis to encompass all fetal and adult cell types, we found that for certain well-known regulators, such as CTCF, MBD2, NFKB, and NFI, there exists a significant correlation between the mean expression of inferred target genes and the mean expression of all regulators within the corresponding family (e.g. NFIA, NFIB, NFIC, NFIX for the NFI motif, Supplementary Fig. 3b). Overall, we conclude that GET possesses the ability to learn meaningful regulatory information that is naturally transferable between cell states.



Supplementary Figure 3. Analysis of GET identified upstream regulators. a. Correlation between motif contribution (Y axis) in Fetal Erythroblast 1 and the predicted target gene expression change (X axis) between Fetal Erythroblast 1 and Fetal HSC. Four motifs relevant to erythroid differentiation are shown. **b.** Correlation between mean transcription factor expression in a motif cluster versus the mean expression of their target genes predicted by GET across fetal cell types.

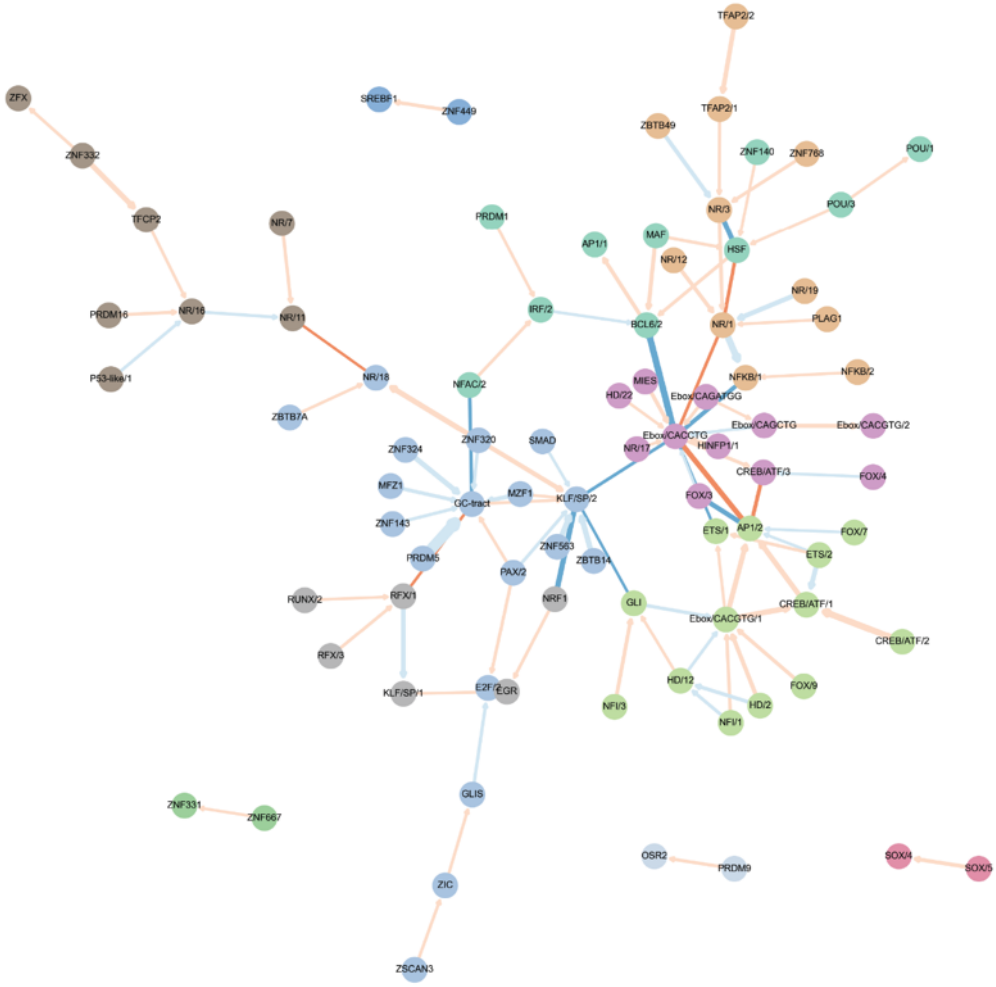


Supplementary Figure 4. Example latent embedding space after various transformer layers. Each dot is a gene colored by expression value.

