Deep genomic models of allele-specific measurements

Xinming Tu[1,*], Alexander Sasse[1,3,4,*], Kaitavjeet Chowdhary[2,*] , Anna E. Spiro[1], Liang Yang [2], , Maria Chikina[5], Christophe O. Benoist[2], Sara Mostafavi[1,6,$]

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington, WA, USA, 98195
[2] Department of Immunology, Harvard Medical School, Boston, MA, USA, 02115
[3] Center for Molecular Biology Heidelberg (ZMBH), Heidelberg University, Germany, 69120
[4] Center for Synthetic Genomics, Heidelberg, Germany, 69120
[5] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA, 16354
[6] Canadian Institute for Advanced Research, Toronto, ON, Canada, MG5 1ZB

*These authors contributed equally

[$]To whom the correspondence should be addressed to: saramos@cs.washington.edu

ABSTRACT

Allele-specific quantification of sequencing data, such as gene expression, allows for a causal investigation of how DNA sequence variations influence *cis* gene regulation. Current methods for analyzing allele-specific measurements for causal analysis rely on statistical associations between genetic variation across individuals and allelic imbalance. Instead, we propose DeepAllele, a novel deep learning sequence-to-function model using paired allele-specific input, designed to learn sequence features that predict subtle changes in gene regulation between alleles. Our approach is especially suited for datasets with few individuals with unambiguous phasing, such as F1 hybrids and other controlled genetic crosses. We apply our framework to three types of allele-specific measurements in immune cells from F1 hybrid mice, illustrating that as the complexity of the underlying biological mechanisms increases from TF binding to gene expression, the relative effectiveness of model's architecture becomes more pronounced. Furthermore, we show that the model's learned *cis*-regulatory grammar aligns with known biological mechanisms across a significantly larger number of genomic regions compared to baseline models. In summary, our work presents a computational framework to leverage genetic variation to uncover functionally-relevant regulatory motifs, enhancing causal discovery in genomics.

Main

Allele-specific quantification (ASQ) enables comparing the functional outcomes of two alleles within the same cellular environment[1–3] (**Fig. 1a**). This approach helps identify loci where genetic variation influences gene regulation and uncovers underlying causal mechanisms. However, achieving this requires pipelining multiple analyses, including associating genetic variation with allelic imbalance across individuals[4], fine-mapping variants to account for linkage disequilibrium[5], and performing motif binding analyses to identify disruptions in transcription factor (TF) binding that drive allelic differences in gene expression[6,7].

In this study, we introduce a novel framework that integrates ASQ with sequence-to-function (S2F) modeling[8,9] to directly derive causal insights into gene regulation. Our model, DeepAllele, extends sequence-based convolutional neural networks (CNNs) by incorporating allele-specific data through paired allele sequences as inputs *and* outputs (**Fig. 1b**). DeepAllele builds on standard sequence-based CNNs, which take genomic DNA subsequences as input and predict functional genomic measurements as output. These models learn sequence features (*i.e.,* sequence motifs) and their combinatorial rules (*cis*-regulatory grammar), which predict experimental measurements such as chromatin accessibility or gene expression in a given cellular context[8,10–12]. However, DeepAllele advances this approach by explicitly modeling allele-specific effects: it takes

paired DNA sequences from two alleles at the same genomic locus as input and predicts allele-resolved signal intensities and their log ratios (**Fig. 1b; Fig. S1**). As we illustrate below, this contrastive-allele design enables DeepAllele to directly infer the impact of genetic variation on functional outputs, enhancing the granularity of *cis*-regulatory grammar learned by S2F models. As a result, DeepAllele provides a more controlled and precise framework for analyzing allele-specific effects to reveal mechanisms of gene regulation.

To develop and evaluate DeepAllele, we utilized datasets from F1 hybrid mice (**Fig. 1a, Table S1**), generated by crossing two genetically distinct inbred strains. This experimental setup is ideal for ASQ analysis due to the high level of heterozygosity (~20 million variants in B6×Cast intercrosses and ~ 40 million variants in B6×Spret intercrosses) and the availability of inbred reference sequences, which facilitates accurate phasing. First, we compared DeepAllele with two baseline model architectures on several F1 datasets, including ChIP-seq for the transcription factor PU.1 in macrophages[13], chromatin accessibility (ATAC-seq) in T regulatory cells (Treg)[14], and RNA-seq (**Table S1**). For each type of data, we included distinct control experiments to ensure that our results are generalizable; for PU.1 ChIP-seq, we assembled data from two distinct F1 mouse strains (B6/PWK and B6/SPRET) providing variation in genetic background; for ATAC-seq, we utilized two independent datasets, generated as bulk and single-cell ATAC-seq data; for RNA-seq, we analyzed two different cell types, macrophages and B cells. Each of these studies essentially provides phased data for two alleles (each of the two inbred strains in a given F1 cross) at each locus (n=locix2 training data points) from the same cellular environment. In our setting, the standard "reference-only" model is trained on genomic DNA from one of the haplotypes (e.g., b6), and "two-genome" model still employs a single input and single output; its key difference is trained on genomic DNA from both alleles, so the number of loci is effectively doubled (**Fig. S1**).

First, as a control experiment, we assessed performance in predicting total counts (normalized by sequencing depth and log transformed) from held-out genomic locations using a leave-out-chromosome scheme (see Methods). We observe that all models perform well on this task (**Fig. 1c, S2a**). Next, we assessed the ability of models to predict allelic differences for the held-out regions, *i.e.,* the difference in allelic expression per region (computed as log allelic ratio). Importantly, the performance on total signal ("counts") and allelic difference measure different aspects of model's generalization; good performance on the latter additionally implies that the model has learned to distinguish functional from non-functional genetic variation and hence must learn a causal model of gene regulation. As shown in **Fig. 1d**, in predicting allelic differences the importance of model architecture is pronounced. Further, we observed that as the complexity of biological process captured by data type increases, from TF binding (PU.1

ChIP-seq), over chromatin accessibility (ATAC-seq) to gene expression (RNA-seq), the gap in model performance widens. On the other hand, different F1-hybrids (PU.1 ChIP-seq) or cell types (RNA-seq) did not influence the performance of the model. Sequence input length did not play a significant role (**Fig. S2b**) – for ChIP- and ATAC-seq most signal is proximal, and for RNA-seq the model likely focuses mainly on proximal regions. Nevertheless, the additional non-linear head for the ratio-predictions significantly improved predictions (**Fig. S2c**), compared to a model that only used the contrastive loss, especially for RNA-seq, likely due to the more complex nature of variants influencing RNA-seq.

We then examined DeepAllele's learned sequence features (motifs). This analysis focused on PU.1 ChIP-seq and ATAC-seq models, given their simpler interpretation. For both models, sequence attributions were robust between model initializations (**Fig. S3a**). For PU.1 ChIP-seq, consistent with the biology of the experiment, our model identifies the PU.1/SPI1 motif (**Fig. S3b**), predicts a significant fraction of sequences (~20%) contain a functional JUN motif, and as well, identifies a few other learned motifs (often partial SPI1 and JUN motifs) with lower functional abundance. The set of learned sequence motifs in ATAC-seq is more diverse, as expected since several TFs combine to open the chromatin at any given region (**Fig. S3c**), including the ETS1-like sequence motif known as a major driver of chromatin accessibility in Treg cells[14,15]. Other learned motifs include KLF2, CTCF, TCF7, and RUNX1.

The difficulty of predicting allelic difference stems from the fact that the impact of natural genetic variants on signal is more subtle than differences between expression of loci across the genome. Thus, an effective model will need to learn more subtle rules of *cis* regulatory grammar to accurately predict allelic variation. To probe the learned sequence grammar of DeepAllele in a more sensitive way, we contrasted the sequence attribution profiles from the two alleles in a given locus. As an example, **Figure 2a** shows sequence attribution profiles[8] for PU.1 binding on the two alleles in the *Kctd1* locus (Chr18:15,042,835, a test sequence) when trained on the ChIP-seq data[13]. At this position, one of the alleles has a significantly higher TF binding signal for PU.1/SPI1. DeepAllele also identifies a predictive JUN motif with a smaller effect size, which is in line with the reported co-activator functionality[16]. DeepAllele predicts the PU.1/SPI1 binding motif to be disrupted due to a C|T variant in the center of the motif. Indeed, amongst the twelve variants in this locus, in-silico mutagenesis (ISM) confirmed that the C|T variant located within the SPI1 motif has the strongest predicted effect. From here on, we will refer to the variant with the strongest predicted effect as the "main variant".

As another example, **Figure 2b** shows attribution maps near the *Tcof1* gene for the ATAC-seq trained model. Here, there appear to be two TF binding motifs in the Cast allele which are both disrupted in B6. DeepAllele predicts the C|T variant located in the

ELK1 motif to be the cause of the differential allelic activity. The secondary G|A variant is located in the flank of the GABPA motif and contributes less strongly to the observed allelic imbalance. Although the effect of the single variant in the GABPA motif appears to be small, the GABPA motif's activity is completely disrupted in the B6 allele, suggesting that the GABPA motif is also impacted by the variant in the ELK1 motif.

Next, we systematically probed how the observed genetic variation between the two inbred strains impacts the learned sequence grammar of gene regulation. We focused on 230 of 1,016 PU.1 ChIP-seq peaks and 346 of 1,081 Treg ATAC-seq peaks in the test sets with substantial allelic variation (|log FC|>=1) for which the model made accurate predictions (Methods, **Fig. S4a-b**). We observed that the ISM effect from the main variant could explain 64% (R=0.8, **Fig. S4c**) and 74% (R=0.86, **Fig. S4d**) of the DeepAllele's ratio predictions for models trained on PU.1 ChIP-seq and in ATAC-seq, respectively, suggesting that in these datasets most of allelic imbalance can be attributed to a single variant. The additive effects of multiple variants in each locus further explained another 6% in PU.1 ChIP-seq but 18% in ATAC-seq (total PU.1 ChIP-seq: 72%, R=0.85; total ATAC-seq: 92%, R=0.96) of the variance in DeepAllele's predictions (**Fig. S4e-f**).

