GET: a foundation model of transcription across human cell types

Xi Fu[1,2+*], Shentong Mo[3,4+], Anqi Shao[5], Anouchka Laurent[6], Alejandro Buendia[1], Adolfo A. Ferrando[6,7], Alberto Ciccia[8], Yanyan Lan[9], Teresa Palomero[6,10], David M. Owens[5,10], Eric P. Xing[3,4*], Raul Rabadan[1,2*]

[1]Program for 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]Department of Dermatology, Columbia University, New York, NY, USA

[6]Institute for Cancer Genetics, Columbia University, New York, NY, USA

[7]Regeneron Genetics Center, Regeneron, Tarrytown, NY, USA

[8]Department of Genetics and Development, Columbia University, New York, NY, USA

[9]Institute for AI Industry Research, Tsinghua University, Beijing, China

[10]Department of Pathology & Cell Biology, Columbia University, New York, NY, USA

+These two authors have contributed equally to this work.

*Co-correspondence:

xf2217@cumc.columbia.edu, epxing@cs.cmu.edu, rr2579@cumc.columbia.edu.

Abstract
Transcriptional regulation, involving the complex interplay between regulatory sequences and proteins, directs all biological processes. Computational models of transcriptions lack generalizability to accurately extrapolate in unseen cell types and conditions. Here, we introduce GET, an interpretable foundation model, designed to uncover regulatory grammars across 213 human fetal and adult cell types. Relying exclusively on chromatin accessibility data and sequence information, GET achieves experimental-level accuracy in predicting gene expression even in previously unseen cell types. GET showcases remarkable adaptability across new sequencing platforms and assays, enabling regulatory inference across a broad range of cell types and conditions, and uncovering universal and cell type specific transcription factor interaction networks. We evaluated its performance on prediction of regulatory activity, inference of regulatory elements and regulators, and identification of physical interactions between transcription factors. Specifically, we show GET outperforms current models in predicting lentivirus-based massive parallel reporter assay readout with reduced input data. In Fetal erythroblast, we identify distal (>1Mbp) regulatory regions that were missed by previous models. In B cell, we identified a lymphocyte-specific transcription factor-transcription factor interaction that explains the functional significance of a lymphoma-risk predisposing germline mutation. In sum, we provide a generalizable and accurate model for transcription together with catalogs of gene regulation and transcription factor interactions, all with cell type specificity.

Main

Transcriptional regulation constitutes a critical yet largely unresolved domain, underpinning diverse biological processes, including those associated with human genetic diseases and cancers[1]. A conserved regulatory machinery orchestrates transcriptional changes, including transcription factors that bind to regulatory sequences, coactivators, mediator, and core transcriptional factors and RNA Polymerase II[2–4]. While different cell types may possess different subsets of regulatory regions, the biochemistry of protein-protein interaction and protein-DNA interaction remains largely the same across cell types when epigenetic conditions are fixed. Clustering of known transcription factor binding site motifs[5] demonstrates great functional redundancy in transcription factor DNA binding domains, further reducing the combinatorial variability of regulatory interactions. However, our understanding of transcription regulation is often limited to specific cell types, and it is not clear how the combinatorial interaction of different transcription factors determines the diversity of expression profiles observed across cell types.

Advances in sequencing technology and the adoption of sophisticated machine learning architectures have enabled the exploration of expression and associated noncoding regulatory features across a broad spectrum of cell types. Traditional methods, such as Expecto[6] and Basenji2[7], utilized convolutional neural networks for shorter input sequences, while state-of-the-art approaches like Enformer[8] extended capabilities with transformer architecture. Nonetheless, existing models present challenges. A key limitation is they can only make predictions on the training cell types, hindering the generalizability and utility of the model.

In the landscape of machine learning and computational biology, foundational models like GPT4[9] and ESM2[10] are emerging as a transformative approach. These models serve as a foundation, upon which specialized adaptations can be built to address specific tasks or challenges. By utilizing extensive pretraining on broad and diverse datasets, foundation models provide a generalized understanding of underlying patterns and relationships. In the field of transcriptional regulation, a foundation model has the potential to synthesize the vast complexities of regulatory mechanisms across various cell types, offering a versatile framework that can be fine-tuned to target specific applications, cell types, or conditions.

Here we introduce the general expression transformer (GET), an interpretable foundation model for transcription regulation across 213 human fetal and adult cell types that exhibits universal applicability and exceptional accuracy. GET learns transcriptional regulatory syntax from chromatin accessibility data across hundreds of diverse cell types and successfully predicts gene expression in both seen and unseen cell types, approaching experimental accuracy. The versatile nature of GET allows it to be transferred to different sequencing platforms and measurement techniques. Additionally, it offers zero-shot prediction of reporter assay readout in new cell types, potentiating itself as a prescreening tool for cell type specific regulatory elements. GET outperforms previous state of the art models in identifying cis-regulatory elements, and identifies novel and known upstream regulators of fetal hemoglobin. Through interpreting GET, we provide rich regulatory insight for almost every gene in 213 cell types. Using coregulation information predicted by GET, we performed causal discovery to pinpoint potential transcription factor-transcription factor interactions and constructed a structural interaction catalog of human transcription factors and coactivators. Using information provided by GET, we successfully identified a lymphocyte-specific transcription factor-transcription factor interaction involving PAX5 and retinoic acid receptor family transcription factors, and highlighted a possible disease driving mechanism of a lymphoma-associated germline-variants through affecting the binding of PAX5 disordered region to the nuclear receptor domain of retinoic acid receptors. Overall, GET's broad applicability and profound understanding of transcription regulation will advance understanding of noncoding genetic variants and guide de novo design of cell-type specific transcriptional regulatory circuits and transcription factors for synthetic biology application.

GET, a foundation model for transcription regulation across 213 human cell types

We embarked on developing GET, a novel foundational model to comprehend the transcription regulation across a diverse range of cell types. Unlike previous models such as Enformer[8], GET employs an extensive effective sequence length exceeding 2 Mbps and is not constrained to making predictions in only training cell types.

The design philosophy of GET is rooted in the conceptual understanding of transcription regulation (**Figure 1b**). At the forefront, Promoter and related contextual regulatory elements can be characterized by how well they bind different transcription factors (motif binding score, **Methods**) and how accessible they are in specific cell types. These features shape a chromatin

environment (p(X)) that governs RNA polymerase II (PolII) functioning. Using an embedding and attention architecture[11] specifically designed for the regulatory elements (**Methods**), we performed self-supervised pretraining to let GET learn how the regions and features interact with each other across diverse cell types. Specifically, by randomly masking out regulatory elements, the model is trained to predict the motif binding scores and optionally accessibility score in the masked region. Subsequently, PolII will translate the chromatin environment p(X) into an expression readout E (**Figure 1b**). A finetuning stage with the same architecture but a different output head will simulate this process. This two-stage design makes it possible to use chromatin accessibility data with no paired expression measurement, greatly improving the diversity of regulation information in the training data.

The pretraining of GET uses of pseudobulk chromatin accessibility gathered from single cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) data across 213 human fetal and adult cell types[12–14]. Out of these, 153 were coupled with expression data, acquired either through a multiome protocol or separate single cell RNA sequencing (scRNA-seq) experiments[15,16] (**Methods and Data Availability**). We calculate the motif binding score using known position weighted matrix and summarized them according to sequence similarity to reduce feature redundancy[5]. Assuming additivity in motif binding score, every sample is a region*feature matrices derived from a continuous range on the accessible genome across different cell types. This design of model input ensures both cell type specificity and generalizability while enabling efficient computational modeling. Strand-specific expression values are assigned to each region based on their overlap with expressed gene's promoters.

GET accurately predicts gene expression in unseen cell types at experimental accuracy

We first assessed GET's ability to accurately predict gene expression in unseen cell types in a setting where one cell type is left out during the expression finetuning process. Remarkably, GET demonstrated the capacity to consistently predict the expression of the left-out cell types at a level of accuracy comparable to experimental standards, even when trained without quantitative accessibility signals. An example can be taken from left-out astrocytes, where the Pearson correlation between GET's predicted expression values and the observed expression reached 0.94 (R^2 = 0.88), a result that is in line with experimental accuracy[17] (Pearson r = 0.92-0.99, **Supplementary Figure 2a**). GET's performance surpasses both transcription start site (TSS) accessibility (r = 0.47, R^2 = -0.23) and gene activity score[18] (r = 0.51, R^2 = -0.67), emphasizing the significance of DNA sequence specificity and distal context information in transcription regulation. Furthermore, GET managed to outperform two robust benchmarks, including top correlated cell type expression (r = 0.83, R^2 = 0.62) and mean expression across training cell types (r = 0.78, R^2 = 0.53; as illustrated in **Figure 2**). Additional validation was carried out, confirming GET's capability to make cell-type-specific predictions, as evidenced by a Pearson correlation of 0.91 (R^2 = 0.82) between predicted and observed log fold change for Fetal astrocyte and Fetal erythroblast expression (see Supplementary **Figure 2b**).

We proceeded to investigate GET's generalizability to adult cell types when trained solely on fetal data. Our findings showed an average R^2 of 0.53 across diverse adult cell types, once

again surpassing the baseline (R^2 = 0.33) obtained using corresponding fetal cell types for prediction (e.g., utilizing Fetal astrocyte to predict Adult astrocyte). The only 3 cell types where we cannot beat the baseline are cell types with low cell counts in either ATAC-seq or RNA-seq label (pancreatic acinar cell). This result reinforces the proposition that GET can extract common regulatory mechanisms that span across various cell types and stages of life.

To ascertain the impact of cross-cell-type pretraining on prediction performance, we finetuned a GET model from random initialization, which exhibited a substantial drop in performance compared to the pretrained version with the same number of training epochs (Pearson r: 0.596; Spearman rho: 0.642). Extending the training period for this baseline failed to enhance its performance (Pearson r: 0.607; Spearman rho: 0.658), highlighting the essential role of pretraining in facilitating model generalization.

In summation, our study demonstrates that by leveraging widely accessible ATAC-seq data and established transcription factor binding motifs, GET acquires a broad understanding of the regulatory code, empowering it to predict unseen cell type expression with experimental precision.

GET can be transferred to different sequencing platforms and measurements

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 a wholly new sequencing platform in such scenarios, we benchmarked its performance on 10x multiome sequencing of the lymph node[19] (**Figure 2a, Methods**) using the leave-cell-type-out evaluation approach. Notably, GET maintained consistent prediction outcomes for both finetuned or left-out cell types.

Given GET's demonstrated adaptability, we explored its applicability to other experimental assays. As a representative example, chromatin accessibility was chosen. We first pretrained GET model using only motif binding scores in accessible regions, and then fine-tuned it to predict quantitative K562 peak-level chromatin accessibility from both ENCODE OmniATAC[20] or NEAT-seq[21] data using a split chromosome evaluation method. Remarkably, the model achieved a Pearson correlation exceeding 0.98 for the left-out chromosome 11 (**Supplementary Figure 3a**). This might be attributed to the intrinsic association between chromatin accessibility and local DNA sequence patterns.

Zeroshot GET prediction of expression-driving regulatory elements in new cell type

Building on the versatility of GET across diverse platforms and measurements, we now venture into examining its capacity for zero-shot prediction of expression-driving regulatory elements in unseen cell types. Lentivirus-based massively parallel reporter assay (lentiMPRA) provides a robust mechanism to test the regulatory activity of numerous genetic sequences by integrating them into the genome, thereby circumventing the limitations inherent in episomal MPRAs and ensuring relevant biological readouts in hard-to-transfect cell lines[22]. Recently utilized to assess over 200,000 sequences in the K562 cell line, this experimental assay has created a

comprehensive benchmark dataset for evaluating whether the GET model can identify regulatory elements in a cell-type specific context[23] (**Figure 2c**).

In an *in silico* procedure akin to the lentiMPRA experiment, we employed the GET model finetuned on K562 chromatin accessibility and expression profile, which has not seen any lentiMPRA data. We then constructed the sequences for insertion, including both the regulatory sequence and minipromoters, randomly inserting these sequences across the genome. Utilizing the GET model, we inferred the activity of the mini promoter within the corresponding chromatin context and averaged over all insertions to obtain a mean readout indicative of the regulatory activity (**Figure 2c, Methods**). We found that best performance is achieved when we combine the mean expression readout with GET-predicted accessibility of the inserted element (**Supplementary Figure 3b**).

Upon examination of the readout distribution for different types of elements (**Figure 2d**), we found our predictions to be consistent with experimental data. Promoter sequences exhibited the highest GET-MPRA readout, followed by chromatin accessibility peak sequences, with heterochromatin sequences registering the lowest readouts, and control sequences spanning a wide range of readout values (**Figure 2d**).