Cross-cell-type embedding analysis

In order to further visualize what GET learns from different cell types, we explored different embedding layers of GET. We found that the embedding from final layers correlates well with expression levels, while earlier layers are more indicative of differences in regulatory grammar (Supplementary Fig. 4, Methods: Regulatory embedding). To investigate whether the embedding tied to regulatory grammar retains cell type-specific information, we gathered the first transformer layer's embeddings for all promoters across cell types. This allows us to capture not only the motif information of the promoters but also that of the CREs owing to the attention mechanism. Intriguingly, a t-distributed stochastic neighbor embedding (t-SNE)[75] visualization of randomly sampled embeddings showed motif separation but no cell type differentiation, suggesting that with a sufficient number of cell types, most regulatory grammar is shared across cell types, although they may be instantiated on different genes (Extended Data Fig. 7c). Nonetheless, when we subset the embeddings to only three specific cell types (fetal astrocyte and two fetal erythroblast subclusters), the embedding visualized with UMAP[68] exhibited distinct clusters for astrocytes and erythroblast genes (Extended Data Fig. 7d). This result further corroborates that GET is proficient at discerning cell-type-specific regulatory information.

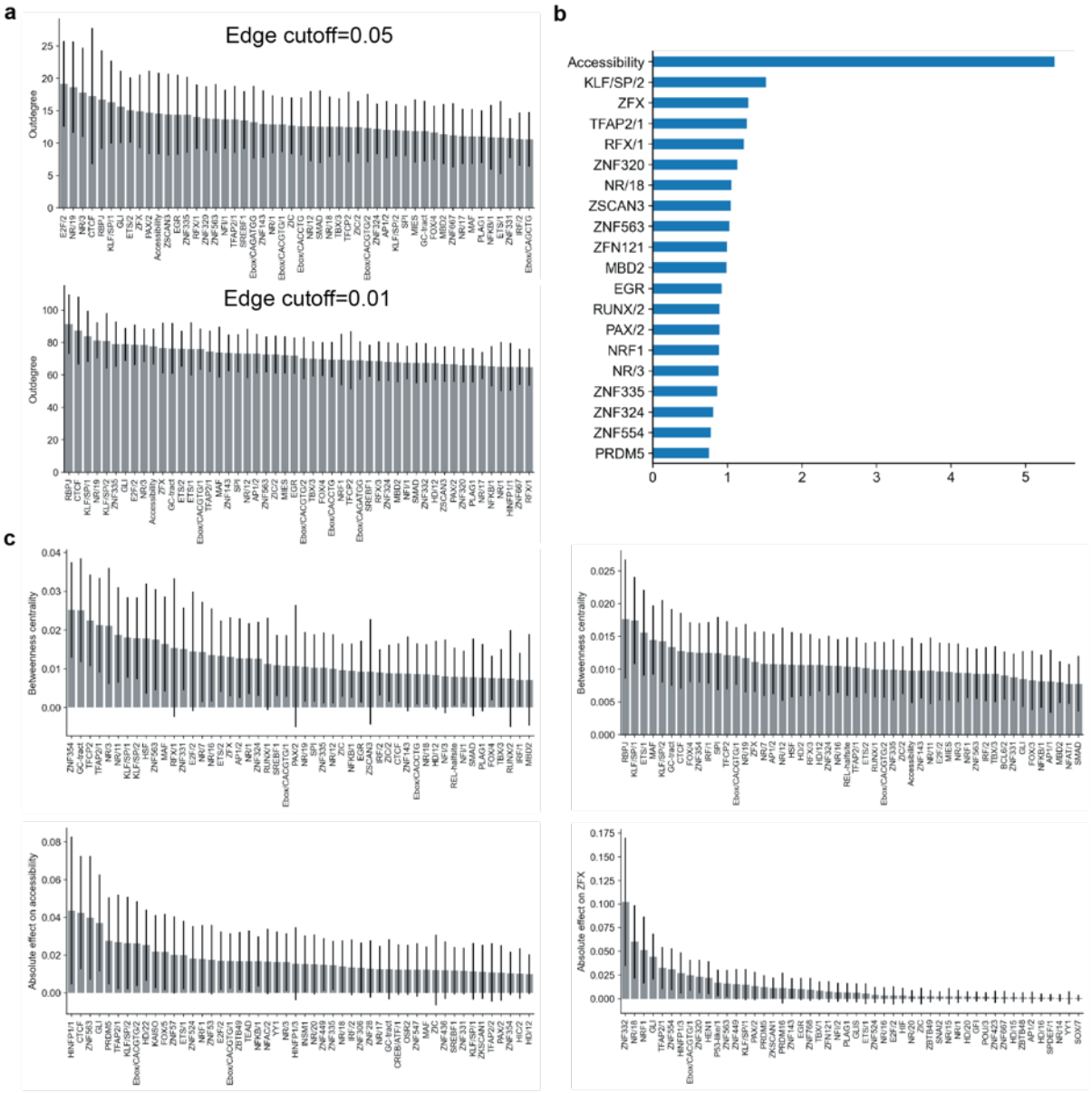
Looking further into the astrocyte-specific gene cluster (cluster #2 in Extended Data Fig. 7e), we discovered that this gene set is particularly enriched in the development of the nervous system and includes astrocyte transcription factors such as NFIA[76,76,77] and GLI3[78] (Extended Data Fig. 7f). Moreover, a comparison of motif contribution across clusters revealed a higher presence of NFI motifs in the astrocyte-specific cluster (Extended Data Fig. 7g), shedding light on the unique regulatory program within astrocytes.