Biologically, we expect that a coherent model distinguishes causal from non-causal variants based on their impact on the learned *cis* sequence grammar, as the predicted main variants that did not fall into a learned sequence motif could represent false positive predictions of the model or noise from ISM. To evaluate DeepAllele's biological coherence, we examined the location of the main variant with respect to the learned sequence motifs. Encouragingly, we observed that most of the main variants fall within an identified motif (90% for PU.1 ChIP-seq, and 79% for ATAC-seq) (**Fig. 2c-d**) in a biologically sensible manner. For example, for the PU.1 ChIP-seq data, 79% of main variants are located in PU.1/SPI1-like motifs (**Fig. 2c, S5a**). Interestingly, 4% of PU.1 PU.1 ChIP-seq binding sites possess their main variant in a JUN-like motif, suggesting that these motifs are not only picked up by the model due to spurious correlation but functionally effect PU.1 binding, further confirming the known interaction between these two regulators[17]. For the Treg ATAC-seq data, 42% of the main variants are located within ETS1-like motifs, known to be a major contributor to T cell and Treg-specific chromatin accessibility[14,15,18]. The model also identified TFs involved in control of state-specific changes in Treg chromatin (e.g. activation), such as TCF7 and RUNX, important for resting Tregs, EGR and IRF families, important for activated Tregs, as well as KLF and SP families, operative primarily at Treg promoter regions[14,15].

Finally, we reasoned that the improved prediction of DeepAllele model over the standard reference-only model architecture should also imply its more subtle and granular understanding of *cis* regulatory grammar of gene regulation. To test this

hypothesis, we extracted motifs from the sequence attributions of both DeepAllele and "reference-only" models and then jointly clustered them to identify differences between learned sequence context for the same main variants (Methods). As an example, **Figure 2e** shows that DeepAllele predicts the main variant of the sequence to be located in motif cluster 281, while there is no significant motif detectable in the baseline model, suggesting that the baseline model simply does not learn this motif or the correct activity in the sequence context. Next, we systematically determined the clusters in which the main variants were located for all well predicted sequences (**Fig. 2f**). We observe that the baseline model does not recognize a significant motif for almost half of the sequences for which DeepAllele determines an active motif (**Fig. 2f, S6**). Upon closer inspection (**Fig. S6b**), we found that in cases where both models detect the relevant motifs, a substantial fraction exhibit differences in predicted effect sizes. This discrepancy may explains DeepAllele's enhanced sensitivity in predicting variant effects.

In summary, we developed DeepAllele to analyze the impact of natural genetic variation using allele-specific measurements. Our model effectively identifies causal transcription factors (TFs) and functional variants that contribute to allelic imbalance in TF binding and chromatin accessibility. We hypothesize that extending this contrastive model architecture could be a promising direction for studying natural genetic variation in personal genomes. While previous research suggests that training on personal genomes does not enable model generalization to unseen loci[19,20], DeepAllele appears to overcome this limitation. We propose that allele-specific quantification of gene expression helps control for non-genetic signals more effectively. Additionally, the constrained sequence representation—limited to two alleles—may mitigate overfitting issues.

Method

Allele-specific quantification and data processing

We adapted the diploid pseudogenome alignment strategy implemented in the lapels/suspenders pipeline[21]. We obtained B6 and Cast pseudogenomes, MOD files, and variant vcf files from the UNC collaborative cross project and Mouse Genome Project. Reads were aligned to both B6 and Cast pseudogenomes, shifted to a common set of B6-based reference coordinates, and assigned to an allele of origin based on which allele has stronger mapping. Non-specific reads with equally strong mapping to both alleles were randomly split in half into B6 and Cast groups to obtain the final reads used for allele-specific analyses. For ATAC-seq datasets, reads were shifted to account for Tn5 insertion sites using deeptools. For Treg ATAC-seq datasets, we mapped reads to a previously defined set of Treg-specific OCRs.

For published TF-binding and chromatin accessibility datasets, FASTQ files were downloaded from the NIH Sequence Read Archive. Reads were trimmed using Trimgalore v0.6.6 (mshttps://github.com/FelixKrueger/TrimGalore), aligned using Bowtie2, filtered to retain high quality, singly-mapped reads with Samtools, and duplicates removed using Picard (http://broadinstitute.github.io/picard/). For F1 datasets, we used the allele-specific mapping strategy described above.

DeepAllele's model architecture

DeepAllele is a CNN framework to identify functional cis-regulatory motifs in paired DNA sequences from two alleles. It uses paired sequences as input for training and analysis, and a multi-head architecture to predict total (log normalized) counts from each allele as well as the log ratio between these sequences (three predictions). DeepAllele consists of three main components: (1) the shared convolution layers, (2) fully connected (FC) layers for predicting allelic ratio (the "ratio head"), (3) FC layers for predicting total counts (the "count head"). Each of these components uses standard convolutional layers, with and without residual connections, standard average pooling, fully connected layers and ReLU activations. The length of the input sequence to the model in this study were 330 bp for ATAC, 551 bp for ChIP, and 5100 bp for RNA-seq. Accordingly, we adjusted the pooling size to 4 for ATAC-seq, 2 for PU.1 ChIP-seq and 4 for RNA-seq. For ATAC-seq, we used 4 repeated residual convolutional blocks, 6 for PU.1 ChIP-seq, and 4 for RNA-seq.

Before both allelic sequences are given separately to the shared convolutional layers, their reverse complement is automatically concatenated to forward strand. This way the

model can make use of the same motif on either one of the strands for its predictions. The two concatenated tensors then go through the shared convolutional layers individually, i.e. both sequences are processed by the same parameters from the same convolutional layers in parallel. The two outputs from the convolutional layer are then flattened into 1D tensors containing each sequences information. Each flattened tensors passes through the count head individually to predict the log-counts of their allele. For the ratio predictions, the flattened tensors are concatenated and then passed through the ratio head layers to predict the log-ratio between alleles from information about both sequences. DeepAllele was trained using Pytorch lighting with Adam optimizer, weight decay = 0.00001, a learning rate of 0.0001, using the learning rate scheduler CyclicLR, base_lr=(learning rate)/2, max_lr=(learning rate)x2. The best model during training was selected based on validation loss. To quantify model performance, we used leave-chromosome out cross-validation strategy, excluding chromosomes 16, 17, and 18 (which is approximately 10% of total data points) for validation and test set. In addition to the final model architecture for DeepAllele, we also investigated a "direct" architecture that predicts log-counts with a shared linear head, and the log-ratio directly from the difference between the two. This model variation still uses both alleles as input but only uses the contrastive loss to improve ratio-predictions without the additional non-linear ratio-head that uses the concatenated sequence representations.

The "standard" or "baseline" model uses the same layers as DeepAllele but only predicts log-counts during training from a single input sequence. We trained models on sequences and counts from a single genome, as well as sequences and counts from two genomes. Log-ratios are only computed afterwards from the predicted counts from sequences from both genomes, and cannot be used by the model during training.

Details on DeepAllele's model architecture

Component	**Sub-component**	**Description of layers**

Shared convolutional layers	First convolutional block	• Conv1D
		  ○ in channels = 4
		  ○ out channels = 256
		  ○ kernel length = 15
		• ReLU
		• BatchNorm

	Second convolutional block	• Conv1D
		  ○ in channels = 256
		  ○ out channels = 256
		  ○ kernel length = 5
		• ReLU

		• BatchNorm

	Repeated residual convolutional blocks with pooling	• Conv1D
		  ○ in channels =256
		  ○ out channels = 256
		  ○ kernel length = 5
		• ReLU
		• BatchNorm
		• Residual connection (addition of input to Conv1D to output of BatchNorm)
		• AveragePool

Count head	First FC layer	• FC layer
		  ○ input size = (sequence length x 256)/pooling size
		  ○ output size = 256
		• ReLU

	2x repeated FC layers	• FC layer
		  ○ input size = 256
		  ○ output size = 256
		• ReLU
		• Dropout
		  ○ dropout = 0.2

	Output FC layer	• FC layer
		  ○ input size = 256
		  ○ output size = 1

Ratio head	2x repeated FC layers	• FC layer
		  ○ input size = 512
		  ○ output size = 512
		• ReLU
		• Dropout
		  ○ dropout = 0.2

	Output FC layer	• FC layer
		  ○ input size = 512
		  ○ output size = 1

Local sequence interpretation using attributions from (T)ISM

For each sequence in the test set, we performed ISM and TISM <cite Sasse et al. 2024 iScience https://github.com/LXsasse/TISM/tree/main> to quantify the contribution of each base-pair in both sequences to the model's predictions and identify regulatory

motifs acting on them. TISM represents an approximation of ISM values from the model's gradient that is processed in the same way as ISM values to return sequence attributions from a single forward pass through the model. Since DeepAllele takes two sequences as input to predict the allelic ratio, when we performed ISM for both sequences independently, i.e we switched one base of only one of the two sequences with another base. We then computed ISM values for each sequence (in B6 and CAST) as the difference between DeepAllele's predicted values of the allelic ratio of the alternated and the original sequence pair. We generated attribution maps from ISM values by subtracting the mean of all four ISM values at each position. We define these "centered ISM" values as sequence attributions which determine the importance of all four bases along the sequence. In figures, we only show the attribution of the base at that position, the three other attributions are hypothetical since they are not present in the sequence.