When benchmarking our model against Enformer, the previous state-of-the-art model that utilized 486 different types of functional genomics data of K562, including transcription factor and histone modifications chromatin co-immunoprecipitation sequencing (ChIP-seq), cap analysis of gene expression (CAGE), and chromatin accessibility measurements, we discovered that our model made more accurate predictions overall (Pearson's r = 0.56 versus 0.44), although Enformer outperformed in peak regulatory activity predictions, which can be attributed to its nucleotide-level modeling architecture and extensive training data specifically targeting K562. Overall, Enformer's predictions tend to have larger across genome variance (**Supplementary Figure 3c**). Ablation GET also presented significant advantages in computational cost. In fact, for this comparison, we had to subsample to 2,000 elements to complete the calculation with Enformer in 3 days. While using the same amount of computing time GET can screen all 200,000 elements.

GET accurately identifies cis-regulatory elements and upstream regulators

Single-cell multiome studies enable the identification of cis-regulatory elements (CRE) in specific cell types, offering potential phenotype intervention targets. Traditional peak-to-gene workflows largely depend on correlating multiome ATAC-seq and RNA-seq counts, with regulator identification necessitating additional filtering by transcription factor (TF) expression[24–26]. Such methods are limited in comprehensiveness due to the nonlinear relationship between accessibility and transcription level and sparsity of single cell data, usually can only produce results for thousands of genes. Through model interpretation techniques (**Methods**), we can efficiently derive region/motif contribution scores for expressed genes across cell types, producing results for virtually all genes in even less abundant cell types (~1,000 cells). Focusing on Fetal erythroblasts, we leveraged published genome base-editing

data to investigate four known fetal hemoglobin regulating loci[27] (*BCL11A, NFIX, KLF1, HBG2*, where first three are transcription factors genes and *HBG2* encodes a fetal hemoglobin subunit, **Figure 3a**).

Applying GET to Fetal erythroblasts yielded interesting insights into the regulation of fetal hemoglobin. We rediscovered the central role of the GATA transcription factor, which, via its binding to an erythroid-specific enhancer, orchestrates the expression of BCL11A, a known modulator of hemoglobin regulation[28,29]. Interestingly, GET also highlighted the role of the SOX family of transcription factors in this enhancer, which were previously linked to fetal hemoglobin[30] but not known to function through this specific enhancer.

Examining all four loci—BCL11A, NFIX, KLF1, and HBG2—we benchmarked GET against established models like Enformer[8], DeepSEA[31], and Activity-by-Contact (ABC)[32,33] . Distinctly, GET outperformed these counterparts, especially in detecting long-range enhancer-promoter interactions (**Figure 3c-d, Supplementary Figure 4a-b**). We also show that while enhancer chromatin accessibility is predictive of regulatory activity for proximal enhancer-promoter relationships, its precision diminishes for long-range interactions. Alternative evaluations using different functional enhancer thresholds (top 10% or 25% mean HbFBase, the gRNA enrichment score defined in the original study[27]) reaffirm GET's precision in this scenario (**Figure 3d**).

GET is able to extract overall motif importance across cis-regulatory elements (CREs) for specific genes. For HBG2, BCL11A, and NFIX, the top motifs identified were consistent with their known transcriptional regulators or hematopoietic transcription factors (**Figure 3e**). For instance, we found significance of NFY and SOX motifs for HBG2 and the reaffirmation of KLF1's influence on BCL11A[27]. Additionally, for NFIX, GET adeptly pinpointed the involvement of TAL1, a known GATA1 binding partner and hematopoietic factor[34].

To determine downstream targets for specific regulators, we developed an *in silico* analysis, taking the GATA motif as a case study. Using the GET motif contribution matrix, we spotlighted the top 10% of genes influenced by the GATA motif. Notably, aligning with GATA1's status as a master regulator of erythroid development, the haematopoiesis biological process was enriched[35] (Enrichment P-value=7.6×10^{-4} with multiple testing correction, **Figure 3f and Methods**) within this gene set. Delving further, the identification of transcription factors within this set laid the foundation for an erythroblast-specific transcription factor regulatory framework centered on GATA. Recognized erythroid lineage transcription factors like KLF1, GATA1, TAL1,and IKZF1 were predicted to be regulated by the GATA motif, underscoring GATA's pivotal role in a multifaceted regulatory network, in line with existing literature[36].

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 receive more gradient from lineage specific transcription factors than those that are indifferent across lineages. Our findings confirmed this, as we noted substantial Spearman correlation between the motif contributions in erythroblast and the differential expression log fold changes

for several erythroblast-related transcription factors, including GATA, ZBTB7A, and MZF1 (**Figure 3g**).

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 Figure 5**). Overall, we conclude that GET possesses the ability to learn meaningful regulatory information that is naturally transferable between cell states.

Cell-type specific regulatory insights through cross-cell-type embedding with GET

Utilizing a cross-cell-type architecture, GET is configured to extract the regulatory context for genes spanning various cell types, embedding them within a shared high-dimensional space (**Methods**). In order to further visualize what GET learns from different cell types, we explored the embedding of different 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.

To investigate whether the embedding tied to regulatory grammar retains cell type-specific information, we gathered the first transformer layer's embedding for all promoters across cell types. This allows us to capture not only the motif information within the promoter but also within the cis-regulatory elements (CREs) due to the attention mechanism employed. Intriguingly, a Uniform Manifold Approximation and Projection (UMAP)[37] 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 (**Figure 4a**).

Nonetheless, when we subset the embeddings to only three specific cell types (fetal astrocyte and two fetal erythroblast subclusters), the UMAP exhibited distinct clusters for astrocytes and erythroblast genes (**Figure 4b**). This result further corroborates that GET is proficient at discerning cell-type-specific regulatory information.

Delving further into the astrocyte-specific gene cluster (cluster #2 in **Figure 4c**), we discovered that this gene set is particularly enriched in the development of the nervous system and includes astrocyte transcription factors such as NFIA[38,38,39] and GLI3[40] (**Figure 4d**). Moreover, a comparison of motif contribution across clusters revealed a higher presence of NFI motifs in the astrocyte-specific cluster (**Figure 4e**), shedding light on the unique regulatory program within astrocytes.

GET-based causal discovery identifies potential transcription factor-transcription factor interaction

Given GET's proficiency in elucidating intricate regulatory mechanisms across diverse cellular contexts, we next investigate whether it learns transcription factor-transcription factor functional interactions implicitly. Using a cell-type agnostic gene-by-motif matrix (**Methods**), we evaluated the correlation between different motif vectors (**Supplementary Figure 6a**). High correlation may represent common genomic targets between different transcription factors. Remarkably, the transcription factor pairs with correlation values in the top decile are more likely to participate in the same biological functions compared to those in the bottom decile (**Supplementary Figure 6b**, Kolmogorov–Smirnov test, P-value 6.78×10^{-82}). For example, MBD2 and MECP2, a high correlation transcription factor pair, act both as readers of DNA methylation[41,42].

We further extended our investigation of the motif-motif interactions by utilizing a causal discovery algorithm, Linear Non-Gaussian Acyclic Model (LiNGAM)[43], to derive a directed acyclic graph from the cell-type agnostic gene-by-motif matrix (**Methods**). The consequent network, displaying interactions with an absolute value greater than 0.1 for clarity in visualization, can be seen in **Supplementary Figure 6c**. Interesting, 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 Figure 7a, Methods**), 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 got similar top out-degree transcription factors (**Supplementary Figure 7b**).

Here we present four subnetworks in **Figure 4g** as examples. Notably, MBD2 and MECP2 has negative interaction with a promoter-enriched motif, GC-tract, which aligns with the well-known repressive effect of promoter methylation on gene expression[44,45]. The other three networks centered around NR/17 (Representative transcription factor: ESR1), NFKB/1 (Representative transcription factor: RELA), and ZFX exemplify the diverse information GET has learned. For example, the pair NR/17-Ebox/CACCTG highlights a functional regulatory complex ESR1-ZEB1[46]. NR/17-GLI is also supported by the known physical interaction between ESR1 and GLI3[47]. NFKB/1-Ebox/CACCTG has a strong interaction with negative effect size, while their representative transcription factor, RELA and SNAI1, has been shown to be interacting using co-immunoprecipitation[48]. ZFX is positively linked to TFAP2/1, and has been shown to co-localize with TFAP2A using ChIP-seq[49,50]. NFKB/1 and NFKB/2 are dimer motifs of NFKB family transcription factors with NFKB/1 motif specifically from NFKB1, NFKB2 and NFKB/2 motifs also contributed by REL, RELA and RELB. The strong link between these two motifs are thus expected.

To quantitatively assess the overlap with currently known physical interactions between transcription factors, we compared the GET motif-motif interaction network with STRING v11[48] database (**Methods**). Our results show a precision (true positive rate) of 5.6% by random chance. However, by selecting the top 1% (793 pairs) causal or correlation pairs from GET's predictions, we achieved precisions of 25.2% and 15.9%, respectively. This confirms the advantage of our innovative causal discovery based model interpretation approach. As a

comparison, a recent mass spectrometry-based transcription factor-transcription factor interaction study[51] reaches 30.4% precision with top 1.25% (990) pairs. This reflects the incompleteness of annotated transcription factor-transcription factor interactions and highlights the GET-predicted motif pairs as a valuable orthogonal information for this task.

A structural catalog of human transcription factor and coactivators

With the predicted causal motif interaction network predicted by GET, we next embarked on building a structural catalog of the human transcription factor interactome using Alphafold2[52]. We started by categorizing transcription factor-transcription factor interactions into several different catalogs: Direct interaction, which includes homodimer, intra-family heterodimer, or inter-family heterodimer, and Cofactor-mediated interaction, which may encompass both cooperative and competitive binding (**Figure 5a**). Starting from the most straightforward intra-family interactions, we first acquire all dimeric structure predictions of more than 1,700 known human transcription factors. To evaluate whether Alphafold2 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 (**Supplementary Figure 8a**). The accuracy of AlphaFold dimer prediction is exemplified by the perfectly aligned TFAP2A structure to experimental results[53] (**Supplementary Figure 8b**), even though there is no other similar template in PDB.

With AlphaFold2's established ability to predict unseen multimer structures, we questioned whether the disordered region in the structure could fold upon binding to partners. Based on causal discovery predicted transcription factor-transcription factor interactions, we sought to identify potential structural interactions using AlphaFold2. Taking TFAP2A and ZFX as an example, we segmented both proteins into four distinct structured or disordered domains based on predicted local distance difference test (pLDDT) (**Figure 5b**) and predicted the multimer structure of all pairwise combinations between these segments. Remarkably, the originally unstructured ZFX intrinsically disordered region (IDR) (**Figure 5c**) folded into a well-defined multimeric structure when paired with TFAP2A structured domains, mainly driven by electrostatic interactions (**Figure 5d**).

To provide another line of evidence, we employed molecular dynamics simulations (**Methods**), discovering that the monomer IDR exhibited a more collapsed structure after 100 ns (**Figure 5e**) and fewer inter-chain hydrogen bonds (**Supplementary Figure 8c**). Moreover, the per-residue pLDDT of ZFX IDR and TFAP2A in the multimer structure correlated strongly with residue instability, as measured by root mean squared distance (RMSD; **Supplementary Figure 8d**), aligning with previous studies indicating AlphaFold's implicit learning of protein folding energy functions. To validate the predicted interactions between these two proteins, we performed co-immunoprecipitation experiments. As shown in **Figure 5f**, we are able to pull down ZFX using a TFAP2A antibody.

When extending our method to negative effect pairs such as SNAI1 (Ebox/CACCTG) and RELA (NFKB/1), the absence of robust structural interactions, despite previous physical evidence, led us to explore cofactor-mediated interactions (**Figure 5i**). Both transcription factors are known to physically interact with EP300[48], and the predicted structures underscored electrostatic interactions with EP300's TAZ1 and TAZ2 domains (**Figure 5j**). This concurs with existing studies on the electrostatic binding of transcription factor IDR to EP300 TAZ domains[54–57].

Broadening our study, we applied the procedure to top 5% transcription factor pairs in each cell type (totalling 1,718 transcription factor pairs or 24,737 pairs of transcription factor segments, see **Methods**) as predicted through GET-based causal discovery and built a structural catalog of transcription factor interactions. Interestingly, the folded conformation of ZFX IDR can also be seen in other transcription factor pairs, for example EGR1 IDR-ZFX IDR (**Supplementary Figure 8e**). We also show that the previously mentioned interaction between ESR1 and ZEB1 could be driven by a confident structural interaction between the ZEB1 C-terminal IDR and ESR1 NR domain (**Supplementary Figure 8f**).