Supplementary Figure 5. Causal motif-motif interaction network inferred from all cell types. Edges with absolute effect size smaller than 0.1 and isolated nodes are removed for visualization purposes. Node colors represent communities. Edge weights represent causal effect sizes. Edge colors represent positive (orange) or negative (blue) effects.

Analysis of the GET causal network

A visualization of the cell-type-agnostic GET causal network with effect size cutoff=0.1 is visualized above in Supplementary Fig. 5. We identified factors such as CTCF, KLF/SP/2 (GC rich motif), TFAP2/1, ZFX, RBPJ, Accessibility, and methylation-associated E2F as having the largest outdegree in the causal network across diverse cell types (Supplementary Fig. 6a), indicating the general importance of these factors in transcription regulation. We also experimented with the GET model trained using both quantitative ATAC signal and motif binding score as input and found similar top outdegree transcription factors (Supplementary Fig. 6b).



Supplementary Figure 6. Overall network statistics of causal motif-motif network. **a.** Outdegree distribution across cell-type-specific causal networks at two different absolute edge weight cutoffs. Error bar shows standard deviation across cell types. **b.** Outdegree distribution of Quantitative-ATAC GET model trained using fetal data. **c.** Betweenness centrality and absolute effect on accessibility or ZFX across cell-type-specific causal networks at edge cutoff 0.01.

We showcase four subnetworks identified by GET in Extended Data Fig. 7i. MBD2 and MECP2 show negative interaction with a promoter-enriched motif, GC-tract, which aligns with the well-known repressive effect of promoter methylation on gene expression[37,38]. The pair NR/17-Ebox/CACCTG highlights a functional regulatory complex ESR1-ZEB1[79]. NR/17-GLI is also supported by the known physical interaction between ESR1 and GLI3[80]. NFKB/1-Ebox/CACCTG has a strong interaction with negative effect size, while their representative transcription factors, RELA and SNAI1, have been shown interactusing co-immunoprecipitation[39]. ZFX is positively linked to TFAP2/1 and has been shown to co-localize with TFAP2A using ChIP-seq[81,82]. The link between NFKB/1 and NFKB/2 is expected as they are commonly annotated to NFKB factors, despite having different sequence content.