Extracting, comparing, and clustering motifs in sequence attribution maps

To compare motifs in the attribution maps from both sequences and different models, we implemented our own tools that use standard metrics for comparing sequence motifs, and standard agglomerative clustering to summarize clusters. First, to extract motifs from sequence attributions, we standardized these values by dividing them by their standard deviation from zero. We used these Z-score standardized attributions of the bases in the input sequences to determine coherently formed motifs from significant base-pair attributions ($|Z(a)| > 1.96$, i.e. $p<0.05$, two sided). We identified sequence motifs as regions with more than four subsequent bases with an absolute standardized attribution over 1.96. We allowed these regions to contain gaps of one base-pair, i.e. one single base-pair with an absolute value under 1.96, followed by another significant base-pair. We performed motif extractions for both sequences independently, and saved their locations in the individual sequences, as well as their location in the alignment. While only the standardized values at the base in the sequence were used to detect motifs, we extracted the standardized attribution maps for all four bases for subsequent clustering with a corrected sign from their mean effect on reference bases, so that motifs with negative effects can be clustered with motifs that have a positive effect, if they look similar on sequence and shape.

To cluster these extracted attribution seqlets (i.e standardized attributions for all four bases, even those that are not present in the sequence to capture the model's learned sequence grammar, not sequence enrichment in the test set sequences), we computed the Pearson's correlation coefficient between all extracted seqlets from both input sequences. To do so, we shifted the extracted seqlets against each other, padded non-aligning overhanging parts in each sequence with 0.25 and computed the Pearson's correlation between the two regions in which at least one motif was had non-padded

entries. We performed this for each pair of attribution seqlets for the forward strand and the reverse complement. The highest correlation between two seqlets and the respective p-value was kept in a distance matrix for clustering. We used the resulting p-value matrix to perform hierarchical clustering with complete linkage, joining all extracted attribution seqlets that correlated with a p-value of less than 0.01 with all other members of their cluster. To visualize the clusters, we created a joint contribution weight matrix (CWM) for each cluster by aligning all seqlets of a cluster to the centroid seqlet, which was defined as the seqlet that is closest in the mean to all the other ones in the cluster. The aligned attribution seqlets of all members of clusters were summed up and divided by the maximum number of seqlets that contributed to a single position in the cluster's joint CWM.

Attribution maps and the joint CWMs contain negative values and do not sum to 1 at each position. To perform motif matching with TomTom, we created PFMs from the joint CWMs by taking the exponential of the CWM values before normalizing every position to 1. We then used TomTom to find known regulatory factors that match the sequence patterns from our analysis. Tomtom matched the PFMs of our clusters to the JASPAR 2020 non-redundant vertebrate database. We assigned every cluster to known regulatory factors if the q-value was below 0.05.

Identifying main variants predictive of allelic imbalance

To identify the main variant that DeepAllele predicts to be casual for allelic imbalance, we performed separate ISMs for all variants between B6 and CAST. Specifically, we inserted single variants of CAST into the B6 sequence, and vice versa B6 variants into the CAST sequence while leaving the other sequence of the input pair intact. We then computed the ISM as the difference in log-ratio between the sequence pair with the variant and the original sequence pair, and averaged over the two ISM values from both independent insertions. We defined the variant with the largest effect size in a sequence pair as the main variant.

Data and Code availability

All the data sets used here are available on the Gene Expression Omnibus (GEO). Details about the data sets can be found in Table S1. F1-hybrid scATAC-seq data for sorted T conventional and Foxp3 reporter+ Treg cells was downloaded from GSE216910. Bulk ATAC-seq of sorted T conventional and Foxp3 reporter+ Treg cells was downloaded from GSE154680. Both ChIP-seq data sets for bone marrow-derived macrophages without KLA stimulation were obtained from GSE109965. RNA-seq data for Mast cells after IL4 stimulation and B-cells after IL4 stimulation were deposited on GEO with this publication at GSE292399.

All code that was used for data processing and analysis is available at https://github.com/mostafavilabuw/DeepAllele-public/.

Author contributions

Conceptualization: X.T., S.M., C.B.; methodology: X.T., K.C., S.M., A.S., A.E.S. ; data processing: K.C., L.Y. ; data curation: K.C., L.Y., X.T. software: X.T., A.S., K.C. A.E.S.; analysis and interpretation of results: X.T., A.E.S., A.S, M.C, K.C. ; supervision: C.B., S.M. ; writing: A.E.S., S.M. ;

Acknowledgements

This research was in part made possible by NIH-R24 (5R24AI072073-18). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

FIGURE LEGENDS

Figure 1. DeepAllele modeling strategy.

a) Schematic of allele-specific quantification (ASQ) of sequencing data across the genome to detect contrasting functional measurements in the same trans-environment. Functional variants cause significant differences between the signal of two alleles **b)** (top) Schematic of the DeepAllele model architecture, using genomic sequence from a pair of alleles as inputs to predict their individual allelic log counts, and their log ratio with separate prediction heads. The convolutional layers are shared between the two input sequences, enabling coherent sequence attribution and motif discovery. (bottom) Schematic of model interpretation to extract learnt sequence grammar and main variant effect predictions from the trained model. **c)** Model performance on predicting log counts in regions from test chromosomes using Pearson correlation for standard single-input trained on reference genome only, single-input trained on both genomes, and DeepAllele model. The standard deviation is computed from five model initializations over predictions on both genomes **d)** Model performance on predicting log ratios in test chromosomes using Pearson correlation. Standard error was computed from five model initializations.

Figure 2. Examples Identifying the mechanism behind allelic imbalances in F1 data.

a) Attribution map from ISM for a sequence-pair to the log ratio for PU.1 ChIP-seq from DeepAllele. The bottom track shows the effect size of the individual variants between both sequences. Detectable motifs are framed and names of associated TFs from JASPAR are shown besides the frame (TomTom $q<0.05$) **b)** Attribution map from ISM for a sequence-pair to the log ratio from DeepAllele ATAC-seq. The three rows show

the same as in a) **c)** Pie-chart showing the percentage of main variants (i.e. variant with the largest predicted effect) in their associated TF-like motif groups in sequence attributions with trusted log-ratio predictions for PU.1 ChIP-seq data. The TF with the largest total enrichment across motif clusters was assigned to all seqlet clusters that had a significant match to this TF (i.e. q-value < 0.05). TF enrichment was computed as the product of the negative log p-value of a TF and the number of seqlets in a cluster. All clusters assigned to the TF are combined as the TF-like group. **d)** Pie-chart showing the percentage of main variants in their associated TF-like motif groups in sequences with trusted log-ratio predictions for ATAC-seq data. **e)** Sequence attributions and variant effect sizes for both alleles from DeepAllele and the standard single-input model. Extracted TF motif and the assigned cluster number is added to the sequence attribution**. f)** Pie-chart showing the motif clusters with the main variant for DeepAllele and the standard model.

Supplementary Figure Legends

Figure S1. DeepAllele and other model architectures.
a) Architecture of the shared convolutional layers, individual modules, and prediction heads that were used in all models. **b)** Model architecture of baseline model with single input sequence. **c)** Model architecture of paired-input direct-ratio prediction model with contrastive loss but without log-ratio head. **d)** DeepAllele model architecture: Convolutional layers are shared for processing both sequences in parallel. Shared count prediction heads use sequence embeddings of each allele to predict counts for their respective genome. The ratio-head uses concatenated sequence embeddings from the shared convolutional layers to predict allelic ratios.

Figure S2. Model performance metrics and ablations.
a) Spearman correlation between predicted and measured log counts and log ratios for standard single-input trained on reference genome only, single-input trained on both genomes, and DeepAllele model in Fig. 1c,d. **b)** Performance comparison (Pearson R for counts, and Pearson R for log ratios) between DeepAllele and Standard models (trained on reference genome and both genomes) from a) for different sequence input window sizes for RNA-seq from B cells and MF cells. Tested sequence inputs were 2.5kb, 5kb, 10kb, and 20kb **c)** Comparison of DeepAllele's log-ratio predictions to a paired-input direct log-ratio prediction model without log-ratio head (see Fig. S1)

Figure S3. Sequence attributions and their motif clusters.
a) Correlation between ISM sequence attributions from two model initializations for ChIP (left) and ATAC-seq (right). **b)** Average linkage dendrogram on correlation distance between aligned motifs for the largest contribution weight matrices from seqlet motif

clusters in attributions from PU.1 ChIP-seq. Barplot shows the percentage of test set sequences with a motif seqlet in that motif cluster. Grey bars show how many times the attribution was positive and red bars show many time it was negative. **c)** Same as in b) but for bulk ATAC-seq data.

Figure S4. Selection of "trusted peak" set and identification of main variants.
a) Measured versus predicted log ratios from DeepAllele for PU.1 ChIP-seq held-out test set. Data points with a measured |log-ratio| > 1 and a significant prediction (i.e. Z-score of prediction value > 1.65), are colored in red and selected for subsequent mechanistic analysis. Contour represents distribution of peaks. Red line shows diagonal. **b)** Same as in a for bulk ATAC-seq test set. **c)** Predicted main variant effect on log ratio from average ISM (over both alleles) versus predicted allelic ratio for PU.1 ChIP-seq of the entire model. Red dots indicate "trusted peaks" with significant allelic imbalance from a). **d)** Same as in c for ATAC-seq. **e)** Sum of all variants predicted average ISM effects versus predicted allelic ratio for PU.1 ChIP-seq from DeepAllele model. Red dots represent trusted peaks from a and grey dots others. **f)** Same as in e) for bulk ATAC-seq.