GET uncovers mechanism of germline mutation in disorder region of transcription factors

To demonstrate the utility of information provided by the GET Catalog, we performed a case study on PAX5, a driver transcription factor of B-cell precursor acute lymphoblastic leukemia (B-ALL)[58]. B-ALL is the most frequent pediatric malignancy, and somatic genetic alterations (deletions, translocations and mutations) in PAX5 occur in approximately 30% of sporadic cases[59]. While most PAX5 somatic missense mutations affect the DNA-binding domain (V26G or P80R), G183S is a recurrent familial germline mutation that confers an elevated risk of developing B-ALL[58–60]. Somatic mutation of G183 and frameshift in a nearby hotspot is also seen in B-ALL patients[61]. Although the pLDDT plot of PAX5 highlights G183 and the octapeptide domain as a small peak in the entire intrinsically disordered region, its functional role remains elusive (**Figure 6a**).

To probe this, we first explored potential interaction pairs involving PAX5 (PAX/2 motif) in fetal B lymphocytes (CXCR5+). We identified promising interactions with several transcription factors including E2F3, MZF1, MECP2, NR4A2, RFX3, and RORA (NR/3 motif, **Figure 6b**). Subsequent exhaustive segment interaction screening revealed a novel interaction between the RORA nuclear receptor (NR) domain and the octapeptide domain of PAX5 (**Figure 6c**). The G183 residue is close to the binding site alpha helix where a mutation to Serine or Valine might introduce spatial clash. This interaction was further corroborated by positive affinity purification-mass spectrometry (AP-MS) data of their paralog PAX2-NR2C2[51], as both the PAX5 octapeptide domain and NR domain are highly conserved and structurally similar across their paralogs.

To elucidate whether PAX/2 and NR/3 motif corregulate genes, we examined the top 10,000 promoters predicted to be most influenced by them. Our analysis uncovered a set of 2,570

genes commonly regulated by both, including surface markers like CD19 and CD79B, as well as known oncogenes implicated in B cell acute lymphoblastic leukemia (B-ALL) including MYC, CEBPD, LMO2, although these oncogenes are also predicted to be strongly repressed by IKZF1 (IKAROS tumor suppressor, with ZNF143 motif), and are not highly expressed (**Figure 6d**). Enrichment analysis revealed an overrepresentation of genes involved in leukocyte activation and genes affected by PAX5 perturbation during B cell differentiation, aligning with previous work on the G183S mutation[58,62–67] (**Figure 6e, Supplementary Figure 9a,b**). On the other hand, the genes that are specifically regulated by PAX/2 or NR/3 are enriched in neuronal pathways and cell cycle respectively. These results are corroborated by the sequence pattern of PAX/2 motif which contains a partial RARA/RORA motif, while another PAX5 motif, PAX/1, contains a partial LHX6 motif which is a neuronal lineage transcription factor (**Supplementary Figure 9c**).

Finally, we used patient tumor RNA-sequencing data to validate our findings. Using data from 15 B-ALL patients[68] without PAX5 somatic coding mutations, we found significant correlations ($P<0.05$) between both PAX5 and RARA/NR4A1 (paralogs of RORA with NR/3 motif) expression levels and the expression of predicted target genes (**Figure 6f, Supplementary Figure 9d**), further supporting the role of the PAX5-nuclear receptor interaction in lymphoma transcriptional programs. In sum, our analysis suggests that the PAX5 G183S germline mutation may confer B-ALL-specific risk by disrupting interactions between the PAX5 intrinsically disordered region and the nuclear receptor domains of other transcription factors, thereby leads to oncogenic transcriptional programs.

Discussion

In this study, we introduced GET, a state-of-the-art foundational model specifically engineered to decipher mechanisms governing transcriptional regulation across a wide range of human cell types. By integrating chromatin accessibility data and genomic sequence information, GET achieves a level of predictive precision comparable to experimental replicates in leave out cell types. Furthermore, GET demonstrates exceptional adaptability across an array of sequencing platforms and assay types. Notably, the model successfully identified long range regulatory elements regulating fetal hemoglobin and their associated transcription factors. Collecting regulatory information from all 213 cell types and synergizing causal interactions deduced through GET with protein structure predictions, we constructed the GET Catalog. Utilizing the PAX5 gene as a case study, we illustrated the catalog's utility in elucidating functional variants in disordered protein domains that were difficult to study. The GET Catalog is publicly accessible via https://huggingface.co/spaces/get-foundation/getdemo.

Future enhancements to GET can be envisioned through the incorporation of multiple layers of biological information, including but not limited to nucleotide-level ATAC footprints, three-dimensional chromatin architecture, and regulator expression profiles. Multiplexed nucleotide-level perturbations or randomizations will be instrumental in calibrating GET for

precise predictions of the functional impact of noncoding genetic variants. The evergrowing single-cell multi-omics datasets offer enormous potential for training GET on continuous cellular trajectories and perturbed states, thereby imparting the model with a dynamic understanding of cell state transitions. Leveraging GET as a computational framework, generative models can be developed to design megabase-scale enhancer arrays and engineer cell-type specific transcription factors or their interaction inhibitors for targeted therapeutic interventions. Collectively, GET represents a pioneering approach in cell type-specific transcriptional modeling, with broad applicability in the identification of regulatory elements, upstream regulators, and crucial transcription factor interactions.

Data availability

Precomputed regulatory inference result, preprocessed data and structure prediction can be viewed at GET website https://huggingface.co/spaces/get-foundation/GET. The full processed data and inference result will be provided in a public AWS S3 bucket.

Code availability

Code for pretraining, finetuning, data preprocessing, and results analyzing will be made available in github (https://github.com/RabadanLab, https://github.com/GET-Foundation) after publication. Pretrained model will be available on Huggingface (https://huggingface.co/get-foundation). Code for the website is available at https://huggingface.co/spaces/get-foundation/GET/tree/main.

Author contributions

X.F. and S.M. initiated the project. X.F. and R.R. conceived of the study and designed the analyses. X.F. and S.M. designed the model with advice from E.X.. X.F. and S.M. implemented the model. X.F. performed data processing. S.M. performed model training, ablation and performance analyses. X.F. performed model interpretation analysis including MPRA, regulatory elements and regulator prediction, network analysis and structural analysis. X.F. and S.M constructed the GET catalog. X.F. build the website with the help from S.M. and A.B.. A.S. and A.L. performed the experiments with advice from D.O. and T.P.. Y.L. provided suggestions and computational resources to a pilot study. A.C. provides critical suggestions to the analysis. A.F. provide critical suggestions and edit the manuscript. E.X., and R.R. supervised the study. X.F., S.M., E.X. and R.R. prepared the manuscript with input from all authors.

Acknowledgements

We gratefully acknowledge funding from NIH (R35 CA253126 to RR, P01 CA174653 to RR, R01 HL159377 to A.F. and R.R., U01 CA243073 to R.R. and T.P.) and SU2C Convergence 3.14 to RR.

Disclosure of potential conflicts of interest

R.R. is a founder of Genotwin and a member of the SAB of DiaTech and Flahy. None of these activities are related to the work described in this manuscript.

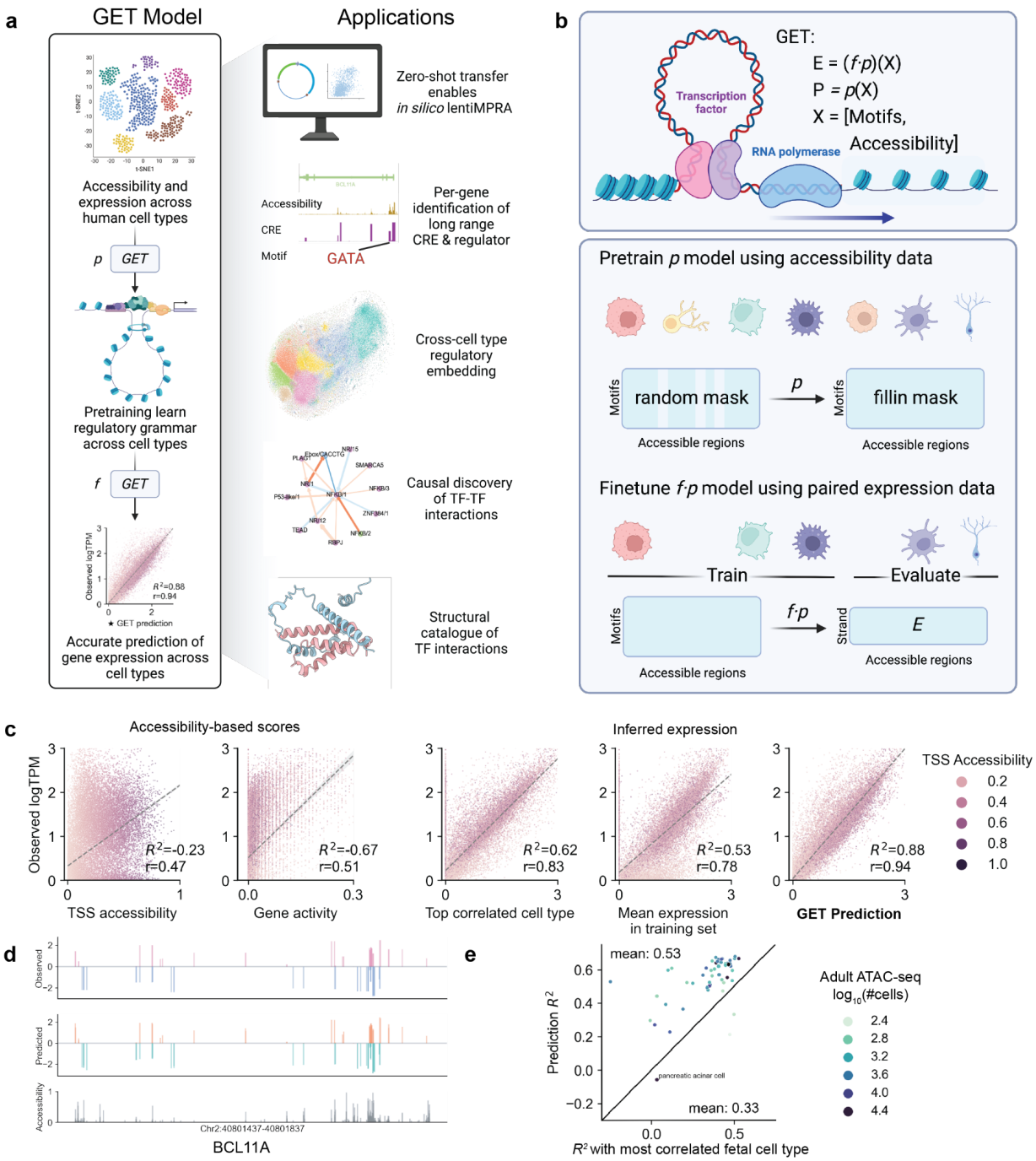

Figure 1. The GET model's universal applicability and exceptional accuracy as a foundational transcription model. a. GET derives transcriptional regulatory syntax (pretrain)

from chromatin accessibility data across hundreds of cell types, furnishing reliable predictions (finetune) of gene expression in both seen and unseen cell types. The model's broad applicability and comprehensibility allow for zero-shot prediction of lentiMPRA measurements, extensive identification of cell-type-specific regulatory elements and upstream transcription factors (transcription factors), universal embedding of regulatory information, and facilitating causal understanding of transcription factor-transcription factor interactions. **b**. Schematic illustration of training scheme of GET. The input of GET is a peak (accessible region) by transcription factor (motif) matrix derived from human single cell (sc) ATAC-seq atlas, summarizing regulatory sequence information across a genomic locus of more than 2 Mbps. Through self-supervised random mask-prediction pretrain of the input data across more than 200 cell types, GET learns transcriptional regulatory syntax (p). Finetuned on paired single cell ATAC-seq/RNA-seq data, GET learns to transform the regulatory syntax to gene expression even in leave-out cell types.(f.p). **c**. Benchmark of GET prediction performance on unseen cell types (Fetal astrocyte). Each point is a gene. Color represents normalized chromatin accessibility in TSS. Gene activity is a score widely used in modern scATAC-seq analysis pipeline[18]. Top correlated cell type is the training cell type whose observed gene expression has the largest correlation with Fetal astrocyte, in this case Fetal inhibitory neuron. Mean cell type is the mean observed gene expression across training cell types. Dash line represents linear fits. **d.** Example visualization of observed expression (top, $\log_{10}$TPM), GET prediction (mid, $\log_{10}$TPM) and chromatin accessibility (bottom, $\log_{10}$CPM) of the BCL11A locus in Fetal erythroblast. Positive (negative) values represent expression on positive (negative) strand on hg38. **e**. GET trained on fetal cell types generalize to adult cell types without retraining, outperforming most correlated celltype baseline. X axis showing R^2 score between GET prediction in adult cell types and observed expression in most similar fetal cell types. Y axis showing R^2 score between GET prediction and observed expression in the adult cell type.