To characterize potential EP300-mediated cofactor binding, we analyzed the cell-type agnostic interaction prediction used in the ChIP-seq colocalization benchmark (Fig. 4b). We defined a motif-motif interaction as potentially EP300-mediated if any TF in each motif cluster exhibited strong ChIP-seq defined colocalization (ChIP-Atlas HepG2 colocalization score = 9). Focusing on motif-motif interactions with an EP300-based annotation (n = 15,252), we found that those in the top 5% of all motif-motif pairs (n = 763) showed enrichment in potentially EP300-mediated interactions (n = 61 among 930 potentially EP300-mediated pairs; Fisher's exact test P value = 0.03). This enrichment strengthened when considering STRING-based physical interactions between TF and EP300 to complement ChIP-seq defined pairs ($\mathrm{P\ value} = 6.53 \times 10^{-7}$, n = 117 among 1490 pairs). Due to computational constraints of AlphaFold-Multimer prediction, we focused on a subset of predicted interacting TFs. For each GET Causal interaction identified across all cell types (with 95th percentile absolute effect size in each cell type), we mapped the motif to the corresponding highest expressed TF and converted the interaction to a TF-TF interaction. We selected TF-TF interactions appearing in more than 5 cell types, resulting in interactions spanning 188 TFs. For each TF, we used AlphaFold2 to model its interaction with EP300 in an all-against-all manner across all pLDDT defined protein domains. We defined the AlphaFold-based EP300 interaction score for each TF as the maximum interface-pTM (ipTM) across all domain-domain multimer predictions with at least 50 prediction confidence (pLDDT). The TF-TF AlphaFold-based EP300 interaction score was defined by multiplying their corresponding scores and mapped back to motif-motif interactions for comparison with GET Causal effect size. When considering all pairs of motifs with a defined AlphaFold-based score, we found that pairs with higher ipTM tended to have higher GET Causal effect size ($\mathrm{P\ value} = 9.73 \times 10^{-3}$, Mann-Whitney U test; Fisher's exact test $\mathrm{P\ value} = 7.3 \times 10^{-3}$, one-sided, odds ratio = 1.57, with the same GET Causal cutoff as above; Extended Data Fig. 8h).

Analysis of AlphaFold prediction on TF structure

To evaluate whether AlphaFold predictions reflect true interactions, we assessed the result on predicting whether a transcription factor family can act as an "intra-family binder" based on the heuristic that intra-family binders should have a higher chance to form homodimers due to very similar structured domains. We found that AlphaFold can reach an area under the receiver operating characteristic (AUROC) of 0.69 and an average precision (AUPR) of 0.41 (Extended Data Fig. 8a). The accuracy of AlphaFold dimer prediction is exemplified by the perfectly aligned

TFAP2A structure to experimental results[83] (Extended Data Fig. 8b), even though there is no other similar template in PDB.

References

74. Flash-Frozen Lymph Node with B Cell Lymphoma (14k sorted nuclei). *10x Genomics* https://www.10xgenomics.com/resources/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard-2-0-0.
75. Simoes-Costa, M. & Bronner, M. E. Insights into neural crest development and evolution from genomic analysis. *Genome Research* **23**, 1069–1080 (2013).
76. Kang, P. *et al.* Sox9 and NFIA Coordinate a Transcriptional Regulatory Cascade during the Initiation of Gliogenesis. *Neuron* **74**, 79–94 (2012).
77. Tchieu, J. *et al.* NFIA is a gliogenic switch enabling rapid derivation of functional human astrocytes from pluripotent stem cells. *Nature Biotechnology* **37**, 267–275 (2019).
78. Petrova, R., Garcia, A. D. R. & Joyner, A. L. Titration of GLI3 Repressor Activity by Sonic Hedgehog Signaling Is Critical for Maintaining Multiple Adult Neural Stem Cell and Astrocyte Functions. *J Neurosci* **33**, 17490–17505 (2013).
79. Mohammadi Ghahhari, N. *et al.* Cooperative interaction between ERα and the EMT-inducer ZEB1 reprograms breast cancer cells for bone metastasis. *Nat Commun* **13**, 2104 (2022).
80. Massah, S. *et al.* Gli activation by the estrogen receptor in breast cancer cells: Regulation of cancer cell growth by Gli3. *Mol Cell Endocrinol* **522**, 111136 (2021).
81. Kenny, C. *et al.* TFAP2 paralogs facilitate chromatin access for MITF at pigmentation and cell proliferation genes. *PLOS Genetics* **18**, e1010207 (2022).
82. Rhie, S. K. *et al.* ZFX acts as a transcriptional activator in multiple types of human tumors by binding downstream from transcription start sites at the majority of CpG island promoters. *Genome Res.* (2018) doi:10.1101/gr.228809.117.
83. Liu, K. *et al.* Structural basis for specific DNA sequence motif recognition by the TFAP2 transcription factors. *Nucleic Acids Research* gkad583 (2023) doi:10.1093/nar/gkad583.
84. Created in BioRender. Fu, X. (2024) BioRender.com/h56n108