Figure S5. Motif clusters affected by main variants.
a) Dendrogram with average linkage on Pearson correlation between shown motifs. Shown are the combined CWMs of the motif clusters that are affected by the main variants and the percentage of trusted sequences with this motif cluster being affected in PU.1 ChIP-seq. Grey bars show the percentage of seqlets that have positive effect on the DeepAllele's count predictions, while red bars show the percentage of seqlets with negative effects **b)** Dendrogram with average linkage on Pearson correlation between shown motifs. Shown are the combined CWMs of the motif clusters that are affected by the main variants and the percentage of trusted sequences with this motif cluster being affected in bulk ATAC-seq. **c)** Pie-chart showing the affected motif clusters by the main variant from trusted sequences in PU.1 ChIP-seq, corresponding to a). **d)** Same as in c) but for bulk ATAC-seq shown in b)

Figure S6. Comparison of the mechanistic interpretation of the main variant location to the standard single input model.
a) Sankey plot comparing the motif cluster assignment between seqlets extracted around the main variant from DeepAllele and the standard single-input model. **b)** Example sequence for which the main variant is located in motif cluster 281 in sequence attributions from DeepAllele while the main variant is located in motif cluster 133 in sequence attributions from the single-input model.

Figure 1

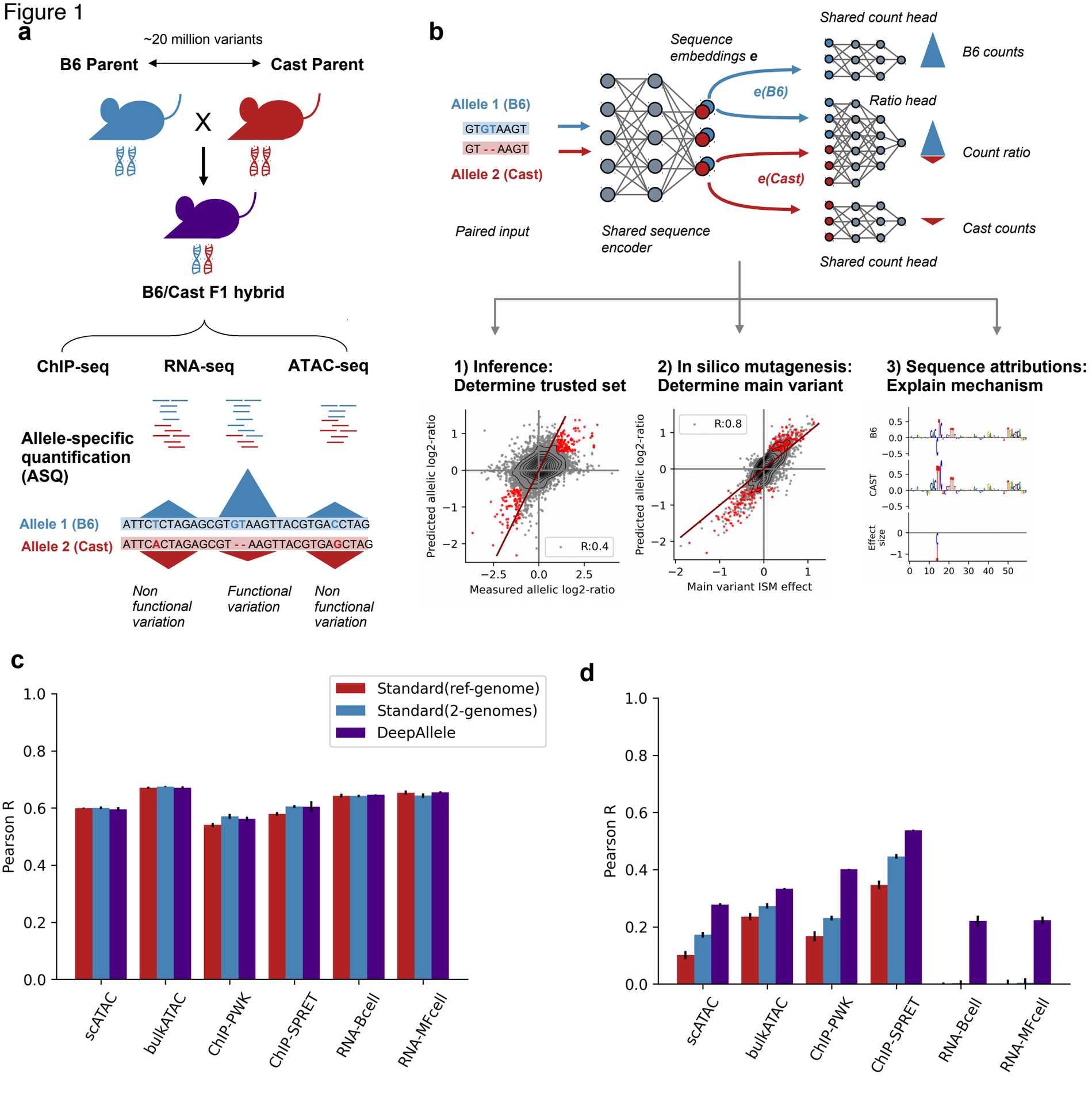

a
~20 million variants
B6 Parent
Cast Parent
X
B6/Cast F1 hybrid
ChIP-seq
RNA-seq
ATAC-seq
Allele-specific quantification (ASQ)
Allele 1 (B6)
Allele 2 (Cast)
ATTCTCTAGAGCGTGTAAGTTACGTGACCTAG
ATTCACTAGAGCGT - - AAGTTACGTGAGCTAG
Non functional variation
Functional variation
Non functional variation
b
Shared count head
Sequence embeddings e
B6 counts
Allele 1 (B6)
GTGTAAGT
GT - - AAGT
Allele 2 (Cast)
e(B6)
Ratio head
Count ratio
e(Cast)
Cast counts
Paired input
Shared sequence encoder
Shared count head
1) Inference: Determine trusted set
2) In silico mutagenesis: Determine main variant
3) Sequence attributions: Explain mechanism
Predicted allelic log2-ratio
Measured allelic log2-ratio
R:0.4
R:0.8
Main variant ISM effect
B6
CAST
Effect size
c
Standard(ref-genome)
Standard(2-genomes)
DeepAllele
Pearson R
scATAC
bulkATAC
ChIP-PWK
ChIP-SPRET
RNA-Bcell
RNA-MFcell
d
Pearson R
scATAC
bulkATAC
ChIP-PWK
ChIP-SPRET
RNA-Bcell
RNA-MFcell

Figure 2

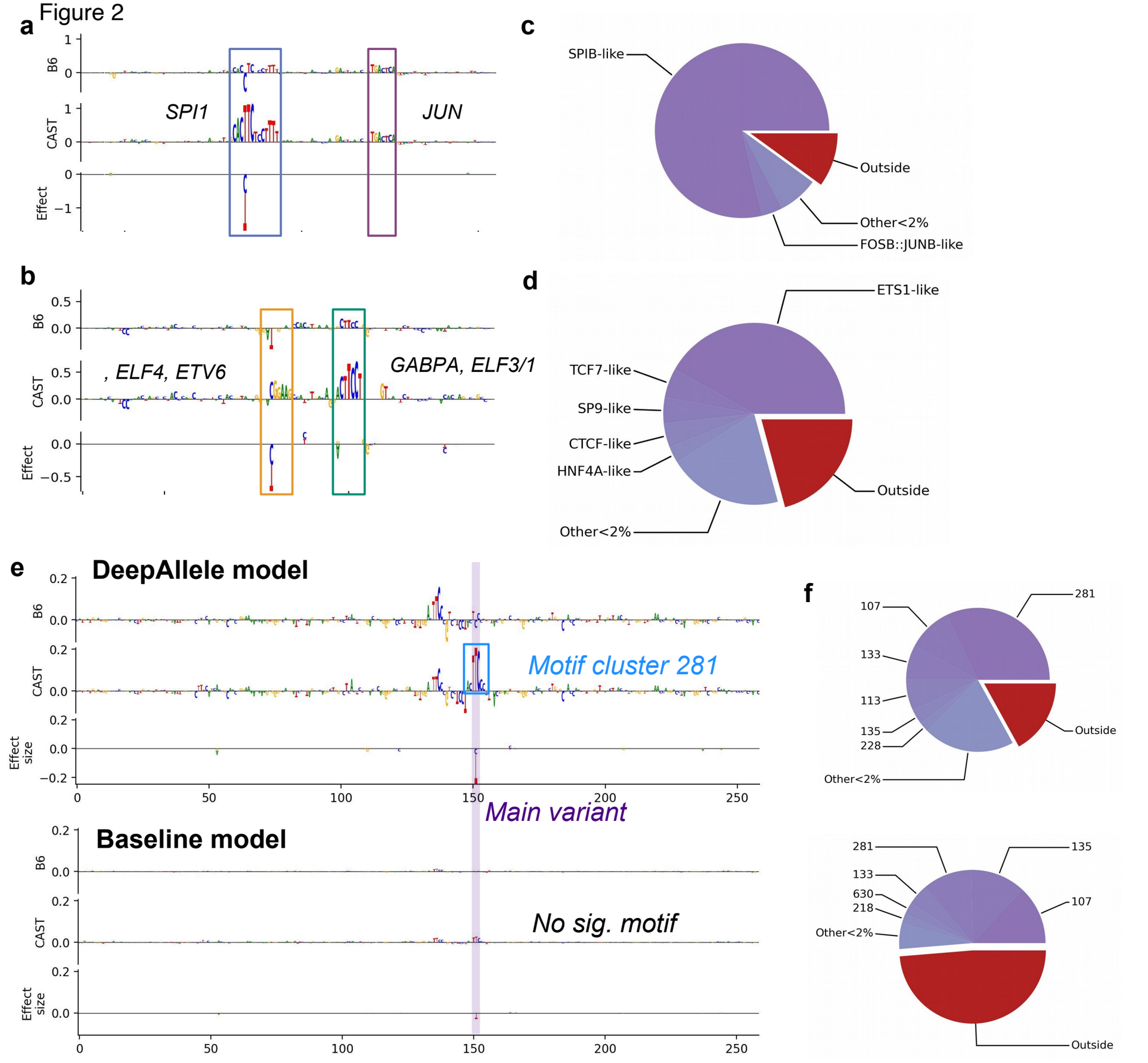

a
B6
CAST
Effect
SPI1
JUN
b
B6
CAST
Effect
, ELF4, ETV6
GABPA, ELF3/1
c
SPIB-like
Outside
Other<2%
FOSB::JUNB-like
d
ETS1-like
TCF7-like
SP9-like
CTCF-like
HNF4A-like
Outside
Other<2%
e
DeepAllele model
B6
CAST
Effect size
Motif cluster 281
Main variant
Baseline model
No sig. motif
f
281
107
133
113
135
228
Outside
Other<2%
281
135
133
630
218
107
Other<2%
Outside