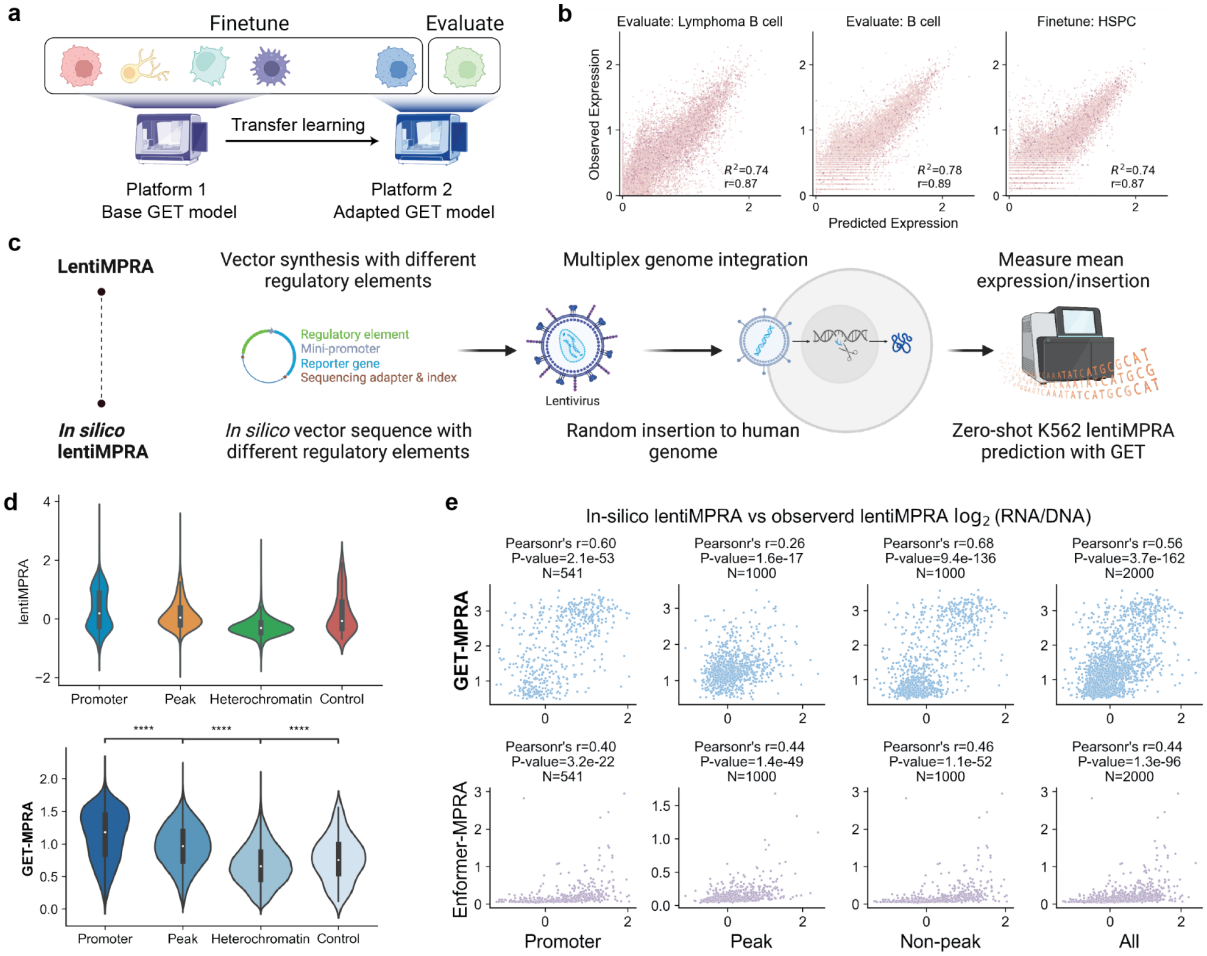

Figure 2. Transfer learning adapts GET to new platforms and measurements. a. Schematic illustration of transferring GET to lymph node 10x multiome dataset **b**. Finetuned GET accurately predicts expression in training and leave-out evaluating cell types. **c**. Schematic workflow of lentiMPRA experiments and *in silico* lentiMPRA using GET model finetuned on K562 multiome data. **d.** Readout distribution of lentiMPRA ($\log_2$RNA/DNA) and GET prediction (mean expression across genomic insertions) for different types of elements. **e.** Benchmark GET lentiMPRA prediction against Enformer on random subset of elements.

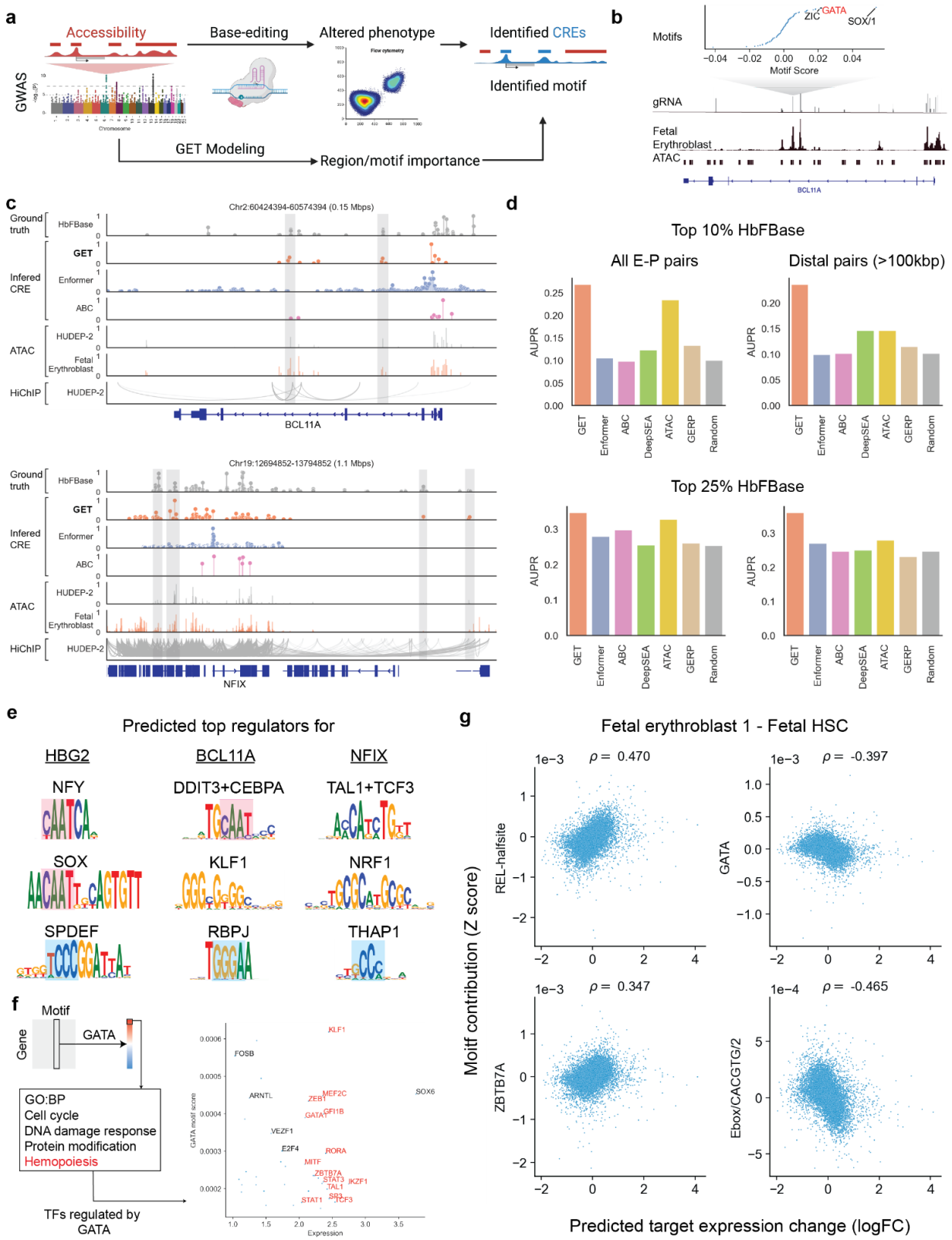
a
Accessibility
GWAS
Base-editing
Altered phenotype
Identified CREs
Identified motif
GET Modeling
Region/motif importance
b
Motifs
ZIC
GATA
SOX/1
Motif Score
gRNA
Fetal Erythroblast ATAC
BCL11A
c
Chr2:60424394-60574394 (0.15 Mbps)
Ground truth
HbFBase
Infered CRE
GET
Enformer
ABC
ATAC
HUDEP-2
Fetal Erythroblast
HiChIP
BCL11A
Chr19:12694852-13794852 (1.1 Mbps)
NFIX
d
Top 10% HbFBase
All E-P pairs
Distal pairs (>100kbp)
Top 25% HbFBase
AUPR
GET
Enformer
ABC
DeepSEA
ATAC
GERP
Random
e
Predicted top regulators for
HBG2
BCL11A
NFIX
NFY
DDIT3+CEBPA
TAL1+TCF3
SOX
KLF1
NRF1
SPDEF
RBPJ
THAP1
f
Motif
Gene
GATA
GO:BP
Cell cycle
DNA damage response
Protein modification
Hemopoiesis
TFs regulated by GATA
GATA motif score
Expression
g
Fetal erythroblast 1 - Fetal HSC
REL-halfsite
GATA
ZBTB7A
Ebox/CACGTG/2
Moitf contribution (Z score)
Predicted target expression change (logFC)

Figure 3. GET model identifies cell-type-specific regulator and cis-regulatory elements. a. Case study of identifying cis-regulatory elements (CRE) and regulators controlling a phenotype, fetal hemoglobin (HbF) level. Four genome-wide association loci (BCL11A, MYB, NFIX, HBG2) have been subjected to genome editing in a previous study, providing the labels for GET benchmark. Region/motif contribution score for each gene can be computed using GET model. **b.** GET identifies GATA motif in erythroid-specific enhancer that upregulates BCL11A, an HbF repressor. Top: motif contribution score for BCL11A expression in the erythroid-specific enhancer. Mid: gRNA enrichment score (HbFBase). Higher score means enrichment in high HbF cells, which implies these edits disturb a cis-regulatory element or regulator binding site that can upregulate BCL11A. Bottom: single cell ATAC-seq signal and peak from Fetal erythroblast. **c.** Genome track of inferred CREs for BCL11A, MYB, NFIX and HBG2. From top to bottom: HbFBase: the gRNA enrichment score from base-editing experiments. GET: GET-infered region importance score (**Methods**). Enformer: Enformer-infered region importance score. ABC: Activity-by-contact prediction collected from the original base-editing study. ATAC/HUDEP-2: Chromatin accessibility track of HUDEP-2, the erythroblast cell line used in base-editing study. ATAC/Fetal Erythroblast: Chromatin accessibility track of Fetal erythroblast, used in the training of GET. HiChIP/HUDEP-2: H3K27ac HiChIP track of HUDEP-2 cell line. **d**. Benchmark results of GET against existing methods and baselines at two HbFBase cutoffs. Left shows results for all enhancer-promoter pairs. Right shows results for only distal enhancer-promoter pairs with distance larger than 100 kbp. AUPR: area under precision and recall curve. **e**. Predicted top 3 three regulators (motifs) for BCL11A, NFIX and HBG2. Similar sequence patterns are highlighted with color shades. **f**. GATA downstream targets inferred by GET (top 10% motif score) show functional enrichment in Hemopoiesis. Scatterplot shows predicted gene expression (X-axis) and GATA-motif score (Y-axis) for GATA-targeted genes with predicted expression larger than one. All transcription factors among these genes are labeled in the plot, where those involved in Hemopoiesis are highlighted in red color. **g**. 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.