Figure S1

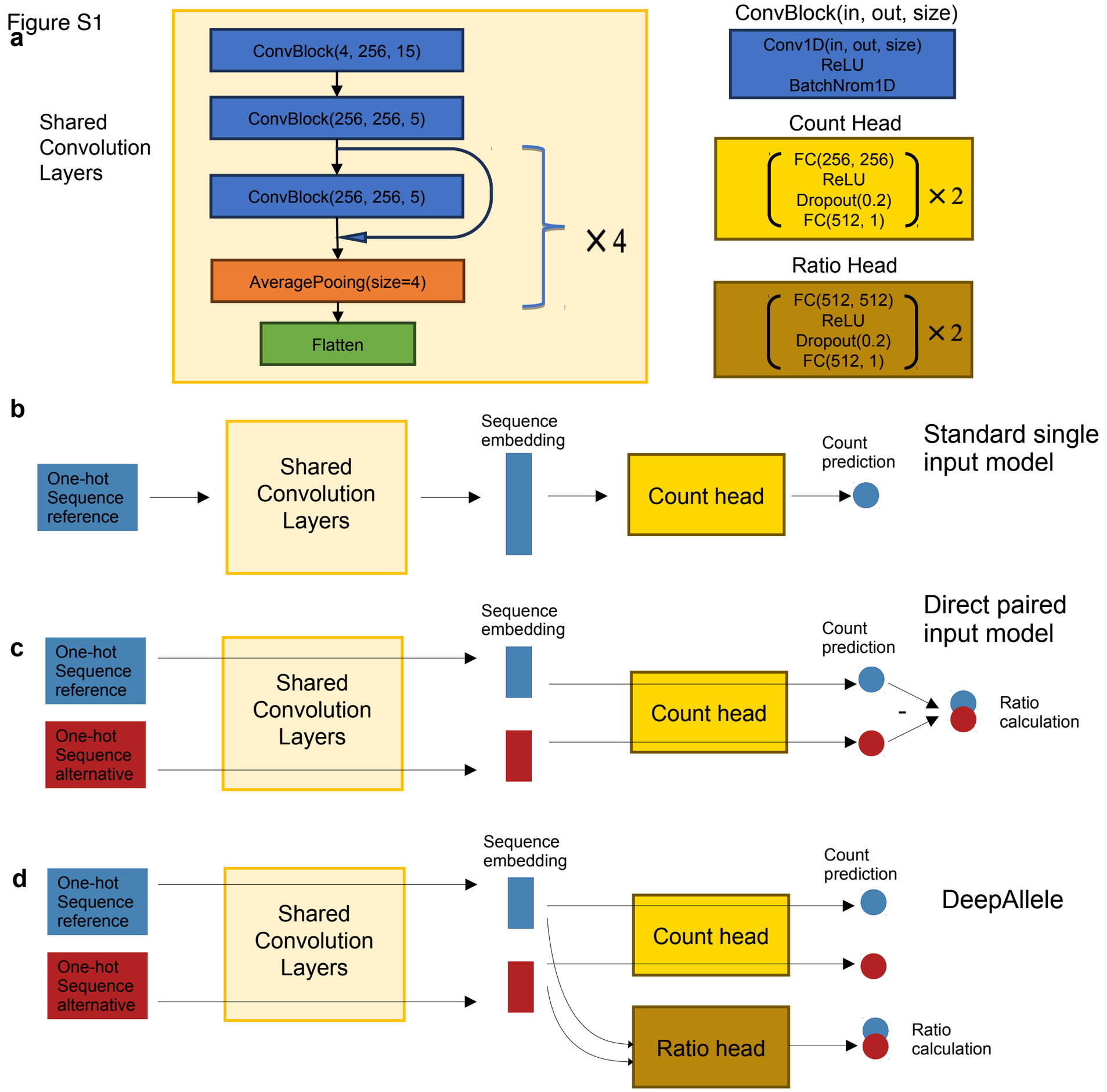

a
Shared Convolution Layers
ConvBlock(4, 256, 15)
ConvBlock(256, 256, 5)
ConvBlock(256, 256, 5)
AveragePooing(size=4)
Flatten
×4
ConvBlock(in, out, size)
Conv1D(in, out, size)
ReLU
BatchNrom1D
Count Head
FC(256, 256)
ReLU
Dropout(0.2)
FC(512, 1)
×2
Ratio Head
FC(512, 512)
ReLU
Dropout(0.2)
FC(512, 1)
×2
b
One-hot Sequence reference
Shared Convolution Layers
Sequence embedding
Count head
Count prediction
Standard single input model
c
One-hot Sequence reference
One-hot Sequence alternative
Shared Convolution Layers
Sequence embedding
Count head
Count prediction
-
Ratio calculation
Direct paired input model
d
One-hot Sequence reference
One-hot Sequence alternative
Shared Convolution Layers
Sequence embedding
Count head
Ratio head
Count prediction
Ratio calculation
DeepAllele

Figure S2

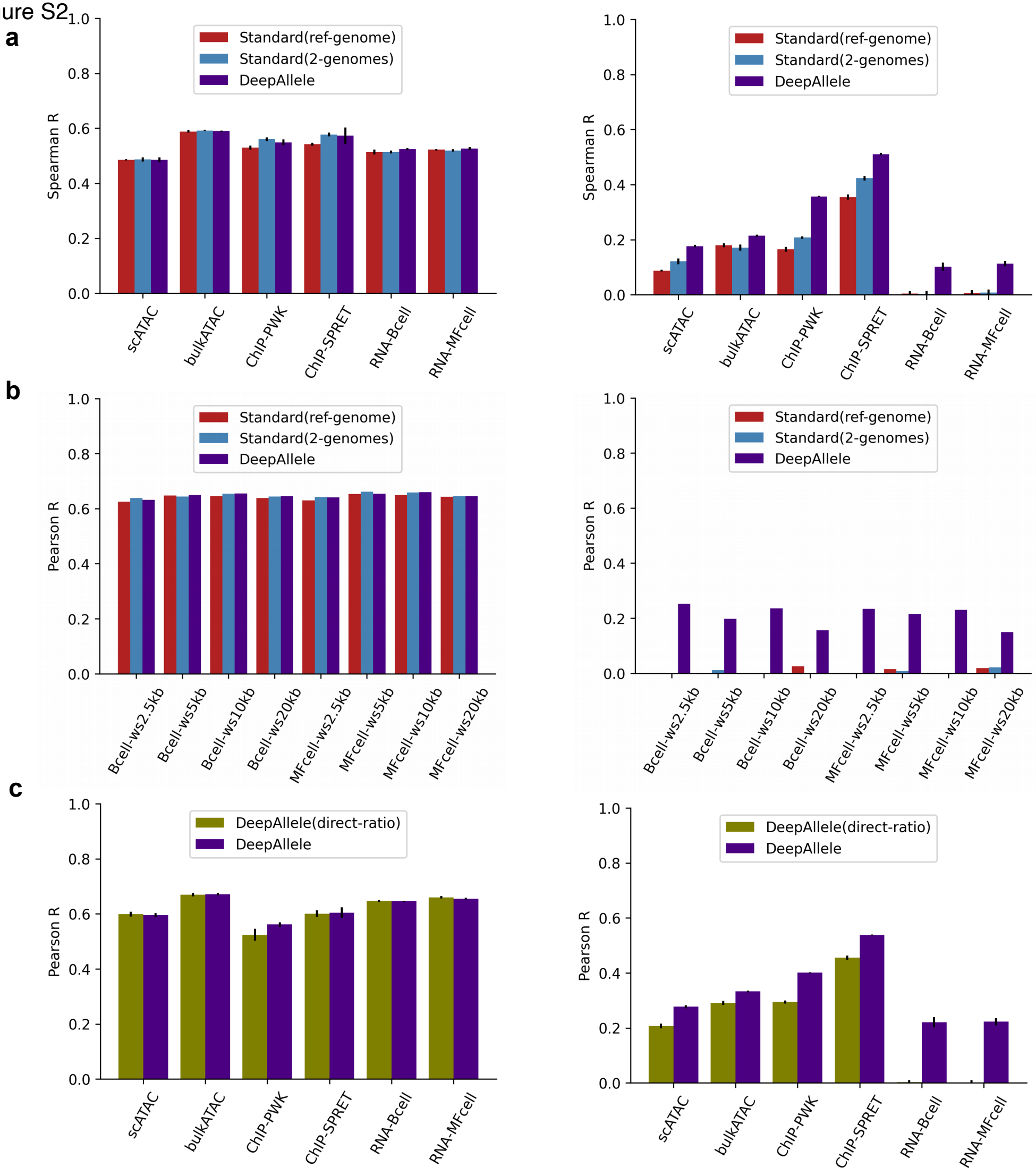

a
Standard(ref-genome)
Standard(2-genomes)
DeepAllele
Spearman R
scATAC
bulkATAC
ChIP-PWK
ChIP-SPRET
RNA-Bcell
RNA-MFcell
b
Pearson R
Bcell-ws2.5kb
Bcell-ws5kb
Bcell-ws10kb
Bcell-ws20kb
MFcell-ws2.5kb
MFcell-ws5kb
MFcell-ws10kb
MFcell-ws20kb
c
DeepAllele(direct-ratio)
DeepAllele

Figure S3

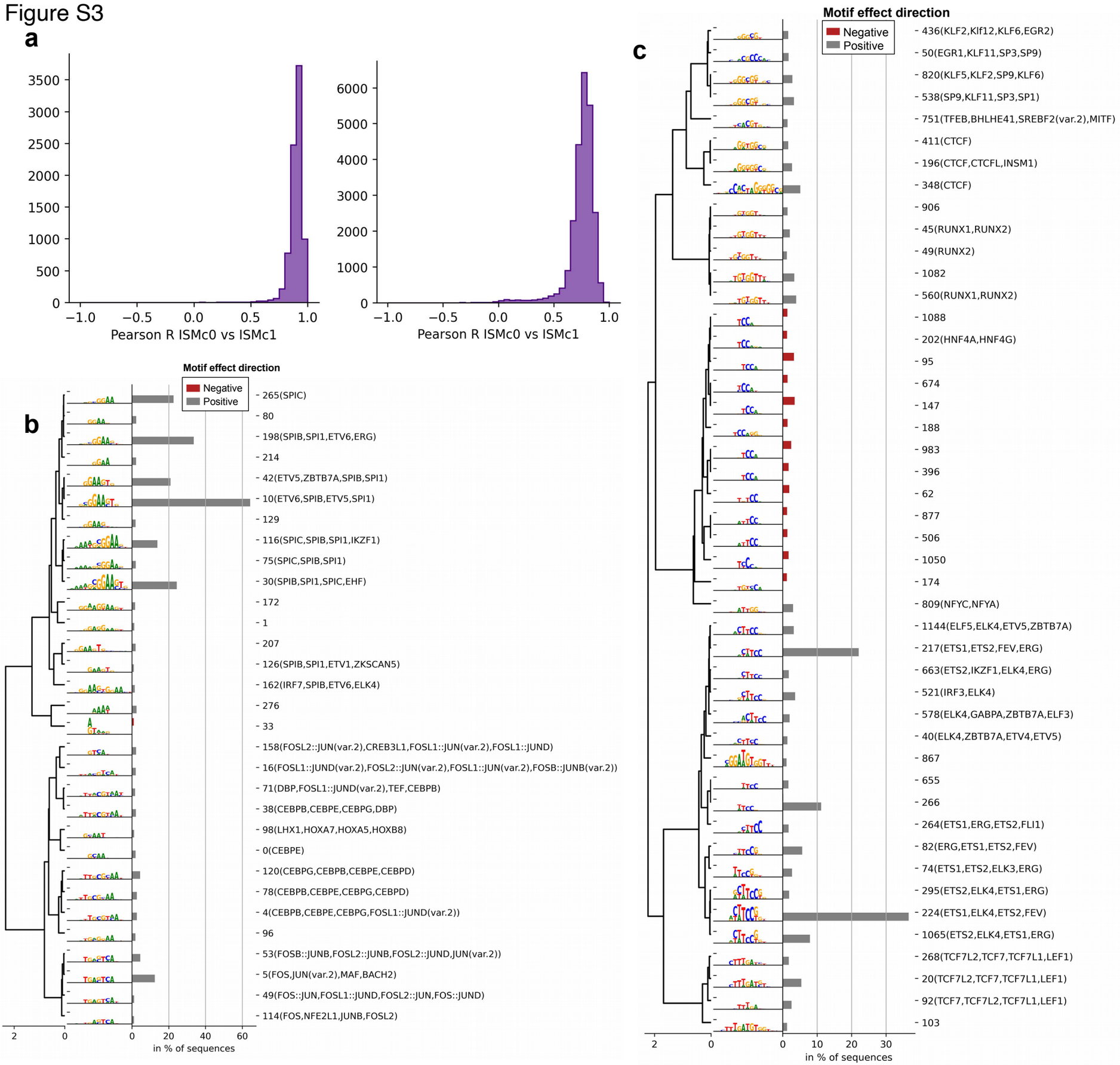

a
Pearson R ISMc0 vs ISMc1
Pearson R ISMc0 vs ISMc1
b
Motif effect direction
Negative
Positive
265(SPIC)
80
198(SPIB,SPI1,ETV6,ERG)
214
42(ETV5,ZBTB7A,SPIB,SPI1)
10(ETV6,SPIB,ETV5,SPI1)
129
116(SPIC,SPIB,SPI1,IKZF1)
75(SPIC,SPIB,SPI1)
30(SPIB,SPI1,SPIC,EHF)
172
1
207
126(SPIB,SPI1,ETV1,ZKSCAN5)
162(IRF7,SPIB,ETV6,ELK4)
276
33
158(FOSL2::JUN(var.2),CREB3L1,FOSL1::JUN(var.2),FOSL1::JUND)
16(FOSL1::JUND(var.2),FOSL2::JUN(var.2),FOSL1::JUN(var.2),FOSB::JUNB(var.2))
71(DBP,FOSL1::JUND(var.2),TEF,CEBPB)
38(CEBPB,CEBPE,CEBPG,DBP)
98(LHX1,HOXA7,HOXA5,HOXB8)
0(CEBPE)
120(CEBPG,CEBPB,CEBPE,CEBPD)
78(CEBPB,CEBPE,CEBPG,CEBPD)
4(CEBPB,CEBPE,CEBPG,FOSL1::JUND(var.2))
96
53(FOSB::JUNB,FOSL2::JUNB,FOSL2::JUND,JUN(var.2))
5(FOS,JUN(var.2),MAF,BACH2)
49(FOS::JUN,FOSL1::JUND,FOSL2::JUN,FOS::JUND)
114(FOS,NFE2L1,JUNB,FOSL2)
in % of sequences
c
Motif effect direction
Negative
Positive
436(KLF2,Klf12,KLF6,EGR2)
50(EGR1,KLF11,SP3,SP9)
820(KLF5,KLF2,SP9,KLF6)
538(SP9,KLF11,SP3,SP1)
751(TFEB,BHLHE41,SREBF2(var.2),MITF)
411(CTCF)
196(CTCF,CTCFL,INSM1)
348(CTCF)
906
45(RUNX1,RUNX2)
49(RUNX2)
1082
560(RUNX1,RUNX2)
1088
202(HNF4A,HNF4G)
95
674
147
188
983
396
62
877
506
1050
174
809(NFYC,NFYA)
1144(ELF5,ELK4,ETV5,ZBTB7A)
217(ETS1,ETS2,FEV,ERG)
663(ETS2,IKZF1,ELK4,ERG)
521(IRF3,ELK4)
578(ELK4,GABPA,ZBTB7A,ELF3)
40(ELK4,ZBTB7A,ETV4,ETV5)
867
655
266
264(ETS1,ERG,ETS2,FLI1)
82(ERG,ETS1,ETS2,FEV)
74(ETS1,ETS2,ELK3,ERG)
295(ETS2,ELK4,ETS1,ERG)
224(ETS1,ELK4,ETS2,FEV)
1065(ETS2,ELK4,ETS1,ERG)
268(TCF7L2,TCF7,TCF7L1,LEF1)
20(TCF7L2,TCF7,TCF7L1,LEF1)
92(TCF7,TCF7L2,TCF7L1,LEF1)
103
in % of sequences