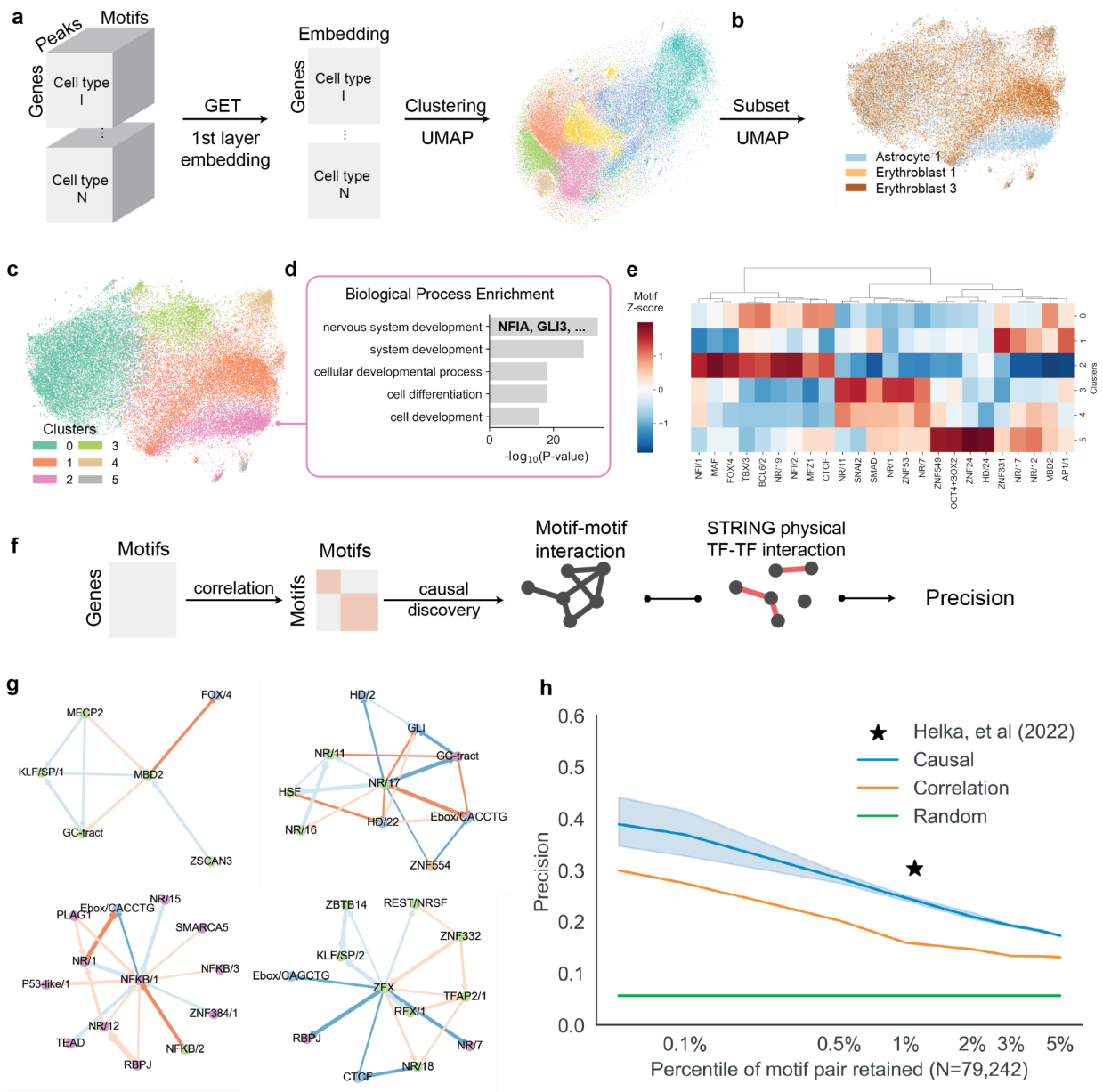

Figure 4. GET captures regulatory information across cell types and informs casual transcription factor-transcription factor interaction. a. Workflow to collect and visualize cross-cell-type regulatory embedding. **b**. The cross-cell-type regulatory embedding reveals cell-type specificity in transcription regulation. Subsampled embedding from Fetal astrocyte (blue) and two Fetal erythroblast (yellow and brown) cell types are visualized with UMAP. **c**. Louvain clustering of subsampled embedding in **b**. Note that cluster 2 is the astrocyte specific cluster. **d**. Gene ontology enrichment of genes in cluster 2. Showing astrocyte-relevant terms and astrocyte marker genes e.g. NFIA, GLI3. **e**. GET motif contribution Z-score (**Methods**, red means higher score comparing to other clusters) for each clusters. Note that cluster 2 has elevated NFI/1 and NFI/2 motifs, which correspond to the NFI family transcription factors. **f**. Causal discovery using the GET motif contribution matrix identifies transcription factor-transcription factor interaction. Physical interactions from STRING databases are used as

a benchmark to calculate the concordance. **g**. Example causal neighbor graph showing interactions (edges) between motifs (nodes). Edge weights means interaction effect size. Edge directions marks casual direction. Blue and red edge color marks negative or positive effect size. Node color marks community detected on the full causal graph. In-community edges are marked by reduced saturation. **h**. Benchmark the concordance of inferred transcription factor-transcription factor interaction using different methods with physical interactions from the STRING database. X axis marks different cutoffs of retained interaction in percentile of 79,242 total possible interactions. Y axis marks the ratio of selected interactions that is also marked as interacted in STRING. Green line marks the random selection background. Orange line marks the result of selection using motif-motif contribution score correlation. Blue line marks the causal discovery result. Shaded area marks standard error across 5 bootstraps. The star marks the result from a recent mass-spectrometry-based transcription factor-transcription factor interaction atlas[51].

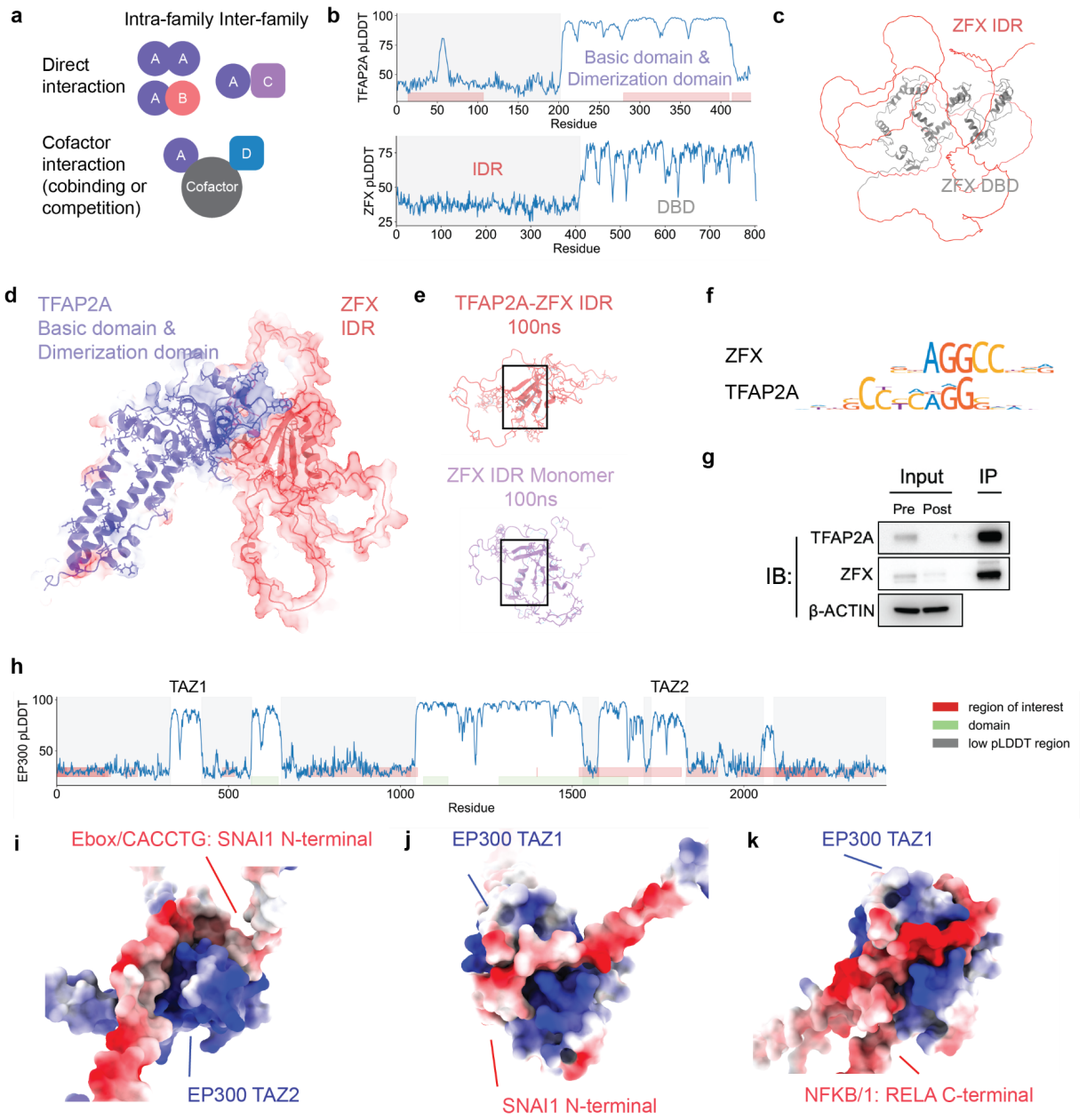

Figure 5. Structural properties of inferred transcription factor-transcription factor interactions through GET causal discovery. a. Catalogs of transcription factor-transcription factor interactions. Direct interaction includes homodimer, intra-family heterodimer or inter-family heterodimer. Cofactor-mediated interaction may include both cooperative and competitive binding. **d**. pLDDT plot for TFAP2A and ZFX, showing correspondence between high pLDDT regions and known protein domains (red rectangles). **e**. Predicted monomer structure of ZFX, showing DNA binding domain (DBD, grey) and intrinsically disordered region (IDR, red). **f**. Predicted structure of TFAP2A structured domains and ZFX IDR. Red and blue color marks negative and positive electrostatic surfaces. **g**. Molecular dynamics simulation of TFAP2A-ZFX

IDR (red) or ZFX IDR monomer (purple). Collapsed structure in ZFX IDR monomer is highlighted in rectangle. **h**. Correlation between pLDDT and residue RMSD across the simulation trajectory of ZFX IDR in the complex structure. Visualized in scatter plot (top) and line plot across the protein sequence (bottom). Yellow and blue shades in the line plot highlight beta sheets or alpha helices. **i**. pLDDT plot for EP300, highlighting TAZ1 and TAZ2 transcription factor interacting domain. Region of interest (red) and domain (green) marks annotated regions on UNIPROT. Low pLDDT regions are highlighted in gray shades. **j**. Predict structural interaction between SNAI1 N-terminal and EP300 TAZ2 domain (left), SNAI1 N-terminal and EP300 TAZ1 domain (mid), and RELA C-terminal and EP300 TAZ1 domain (right). Red and blue color marks negative and positive electrostatic surfaces.