Figure S5

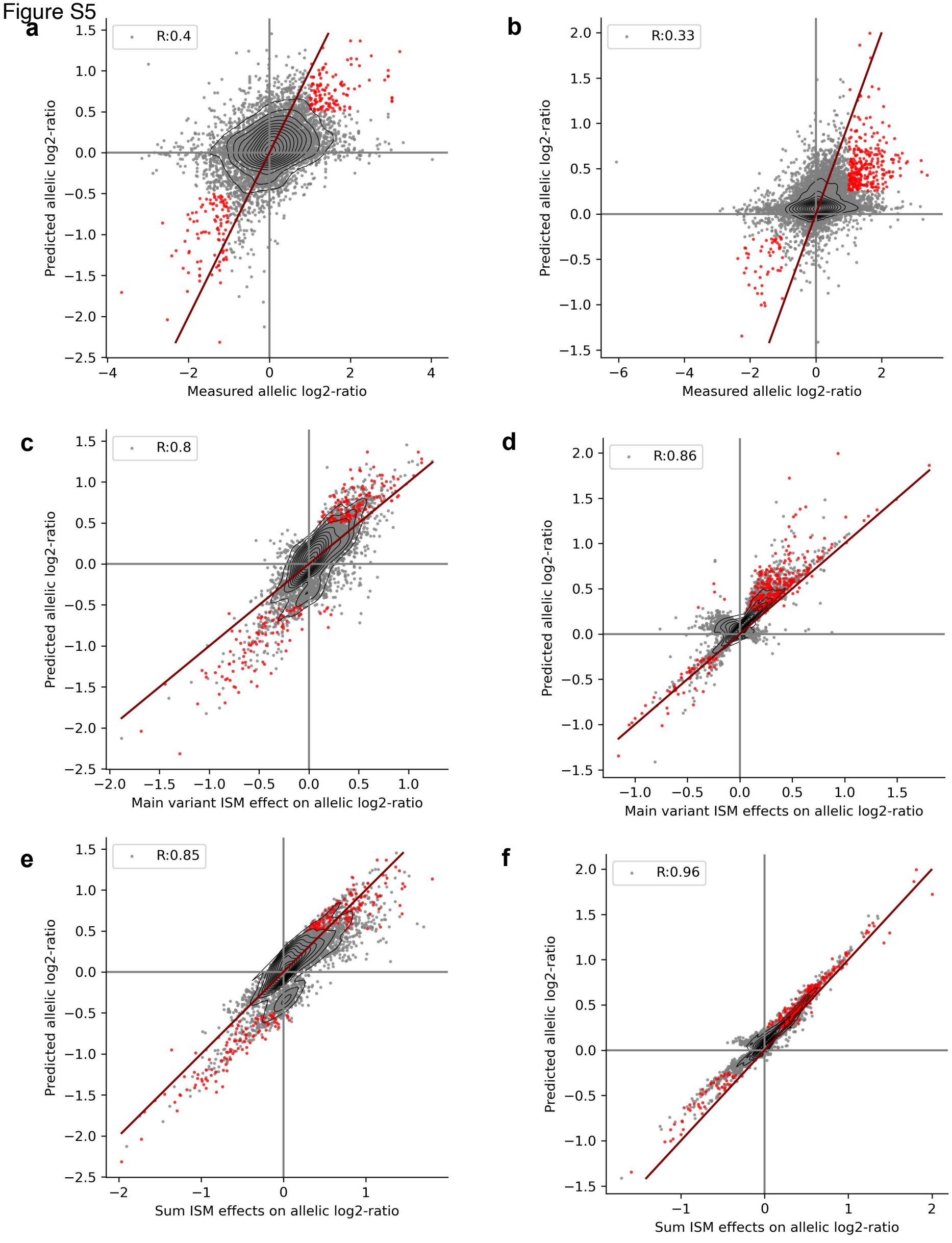
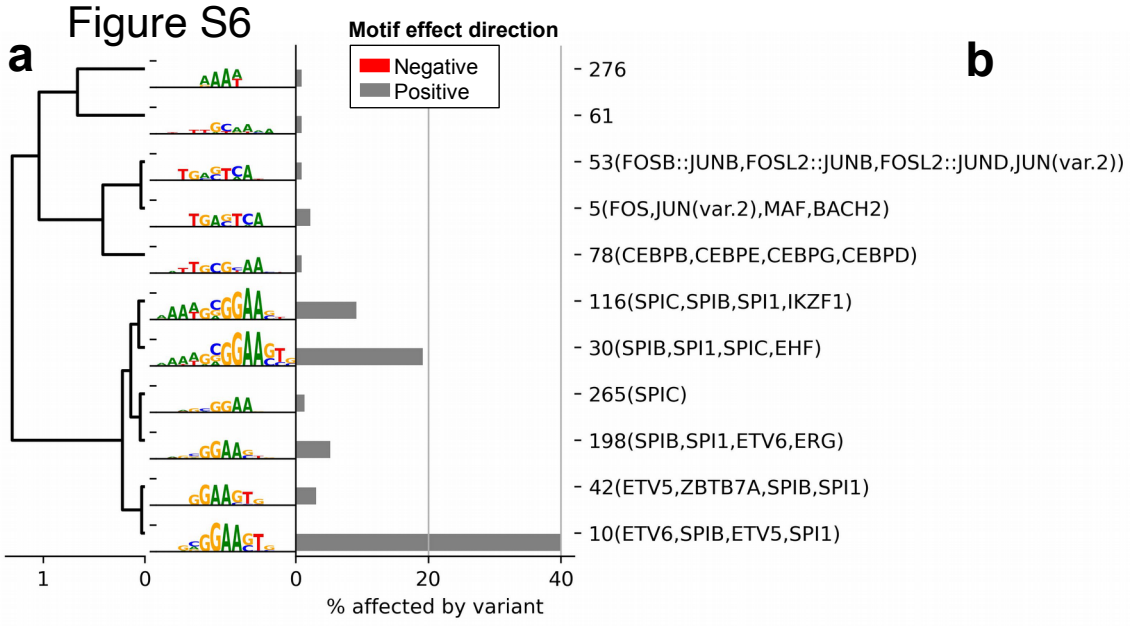
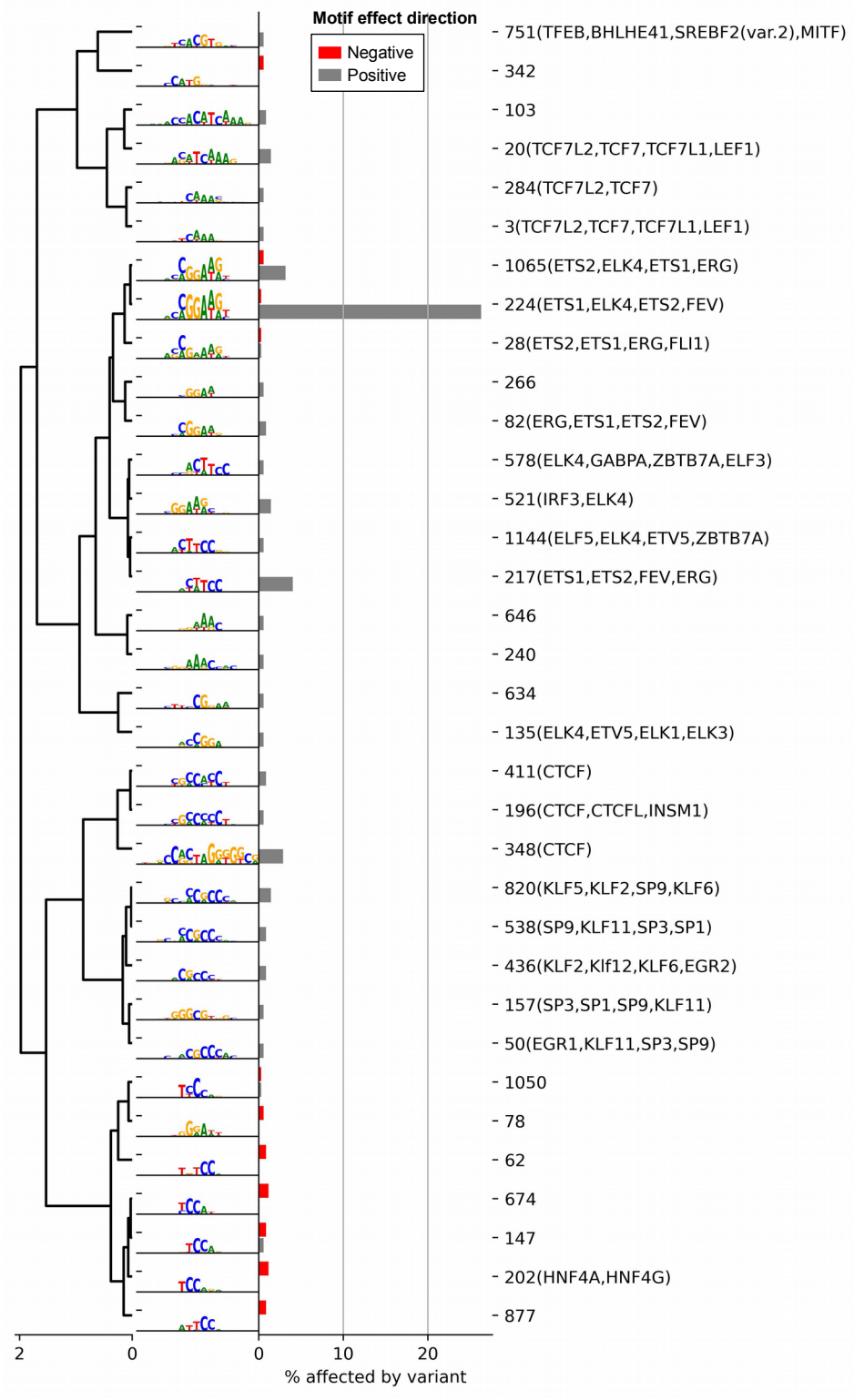
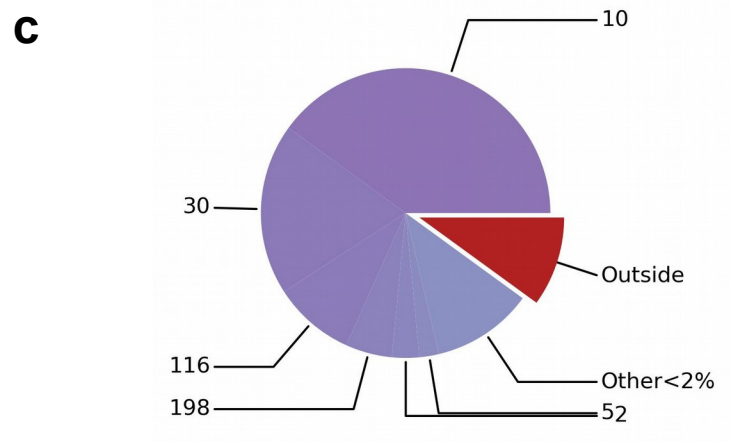

a
R:0.4
Predicted allelic log2-ratio
Measured allelic log2-ratio
b
R:0.33
Predicted allelic log2-ratio
Measured allelic log2-ratio
c
R:0.8
Predicted allelic log2-ratio
Main variant ISM effect on allelic log2-ratio
d
R:0.86
Predicted allelic log2-ratio
Main variant ISM effects on allelic log2-ratio
e
R:0.85
Predicted allelic log2-ratio
Sum ISM effects on allelic log2-ratio
f
R:0.96
Predicted allelic log2-ratio
Sum ISM effects on allelic log2-ratio