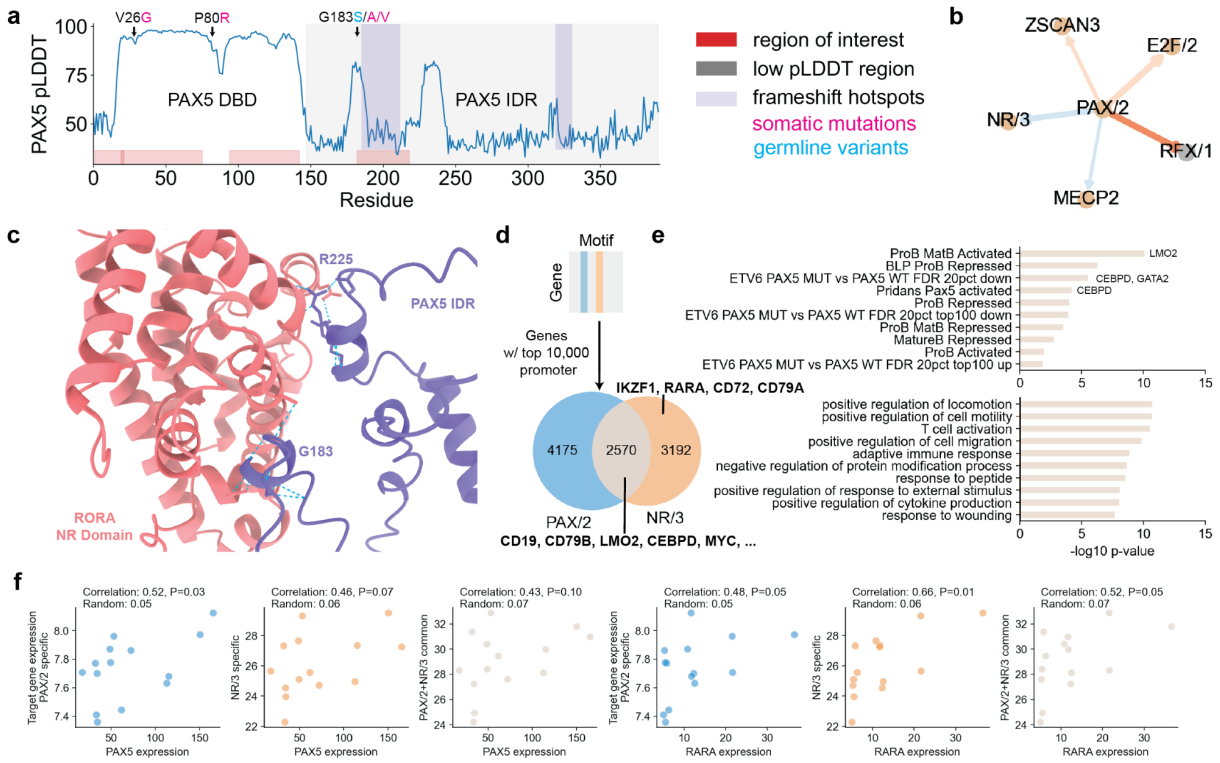

Figure 6. GET identifies a cell type specific transcription factor-transcription factor interaction affected by a cancer-associated germline variant. a. pLDDT plot for PAX5. Showing three mutational hotspots: V26G, P80R, G183S/V/A and two frameshift hotspots[61]. Region annotations from UNIPROT are shown in the figure as 'region of interest'. **b**. B-cell specific motif interactions of PAX/2. PAX5 is the highest expressed transcription factor with PAX/2 motif. RORA is the highest expressed transcription factor with the NR/3 motif. Color scheme follows Figure 4**g**. **c.** Predicted multimer structure of PAX5 IDR and RORA NR domain. Showing contacts around G183 and R225. **d.** Venn diagram of identified PAX/2 and NR/3 specific and common regulatory targets using GET gene-by-motif importance matrix. **e.** Enrichment analysis using B-cell associated gene sets in Shah et al.[58] (top) and biological process gene ontology gene sets (bottom). Results for the PAX/2-NR/3 common genes are shown in this figure. Results for PAX/2 or NR/3 specific genes are shown in Supplementary Figure 9. **f.** Spearman correlation between PAX5 (PAX/2), RARA (NR/3) and the average expression of PAX/2-specific (blue), NR/3 specific (orange) and common (light brown) target genes in B-ALL patients without PAX5 somatic coding mutations.

References

1. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nat Biotechnol* **35**, 732–746 (2017).
2. Richter, W. F., Nayak, S., Iwasa, J. & Taatjes, D. J. The Mediator complex as a master regulator of transcription by RNA polymerase II. *Nat Rev Mol Cell Biol* **23**, 732–749 (2022).
3. Malik, S. & Roeder, R. G. Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nat Rev Genet* 1–16 (2023) doi:10.1038/s41576-023-00630-9.
4. Wang, H., Schilbach, S., Ninov, M., Urlaub, H. & Cramer, P. Structures of transcription preinitiation complex engaged with the +1 nucleosome. *Nat Struct Mol Biol* **30**, 226–232 (2023).
5. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
6. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics* **50**, 1171–1179 (2018).
7. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology* **16**, e1008050 (2020).
8. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).
9. OpenAI. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
10. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
12. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, (2020).

13. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e19 (2021).

14. Joung, J. *et al.* A transcription factor atlas of directed differentiation. *Cell* **186**, 209-229.e26 (2023).

15. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).

16. Consortium*, T. T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* (2022) doi:10.1126/science.abl4896.

17. Li, J. *et al.* Conservation and divergence of vulnerability and responses to stressors between human and mouse astrocytes. *Nat Commun* **12**, 3958 (2021).

18. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403–411 (2021).

19. 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.

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

21. Chen, A. F. *et al.* NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* **19**, 547–553 (2022).

22. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat Protoc* **15**, 2387–2412 (2020).

23. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. 2023.03.05.531189 Preprint at https://doi.org/10.1101/2023.03.05.531189 (2023).

24. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat Methods* 1–11 (2023) doi:10.1038/s41592-023-01971-3.

25. Kamimoto, K. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* **614**, 742–751 (2023).

26. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 1–13 (2023) doi:10.1038/s41592-023-01938-4.

27. Cheng, L. *et al.* Single-nucleotide-level mapping of DNA regulatory elements that control fetal hemoglobin expression. *Nat Genet* **53**, 869–880 (2021).

28. Basak, A. & Sankaran, V. G. Regulation of the fetal hemoglobin silencing factor BCL11A. *Ann N Y Acad Sci* **1368**, 25–30 (2016).

29. Martyn, G. E. *et al.* Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. *Nature Genetics* **50**, 498–503 (2018).

30. Listì, F. *et al.* Study on the Role of Polymorphisms of the SOX-6 and MYB Genes and Fetal Hemoglobin Levels in Sicilian Patients with β-Thalassemia and Sickle Cell Disease. *Hemoglobin* **42**, 103–107 (2018).

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

32. Fulco, C. P. *et al.* Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).

33. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

34. Wu, W. *et al.* Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res* **24**, 1945–1962 (2014).

35. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191–W198 (2019).

36. Welch, J. J. *et al.* Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood* **104**, 3136–3147 (2004).

37. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).

38. Kang, P. *et al.* Sox9 and NFIA Coordinate a Transcriptional Regulatory Cascade during the Initiation of Gliogenesis. *Neuron* **74**, 79–94 (2012).

39. 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).

40. 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).

41. Baubec, T., Ivánek, R., Lienert, F. & Schübeler, D. Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. *Cell* **153**, 480–492 (2013).

42. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).

43. Shimizu, S., Hoyer, P. O., Hyv&#228, A., rinen & Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* **7**, 2003–2030 (2006).

44. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).

45. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**, 23–38 (2013).

46. 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).

47. 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).

48. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2019).

49. Kenny, C. *et al.* TFAP2 paralogs facilitate chromatin access for MItranscription factor at pigmentation and cell proliferation genes. *PLOS Genetics* **18**, e1010207 (2022).

50. 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.

51. Göös, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022).

52. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021) doi:10.1038/s41586-021-03819-2.

53. 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.

54. De Guzman, R. N., Wojciak, J. M., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. CBP/p300 TAZ1 domain forms a structured scaffold for ligand binding. *Biochemistry* **44**, 490–497 (2005).

55. Miller Jenkins, L. M. *et al.* Characterization of the p300 Taz2-p53 TAD2 complex and comparison with the p300 Taz2-p53 TAD1 complex. *Biochemistry* **54**, 2001–2010 (2015).

56. Lochhead, M. R. *et al.* Structural insights into TAZ2 domain-mediated CBP/p300 recruitment by transactivation domain 1 of the lymphopoietic transcription factor E2A. *J Biol Chem* **295**, 4303–4315 (2020).

57. Ferrie, J. J. *et al.* p300 Is an Obligate Integrator of Combinatorial Transcription Factor Inputs. *bioRxiv* 2023.05.18.541220 (2023) doi:10.1101/2023.05.18.541220.

58. Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell

acute lymphoblastic leukemia. *Nat Genet* **45**, 1226–1231 (2013).

59. Escudero, A. *et al.* Clinical and immunophenotypic characteristics of familial leukemia predisposition caused by PAX5 germline variants. *Leukemia* **36**, 2338–2342 (2022).

60. Auer, F. *et al.* Familial Predisposition to B-Cell Precursor Acute Lymphoblastic Leukemia Mediated By PAX5 Germline Variants. *Blood* **140**, 8888–8889 (2022).

61. Gu, Z. *et al.* PAX5-driven Subtypes of B-progenitor Acute Lymphoblastic Leukemia. *Nat Genet* **51**, 296–307 (2019).

62. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).

63. Pridans, C. *et al.* Identification of Pax5 Target Genes in Early B Cell Differentiation1. *The Journal of Immunology* **180**, 1719–1728 (2008).

64. Revilla-i-Domingo, R. *et al.* The B-cell identity factor Pax5 regulates distinct transcriptional programmes in early and late B lymphopoiesis. *The EMBO Journal* **31**, 3130–3146 (2012).

65. Delogu, A. *et al.* Gene Repression by Pax5 in B Cells Is Essential for Blood Cell Homeostasis and Is Reversed in Plasma Cells. *Immunity* **24**, 269–281 (2006).

66. Schebesta, A. *et al.* Transcription Factor Pax5 Activates the Chromatin of Key Genes Involved in B Cell Signaling, Adhesion, Migration, and Immune Function. *Immunity* **27**, 49–63 (2007).

67. Holmfeldt, L. *et al.* The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet* **45**, 242–252 (2013).

68. Oshima, K. *et al.* Mutational and functional genetics mapping of chemotherapy resistance mechanisms in relapsed acute lymphoblastic leukemia. *Nat Cancer* **1**, 1113–1127 (2020).

Methods

ATAC-seq Data processing

Pseudobulk

To get the chromatin accessibility score for each region, we used the scATAC-seq count table and cell annotation table from each studies. To group single cells into pseudobulk "cell types", we used the louvain clustering result from each study. The cell type annotation of each cluster is used to define the biological cell type. We empirically used cell count > 600 as a threshold to select cell clusters with enough sequencing depth. We have compiled a table of pseudobulk cell types used in the training process below:

Cell-type-specific accessible region identification

For the identification of cell-type-specific accessible regions, the peak calling results from the original studies of each dataset were followed to obtain a union set of peaks. Subsequently, to compile a list of accessible regions specific to each cell type, peaks with no counts were filtered out.

In the context of the human fetal and adult chromatin accessibility atlas, we employed the peak set produced by Kai Zhang et al.[1], incorporating the fetal chromatin accessibility atlas originally published by Silvia Domcke et al.[2]. We have also trained a version of fetal-only GET model using the original peak calling and cell type annotation from Silvia Domcke et al., resulting in comparable expression prediction and regulatory analysis performance. For the 10x multiome data, we used the provided peak fragment count matrix. For the K562 NEAT-seq and bulk chromatin accessibility data, a more permissive version of peaks was called using MACS2[3], and different logTPM cutoffs were applied to the resulting peak set to select accessible regions. This accessibility-based data augmentation enhances the diversity of input data and fine-tunes the GET model for data from a single cell type. The code for processing is publicly available in our Github repository at atac rna data processing.

Accessibility features

In our study, the chromatin accessibility score for a specific genomic region is defined by the count of fragments located within that region for a given cell type pseudobulk. To enhance the model's generalizability, these counts are further normalized through the logTPM (Log Transcripts Per Million) procedure. Specifically, let t be the total fragment count in a pseudobulk, and c_i be the fragment count in region i. Then, the accessibility score s_i is computed as:

$$s_i = \log_{10}\left(\frac{c_i}{t} + 1\right), \quad t = \sum_i c_i$$

For the majority of the regulatory analysis, the 'Without-ATAC' version of the GET (Gene Expression Tracking) model is utilized to comprehensively evaluate the regulatory influences exerted by transcription factors. In both the training and inference phases of this specific model version, the accessibility scores are uniformly set to 1 if the region is identified as a chromatin accessibility peak. This equates to assuming binary chromatin accessibility states within the studied scenario.

Motif features

To calculate the motif binding score within a specific genomic region, the corresponding sequence is scanned against the hg38 reference genome. This procedure involves utilizing 2,179 transcription factor motif position-weighted matrices (PWMs), as previously compiled by Jeff et al.[4], accessible at Vierstra's resources. For the scanning process, the MOODS tool is used with default threshold[5].

More specifically, to represent sequence information while mitigating feature redundancy, a specialized motif scoring process is implemented. Building on Jeff's prior research, these 2,179 motifs are categorized into 282 motif clusters, a classification determined by PWM similarity. By using this established clustering definition, nucleotide-level motif matches that are redundant are eliminated, retaining only the match with the highest score within overlapping matches belonging to the same motif cluster.

Subsequently, the scores of all non-overlapping motif matches within each motif cluster are summed, yielding one cumulative score for each of the 282 clusters. As a final step, motif binding scores for all regions within a given cell type are determined and subjected to min-max normalization across regions. This normalization facilitates model generalization and the training process, ensuring that each motif cluster's score is processed in a standardized manner.

Input data

GET is designed to capture the interaction between different regions and regulators. To facilitate that, we need each input sample to contain a certain number of consecutive accesible regions, mimicking the "reception field" of an RNA Polymerase II. Through previous experiment we found that ideally the equivalent genome coverage should be around or more than 2 Mbp, a range where most of the chromatin contact happens. As a result, based on our current data prepossessing pipeline we choose to

use 200 as the input region numbers for one training sample. We acquired non-overlapping samples from the genome to use as our pretrain input, and from only the training chromosomes to use as our finetune input.