Figure S6


a
Motif effect direction
Negative
Positive
276
61
53(FOSB::JUNB,FOSL2::JUNB,FOSL2::JUND,JUN(var.2))
5(FOS,JUN(var.2),MAF,BACH2)
78(CEBPB,CEBPE,CEBPG,CEBPD)
116(SPIC,SPIB,SPI1,IKZF1)
30(SPIB,SPI1,SPIC,EHF)
265(SPIC)
198(SPIB,SPI1,ETV6,ERG)
42(ETV5,ZBTB7A,SPIB,SPI1)
10(ETV6,SPIB,ETV5,SPI1)
1
0
0
20
40
% affected by variant


b
Motif effect direction
Negative
Positive
751(TFEB,BHLHE41,SREBF2(var.2),MITF)
342
103
20(TCF7L2,TCF7,TCF7L1,LEF1)
284(TCF7L2,TCF7)
3(TCF7L2,TCF7,TCF7L1,LEF1)
1065(ETS2,ELK4,ETS1,ERG)
224(ETS1,ELK4,ETS2,FEV)
28(ETS2,ETS1,ERG,FLI1)
266
82(ERG,ETS1,ETS2,FEV)
578(ELK4,GABPA,ZBTB7A,ELF3)
521(IRF3,ELK4)
1144(ELF5,ELK4,ETV5,ZBTB7A)
217(ETS1,ETS2,FEV,ERG)
646
240
634
135(ELK4,ETV5,ELK1,ELK3)
411(CTCF)
196(CTCF,CTCFL,INSM1)
348(CTCF)
820(KLF5,KLF2,SP9,KLF6)
538(SP9,KLF11,SP3,SP1)
436(KLF2,Klf12,KLF6,EGR2)
157(SP3,SP1,SP9,KLF11)
50(EGR1,KLF11,SP3,SP9)
1050
78
62
674
147
202(HNF4A,HNF4G)
877
2
0
0
10
20
% affected by variant


c
10
30
Outside
116
Other<2%
198
52

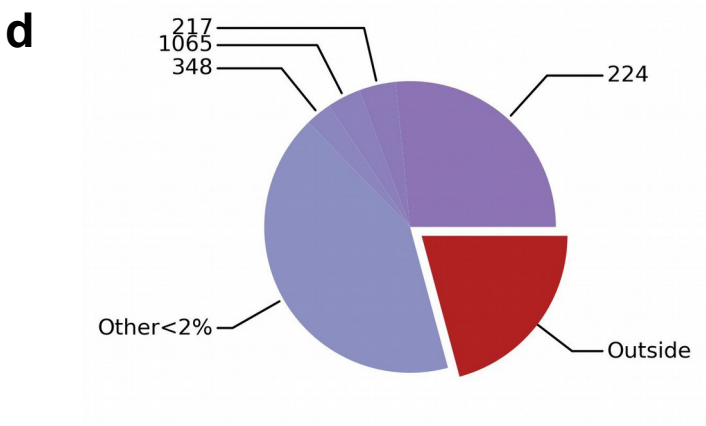
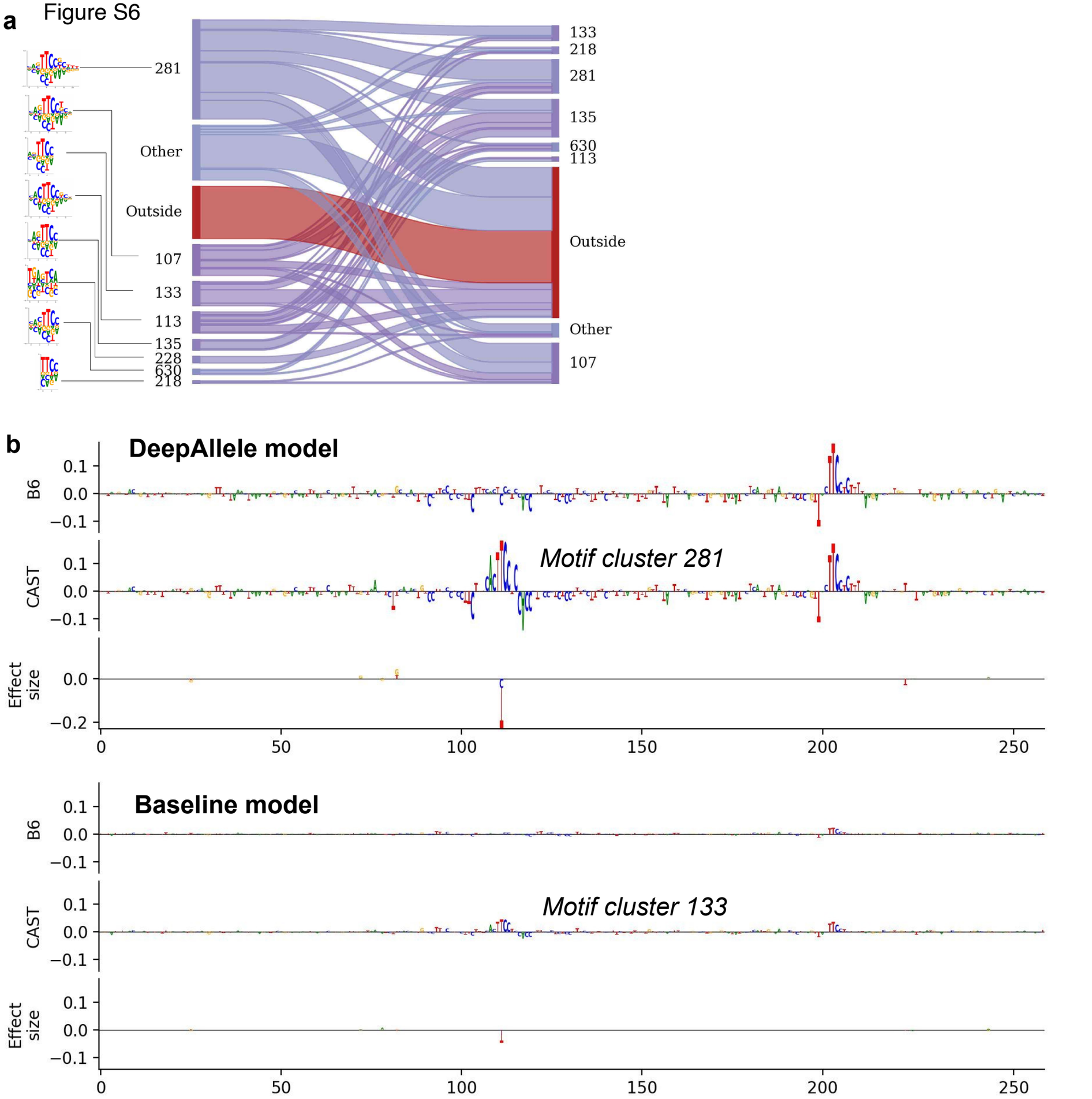

d
217
1065
348
224
Other<2%
Outside

Figure S6


a
281
Other
Outside
107
133
113
135
228
630
218
133
218
281
135
630
113
Outside
Other
107
b
DeepAllele model
B6
CAST
Effect size
Motif cluster 281
Baseline model
Motif cluster 133
0.1
0.0
−0.1
−0.2
0
50
100
150
200
250

References:

1. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).

2. Castel, S. E. *et al.* A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, (2020).

3. Andergassen, D. *et al.* Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. *Elife* **6**, e25125 (2017).

4. St Pierre, C. L. *et al.* Genetic, epigenetic, and environmental mechanisms govern allele-specific gene expression. *Genome Res.* **32**, 1042–1057 (2022).

5. Zou, J. *et al.* Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLoS Genet.* **15**, e1008481 (2019).

6. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).

7. Shen, Z., Hoeksema, M. A., Ouyang, Z., Benner, C. & Glass, C. K. MAGGIE: leveraging genetic variation to identify DNA sequence motifs mediating transcription factor binding and function. *Bioinformatics* **36**, i84–i92 (2020).

8. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00532-2.

9. Sasse, A., Chikina, M. & Mostafavi, S. Unlocking gene regulation with sequence-to-function models. *Nat. Methods* **21**, 1374–1377 (2024).

10. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).

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

12. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* vol. 26 990–999 Preprint at https://doi.org/10.1101/gr.200535.115 (2016).

13. Link, V. M. *et al.* Analysis of genetically diverse macrophages reveals local and domain-wide mechanisms that control transcription factor binding and function. *Cell* **173**, 1796-1809.e17 (2018).

14. Chowdhary, K., Léon, J., Mathis, D. & Benoist, C. An integrated transcription factor framework for Treg identity and diversity. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2411301121 (2024).

15. van der Veeken, J. *et al.* The transcription factor Foxp3 shapes regulatory T cell identity by tuning the activity of trans-acting intermediaries. *Immunity* **53**, 971-984.e5 (2020).

16. Grondin, B. *et al.* c-Jun homodimers can function as a context-specific coactivator. *Mol. Cell. Biol.* **27**, 2919–2933 (2007).

17. Zhao, X. *et al.* PU.1-c-Jun interaction is crucial for PU.1 function in myeloid development. *Commun. Biol.* **5**, 961 (2022).

18. Zhong, Y. *et al.* Hierarchical regulation of the resting and activated T cell epigenome by major transcription factor families. *Nat. Immunol.* **23**, 122–134 (2022).

19. Mostafavi, S. *et al.* A scalable approach to investigating sequence-to-expression prediction from personal genomes. *Research Square* (2025) doi:10.21203/rs.3.rs-6030940/v1.

20. Drusinsky, S., Whalen, S. & Pollard, K. S. Deep-learning prediction of gene expression from personal genomes. *bioRxiv* (2024) doi:10.1101/2024.07.27.605449.

21. Holt, J., Huang, S., McMillan, L. & Wang, W. Read annotation pipeline for high-throughput sequencing data. in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics* (ACM, New York, NY, USA, 2013). doi:10.1145/2506583.2506645.