GET model is engineered to encapsulate the interactions between neighboring regions. To achieve this, it is essential for each input sample to encompass a specific number of consecutive accessible regions, simulating the "reception field" of an RNA Polymerase II. Through empirical research, we determined that an optimal equivalent genome coverage for this purpose is approximately 2 Mbp or greater, a span within which the majority of chromatin contacts occur. Consequently, in line with our existing data preprocessing pipeline, we selected 200 as the quantity of input regions for a single training sample. Non-overlapping samples were extracted from the genome for pretraining, and exclusively from training chromosomes for the finetuning phase.

RNA-seq Data Processing

Cell Type Matching

For experiments encompassing multiomics, the correspondence between accessibility and expression is inherently determined through cell barcodes. In pseudobulk cases, where accessibility and expression are assessed independently, cell type annotations are utilized to facilitate the mapping. Specifically, the fetal expression atlas from Cao et al.[6] is employed for fetal cell types, while adult data is extracted from Tabula Sapiens[7]. When several ATAC pseudobulk share the same cell type annotation, identical expression labels are assigned. This approach, while a compromise, is necessitated by the current dearth of multiome sequencing data, a situation expected to change dramatically in the near future.

Expression Values

Expression values are allocated to each region within our input. Constrained by poly-A scRNA-seq, only aggregated mRNA levels can be captured, resulting in values that are not reflective of the nascent transcription rate more closely tied to regulatory events. Nonetheless, these values furnish valuable cell-type-specific information. The process begins by intersecting the input region list with Gencode V40 transcripts annotation to pinpoint promoters, followed by the assignment of log count per million (CPM) values to regions corresponding to these promoters. All remaining regions are assigned a value of 0. Although this does not perfectly represent all transcription events happening in a cell, we believe the zero label on non-promoter region helps in delivering informative negative labels to the model.

Input Target

In alignment with the 200×283 input matrix, the target input is a 200×2 matrix, symbolizing the transcription levels of the corresponding 200 regions across both positive and negative strands.

Model architecture

The GET architecture consists of three parts: 1) A regulatory element (RE) embedding layer, 2) 12 RE-wise attention layers, and 3) a linear layer as the expression prediction head (Supplementary Figure 1).

Our GET takes 200 regulatory elements, each with 282 motif binding scores and optionally one accessibility score as a sample as the input. As a result, the input is a 200×283 matrix. When we choose to not using the quantitative accessibility score, we set the 283-th column to 1.

Then we feed the sample into the RE embedding layer to generate the regulatory element embedding with a dimension of 768 for each peak. Since we do not want to lose information in the input of the original regulatory element, we apply a linear layer to capture the general information in the different classes of transcription factor binding sites. To learn the cis- and trans-interactions between regulatory elements and transcription factors, we apply 12 RE-wise Attention (REA) layers with a multi-head attention mechanism on the RE embeddings along the regulatory element.

Suppose N_h, d_v, d_k denote the number of heads, the depth of values, and the depth of keys. The output from each head h is computed as

$$O_h = Softmax\left(\frac{X'W_q(X'W_k)^T}{\sqrt{d_k}}\right)(X'W_v) \tag{1}$$

where $W_q, W_k \in \mathbb{R}^{(n\times D)\times d_k}, W_v \in \mathbb{R}^{(n\times D)\times d_v}$ are learnable linear transformations.

Then we concatenated the output from each head h for the RE-wise Attention block. The Layer Normalization (LN), Feed-forward Network (FFN), and Residual Connections are finally utilized to generate the output for each layer. Thus, the mechanism behind the RE-wise attention block is summarized as:

$$\mathbf{z}'_l = \mathrm{MHA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}; \mathbf{z}_l = \mathrm{FFN}(\mathrm{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l \tag{2}$$

where $\mathbf{z}'_l, \mathbf{z}_{l-1})$ denote the intermediate representation in the block l and the output from the block $l-1$. We apply two linear layers with a GELU[8] activation layer in the FFN layer.

The GET architecture is similar to the state-of-the-art model Enformer[9]. However, the following changes helped us improve and exceed its performance: GET uses the regulatory element (RE) embedding layer to capture the general information of regulatory elements in the different classes of transcription factor binding sites. Moreover, a masked regulatory element mechanism was utilized to learn the general cis- and trans-interactions between regulatory elements and transcription factors from different kinds of human cell types.

Specifically, a random set of positions was uniformly selected to mask out $\mathbf{M} = \{\mathbf{m}_i\}_{i=1}^{k}$ with a mask ratio of $r = k/n$. We replaced the regions in the selected positions with a $[\texttt{MASK}]$ regulatory element, and the masked input regulatory element is denoted as $X^{\text{masked}} = (X, \mathbf{M}, [\texttt{MASK}])$. where $X = \{\mathbf{x}_i\}_{i=1}^{n}$ is the input sample with n regulatory elements. The training goal is to predict the original values of the masked elements $\mathbf{M}$. Specifically, we take masked regulatory element embeddings X^{masked} as input to our GET, while a simple linear layer is appended as the prediction head. Therefore, the overall objective of self-supervised training is formulated as:

$$\mathscr{L} = \mathbb{E}\left(\sum_{i \in \mathbf{M}} -\log p(\mathbf{x}_i | X^{\text{masked}})\right) \tag{3}$$

where $\mathbf{x}_i$ denote the masked region to be predicted.

Training scheme

We conduct pre-training in the large-scale single-cell Chromatin accessibility data. Then we fine-tune the pre-trained model on the Paired chromatin accessibility-gene expression data with the same Poisson negative log-likelihood loss function as Enformer[9]. Expression values are represented as normalized transcript per million (TPM). We then match the cell types between RNA and ATAC datasets by annotating cell type names and ignoring those that cannot be matched. To improve training stability, we log-transform the expression values as $\log_{10}$(TPM+1). We then map the gene expression to accessible regions using the following approach: if a region overlaps with a gene's transcription start site (TSS), the gene's expression value is assigned to that region as a label; if a region overlaps with multiple gene's TSS, the expression values of the corresponding gene are summed up and used as the label of that region; if a region does not overlap with any TSS, the corresponding expression label is set as 0. Finally, each regulatory element is assigned to an expression target value.

The GET implementation is based on PyTorch[10] framework. For the first training stage, we applied AdamW[11] as our optimizer with a weight decay of 0.05 and a batch size of 256. The model was trained for 800 epochs with 40 warmup epochs for linear learning rate scaling. We set the maximum learning rate to 1.5e-4. For the second fine-tuning stage, we used AdamW[11] as our optimizer with a weight decay of 0.05 and a batch size of 256. The model was trained for 100 epochs.

Model evaluation

We use pearson correlation, spearman correlation, and R^2 to evaluate the prediction performance in all settings. For evaluation of cell type specific gene prediction, we compare the observed and predicted log fold change between two cell types using the same metrics.

Cross-cell-type expression prediction

In cross-cell-type prediction setting, we pretrained on ATAC-seq data from all cell types, and finetuned with expression data from only training cell types, hiding the evaluation cell type expression label from the model. On average, GET achieves 0.799 peasrson s r and 0.845 spearman correlation across different leave-out-cell type settings. Furthermore, GET is able to get similar performance when chromosomes (chr4,chr14) are also leaved out (cell-type & chromosome leave-out, Spearman rho: 0.938, R2: 0.868, Pearson's r: 0.935).

Platform transfer prediction

In order to transfer to a new sequencing platform, there are multitude of domain shift that need to be addressed. This including but not limit to: 1. Sequencing depth: as lower depth will lead to less captured peaks. It will also affect the signal to noise ratio in the accessibility quantification; 2. Peak calling threshold and software; 3. Technical bias due to different library constructing and sequencing method; and 4. biological differences

Due to these biases, it's hard to directly apply a model trained on one dataset to a new platform without finetuning. Thus, for a new dataset with multiple cell type avaliable, we took a leave-out cell type approach of finetuning. For a dataset of sorted cell types where only one cell type is avaliable, we used leave-out chromosome training.

LentiMPRA zeroshot prediction

The experimental procedure involves designing a library of lentivirus vector which contains both desired sequence elements and a minipromoter; then the vector will be randomly inserted on the genome through viral infection; the regulatory activity is then measure through sequencing and counting the log copy number of transcribed RNAs and integrated DNA copies. To

simulate this approach using GET, we first collect the sequence element library and constructed the vector sequence with mini promoter. We then follows the same data preprocessing procedure to get the motif score of the inserted elements. For each element, we perform 'in silico insertion' by sum up its motif score with a existing region on the genome. The +/- 100 regions centered around the insertion region where then used as a input sample for GET to make expression prediction. The mean predicted expression ($\log_{10} TPM$ were multiplied with the predicted accessibility (using a GET model finetuned to predict ATAC logTPMs) as the predicted regulatory activity. For each region, we perform 600 insertion across the genome to match with the experimental insertion count. We used the GET model finetuned on K562 multiome and bulk ATAC- and RNA-sequencing data to perform the inference. For Enformer, we performed the same analysis, with the only difference is that we integrate the vector sequence to a random position on the genome and collected a 196,608 bp sequence centered around the insertion site. Enformer is trained on 5,313 human epigenome track, with 486 experiments specifically for K562. To compute the regulatory activity, we selected the output from the K562 CAGE track, which is a quantitative and nucleotide-level map of 5' of transcripts. Following the practice of the original study, we used the average output of the 3 bins in the center of sequence as the predicted expression for a sample. Each elements were also inserted into 600 random genome locations to compute the final averaged regulatory activity. We were only able to perform this experiments for 1,000 enhancers and 1,000 non-enhancer elements due to the time complexity of Enformer inference. The comparison with GET is performed on the same set of elements.

Model interpretation and analysis

Calculation of jacobian matrix

We used multiple feature attribution methods in different analysis.

The gradient of the model's output with respect to the input features, represented by the vector $\nabla f(\mathbf{x})$, tells that how much the model output (Expression) will change when we change a small amount of the input along a dimension (e.g. a certain motif in a cisregulatory region).

The generalization to multiple outputs in the context of neural network feature attribution extends to the Jacobian matrix:

$$\mathbf{J}_{i,j} = \frac{\partial f_i}{\partial x_j}, \tag{4}$$

where f_i is the i-th output, representing the transcription level on either the positive or negative strand, and x_j is the j-th input feature, comprising scanned and summarized binding scores for 282 TF motif clusters, and an additional dimension for accessibility scores.

For our specific analysis, the input is a region-by-feature matrix of dimensions $(200, 283)$, including 282 features for TF motif clusters and 1 for accessibility scores. Two models are considered:

- **With-ATAC Model**: Accessibility scores are set to the normalized TPM of Tn5 insertion count in the given accessible region.
- **Without-ATAC Model**: Accessibility scores for all regions are set to 1, focusing solely on chromatin accessibility peaks.

The corresponding output is a region-by-strand matrix of dimensions $(200, 2)$, capturing the transcription levels on both positive and negative strands. This formulation enables the computation of the Jacobian matrix, vital for understanding the influence of individual features on the transcription levels.

Integrated Gradients (IG) provides importance scores by approximating the integral of gradients along a path from given baselines to inputs. The mathematical formulation is:

$$\mathrm{IG}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_{\text{baseline}}) \cdot \int_0^1 \nabla f(\mathbf{x}_{\text{baseline}} + \alpha(\mathbf{x} - \mathbf{x}_{\text{baseline}}))\, d\alpha, \tag{5}$$

where $\mathbf{x}_{\text{baseline}}$ is the baseline input[?].

DeepLIFT (Deep Learning Important FeaTures) attributes the difference in activation to each input feature, based on a reference input.

These methodologies offer unique perspectives on feature importance, with choices guided by computational efficiency, granularity, and the specific modeling context.

Identifying important regions and regulators

We first gather inference samples across the genome by producing 200-region windows that centered around each genes promoter. Given a specific gene g on strand $s \in \{0, 1\}$, the expression value can be inferred using the General Expression

Method	Conceptual Overview
Gradients/Jacobians	Provides a first-order approximation of feature influence on the output.
Integrated Gradients[12]	Approximates the integral of gradients along a path, providing smoother and more detailed attribution, with baseline comparisons.
DeepLIFT[13]	Focuses on differences in activation, running faster than Integrated Gradients, with support for specific non-linear activations.

Table 1. Summary of Feature Attribution Approaches

Transformer (GET) model f applied to an input matrix $\mathbf{X} \in \mathbb{R}^{r \times m}$, where r denotes the number of regions, and m includes motifs and optionally accessibility features:

$$\mathbf{E} = f(\mathbf{X}) \tag{6}$$

$$E_g = \mathbf{E}[r//2, s] \tag{7}$$

Where $[\cdot,\cdot]$ is the indexing operator, s is the strand of the gene.

The jacobian matrix (tensor) $\mathbf{J}_X \in R^{r \times 2 \times r \times m}$ of f at the point $(\mathbf{E}, \mathbf{X})$ evaluates how each output dimension will change when each input dimension changes a small quantity. We specifically pick the output dimension and strand that correspond to the given gene, represented as $\nabla g \in \mathbb{R}^{r \times m}$:

$$\nabla \mathbf{g} = \mathbf{J}_X[r//2, s] \tag{8}$$

$$\mathbf{J_X} = \frac{\partial \mathbf{E}}{\partial \mathbf{X}} \tag{9}$$

The feature (motif) importance vector $v_g \in \mathbb{R}^m$ is obtained by multiplying the gradient element-wise with the original input and summarizing across regions:

$$v_g = \sum_{i=1}^{r} (\nabla \mathbf{g} \odot \mathbf{X})[i,:] \tag{10}$$

where $\odot$ signifies the element-wise or Hadamard product. Since the gene-by-motif matrix is mostly used for feature-feature interaction analysis, we use the $\mathbf{X}$ with quantitative ATAC signal even when we infer $\mathbf{J_X}$ using a 'Without-ATAC' model. This helps us to study the relationship between regulators and observed chromatin accessibility.

The cell type c specific genome-wide gene-by-motif matrix $\mathbf{V_c}$ is acquired by concatenating the v_g across the genome. And the same process can be applied to different cell types.

Similarly, the region importance vector $l_g \in \mathbb{R}^r$ is given by:

$$l_g = \sum_{j=1}^{m} (\nabla \mathbf{g} \odot \mathbf{X})[:, j] \tag{11}$$

Regulator top targets

Based on the gene-by-motif matrix V_c, we can choose a TF/motif (in our case, GATA) and ask what genes will be mostly affected by this TF by identify the largest entries in the motif column. We choosed top 1,000 genes and performed gene ontology enrichment analysis using go:Profiler with the default "g_SCS" multiple hypothesis testing correction. To avoid general terms we filtered the result with term size (gene number in a term definition) larger than 500 and smaller than 1000. Terms with adjusted P-value smaller than 0.05 are retained as significant terms. We further selected TFs in the "Hemopoiesis" term with expression $\log_{10}$TPM > 1 for visualization against 'GATA' motif score.

Transcription Factor and Target Gene Correlation Analysis in Fetal Cell Types

In this analysis, we sought to elucidate the relationship between transcription factors (TFs) and their target gene expression across different cell types. Gene-by-motif files were aggregated and organized into a unified structure comprising genes, motifs, and corresponding cell features. We identified the target genes for each TF within predefined motif clusters, and computed the mean expressions of both the target genes and the corresponding TFs. To avoid potential artifacts caused by experimental batch effect in expression measurement, we performed the analysis both in adult+fetal cell types and also in only fetal cell types and get similar results.

The relationship was assessed using Spearman correlation analysis, with the results visualized through scatter plots. The x-axis represented the mean expression of the target genes, and the y-axis represented the mean expression of the TFs. Each plot was annotated with the Spearman correlation coefficient and the associated p-value, providing a statistical assessment of the correlation.

The analysis was performed iteratively for all TFs within the motif clusters specific to fetal cell types. The correlation coefficients and p-values were compiled, and the visualizations were saved as individual image files. This comprehensive view of the relationship between TFs and their target genes offers valuable insights into the regulatory dynamics within the context of fetal development.

Regulatory embedding

We collected the embedding of each gene after each transformer block of GET. For a gene g, it's embedding is defined as the embedding vector of the promoter in the output of i-th block. The embedding contains not only promoter information but also information from surrounding regions owing to the attention mechanism. In general, the deeper the layer, the more its space is dominated by the expression output. UMAP[14] was used to visualize the embedding. Louvain clustering was performed on the embedding space to colorize the UMAP visualization. Resolution is arbitrarily chosen to keep the cluster number around 10 and close to the UMAP density.

We computed the embedding in two different settings: cell type specific setting, where each dot is a gene embedding from a specific cell; cell type agnostic setting, where each dot is a gene embedding randomly sampled from all cell types. 50,000 embedding is sampled in the second case to make UMAP computation feasible.

Causal discovery of regulator interaction

We performed pairwise Spearman correlation using the gene-by-motif matrix also in both cell type specific and agnostic settings. Input*gradient score were used to constructed the matrix for comuputational efficiency. For the cell type agnostic settings, all genes with their promoter overlaps with a open chromatin peaks from all cell types are involved in the correlation calculation. Causal discovery was performed on the gene-by-motif matrix using LiNGAM[15]. For the cell type agnostic settings, 50,000 genes were randomly sampled from all cell types, the resulting matrix were subjected to LiNGAM algorithm implemented in the Causal Discovery Toolbox python package.

To benchmark the predicted causal edges in the cell type agnostic setting, we downloaded known physical interaction subnetwork from STRING V11 database[16] and kept interactions with a combined score larger than 400 as the ground truth label. Since the pairs predicted by GET is on motif cluster level, we mapped the physical interactions between TFs onto the motif clusters based on the motif cluster annotation. The resulting motif-motif physical interaction network were then compared with our prediction to calculate the precision. We also downloaded and compiled all significant interactions determined by mass spectroscopy from the Human Transcription Factor protein interaction network[17] and mapped them also to motif-motif interactions for comparison.

For our TF interaction database, we performed the LiNGAM analysis using cell-type-specific gene-by-motif table. Interactions with top 5% absolute effect size are retained in the final database. For each interaction, we performed structural analysis between the two TFs with highest expression in the corresponding cell types.

Structural analysis

Alphafold benchmark on intra-family binder prediction

We classify a TF as a intra-family binder if any of its member TFs have a known physical interaction annotated in STRING V11 database. Based on the hypothesis that if a TF can bind as a heterodimer, due to sequence and structure similarity they should also have the potential of binding as a homodimer, although the dimerization affinity might be different. We thus used AlphaFold to predict the 'hypothetical homodimer structure' of all known TFs, and try to predict whether a TF could be a intra-family binder based on various AlphaFold-based metrics. Among several different Alphafold-based metrics, including `mean_plddt` (average predicted Local Distance Difference Test score across all residues), `pAE` (predicted Aligned Error across all inter-chain interactions), `pDockQ` (predicted DockQ metric using interface pLDDT), and `pDockQ × pAE`. We found that `pDockQ× pAE` led to the best AUROC (0.69) and AUPR (0.41) when classifying intrafamily binder TFs.

Protein sequence segmentation

pLDDT from Alphafold is a reliable protein domain caller due to its accurate structure prediction performance. We segment each TF protein sequence to low and high pLDDT regions. Empirically, we found that 80% (recall) of known DNA-binding domains can be easily identified using high pLDDT regions plus high ratio of positive charged residues. More specifically, we first computed smoothed pLDDT using a 10 aa moving-average kernel and then normalize the score by dividing the max. After that, any region that has a smoothed pLDDT score less than 0.6 is defined as a low pLDDT region. If two low pLDDT regions are close (<30 aa) they will be merged as one. Any region that is not a low pLDDT region will be labeled as a high-pLDDT region.

Multimer structure prediction

LocalColabFold and ColabFold is used to predict multimer structure with AlphaFold Multimer v2.3 model. For homodimer prediction, we used all 5 models with 3 recycles. For our large scale interaction screening, we used model 3 and 3 recycles for each prediction. Predicted aligned error (PAE) and predicted LDDT were stored for downstream analysis. pDockQ were calculated following code from FoldDock[18].

For the large score interaction screening, we performed exhaustive multimer prediction between all possible low/high-pLDDT segment pairs of the two protein in a pair. We then compare the new pLDDT of each segment in the multimer structure with their original pLDDT in the monomer or homodimer structure. If a segment showing higher

Molecular dynamics simulation

The initial configuration was prepared from the Alphafold predicted PDB file. The Amber99SB-dispersion (a99SBdisp) force field was employed for system parameterization. A cubic simulation box was defined with a box size of 1 nm. Subsequently, the system was solvated using the TIP4P water model through the solvate module. To neutralize the system and generate physiological ion concentrations, sodium (Na+) and chloride (Cl-) ions were added using the genion module. The energy minimization terminates upon reaching a maximum force below 1000.0 kJ/mol/nm. Each minimization iteration utilizes a step size of 0.01 and is configured to run for a maximum of 50,000 steps. The system was then equilibrated in two steps: first in the NVT (Constant Number, Volume, Temperature) ensemble and then in the NPT (Constant Number, Pressure, Temperature) ensemble for 100 ps of simulation time. A 100-ns production run was then performed and trajectories and energy profiles were stored for subsequent analysis. All configs of these are avaliable at the Proscope repo.

Structure visualization

ChimeraX was used to visualize the predicted structures. VMD were used to generate the movie of molecular dynamics simulation trajectory.

Biological experiments

Cell Culture

HeLa cells were purchased from ATCC (CCL-2). HeLa cells were cultured in DMEM (Gibco, 11965) supplemented with 10% defined FBS (HyClone, SH30070), at 37°C/ 5% CO2.

TFAP2A co-immunoprecipitation

HeLa cell protein lysates were generated with 0.5% NP-40 lysis buffer (50mM Tris-HCl, 150mM NaCl, 0.5% NP-40) with phosphatase and protease inhibitor cocktail (Sigma-Aldrich, PPC1010). Samples were incubated with 5 µg agarose-conjugated TFAP2A primary antibody (Santa Cruz Biotechnology, sc-12726 AC) overnight at 4°C. Beads were washed, then boiled in Laemmli loading buffer (BioRad, 1610737). Proteins were separated on 10% Tris-Glycine gels (ThermoFisher, XP00100), transferred to PVDF (Immobilon-P, IPVH00010) and probed with primary antibodies against TFAP2A (ABclonal, A2294), ZFX (ThermoFisher, PA5-34376) and β-ACTIN (Santa Cruz Biotechnology, sc-47778) followed by chemiluminescence detection.

Data avaliability

Bulk RNA-sequencing of B-ALL Patients published in our previous study[19] is acquired at SRA (PRJNA534488). Human transcription factor protein interaction networks are downloaded from supplementary data of Helka, et al[17]. Training data and trained model will be open sourced upon publication.

References

1. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19, 10.1016/j.cell.2021.10.024 (2021).
2. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, 10.1126/science.aba7612 (2020). Publisher: American Association for the Advancement of Science Section: Research Article.
3. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137–R137, 10.1186/gb-2008-9-9-r137 (2008).
4. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736, 10.1038/s41586-020-2528-x (2020). Number: 7818 Publisher: Nature Publishing Group.
5. Korhonen, J. H., Palin, K., Taipale, J. & Ukkonen, E. Fast motif matching revisited: high-order PWMs, SNPs and indels. *Bioinformatics* **33**, 514–521, 10.1093/bioinformatics/btw683 (2016). _eprint: https://academic.oup.com/bioinformatics/article-pdf/33/4/514/49037769/bioinformatics_33_4_514.pdf.
6. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, 10.1126/science.aba7721 (2020). Publisher: American Association for the Advancement of Science Section: Research Article.
7. Consortium*, T. T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 10.1126/science.abl4896 (2022). Publisher: American Association for the Advancement of Science.
8. Hendrycks, D. & Gimpel, K. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415* (2016).
9. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
10. Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of Advances in Neural Information Processing Systems* (2019).
11. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019).
12. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks, 10.48550/arXiv.1703.01365 (2017). ArXiv:1703.01365 [cs].
13. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences, 10.48550/arXiv.1704.02685 (2019). ArXiv:1704.02685 [cs].
14. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 10.48550/arXiv.1802.03426 (2020). ArXiv:1802.03426 [cs, stat].
15. Shimizu, S., Hoyer, P. O., Hyv&#228, A., rinen & Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.* **7**, 2003–2030 (2006).
16. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613, 10.1093/nar/gky1131 (2019).
17. Göös, H. *et al.* Human transcription factor protein interaction networks. *Nat. Commun.* **13**, 766, 10.1038/s41467-022-28341-5 (2022). Number: 1 Publisher: Nature Publishing Group.
18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265, 10.1038/s41467-022-28865-w (2022). Number: 1 Publisher: Nature Publishing Group.
19. Oshima, K. *et al.* Mutational and functional genetics mapping of chemotherapy resistance mechanisms in relapsed acute lymphoblastic leukemia. *Nat. Cancer* **1**, 1113–1127, 10.1038/s43018-020-00124-1 (2020).