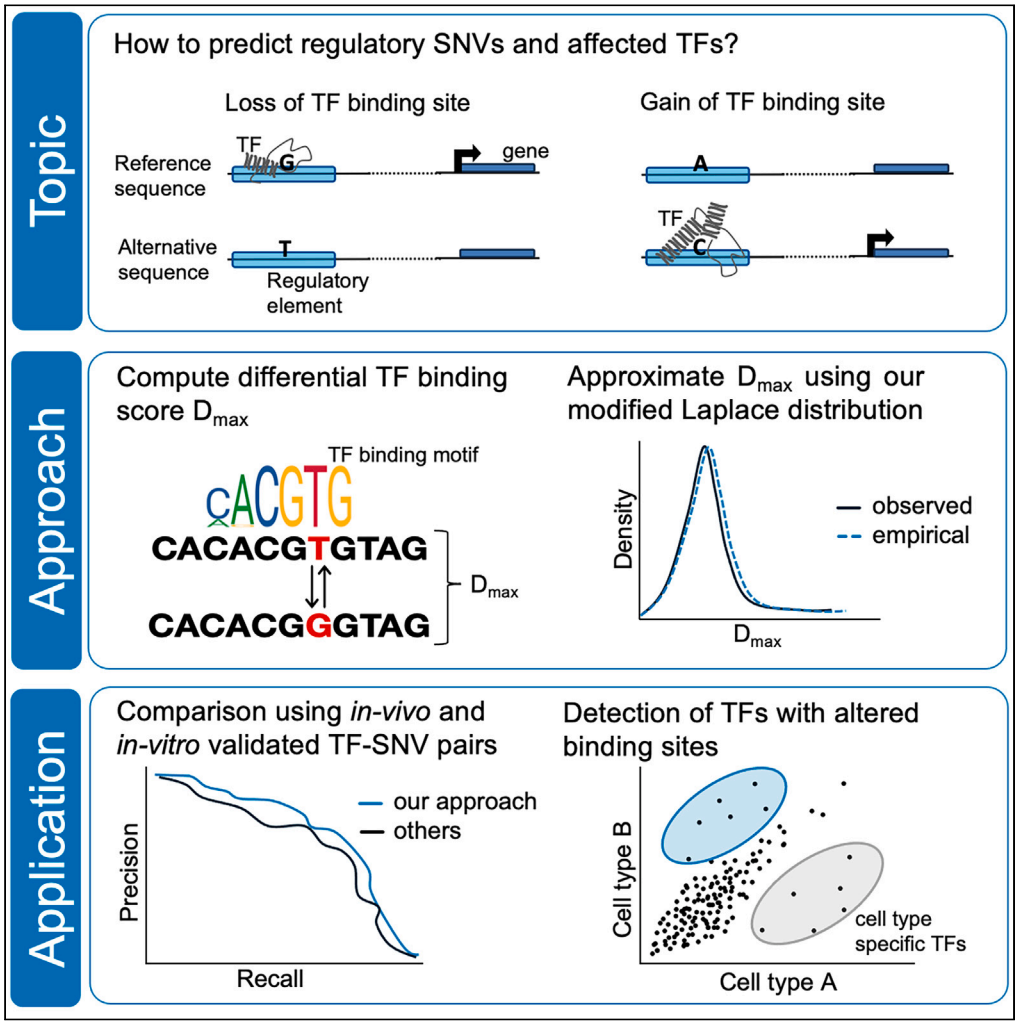
iScience

CellPress
OPEN ACCESS

Article

A statistical approach for identifying single nucleotide variants that affect transcription factor binding



Nina Baumgarten, Laura Rumpf, Thorsten Kessler, Marcel H. Schulz

marcel.schulz@em.uni-frankfurt.de

Highlights

Single nucleotide variants (SNVs) may affect transcription factor (TF) binding

Fast statistical approach to assess significance of differential TF binding for SNVs

Validate new approach on *in vitro* and *in vivo* TF binding assays

Applications on GWAS SNVs and large eQTL studies illustrate utility

Baumgarten et al., iScience *27*, 109765
May 17, 2024
https://doi.org/10.1016/j.isci.2024.109765

iScience

CellPress
OPEN ACCESS

Article

A statistical approach for identifying single nucleotide variants that affect transcription factor binding

Nina Baumgarten,[1,2,3,4] Laura Rumpf,[1,2,3,4] Thorsten Kessler,[5,6] and Marcel H. Schulz[1,2,3,4,7,*]

SUMMARY

Non-coding variants located within regulatory elements may alter gene expression by modifying transcription factor (TF) binding sites, thereby leading to functional consequences. Different TF models are being used to assess the effect of DNA sequence variants, such as single nucleotide variants (SNVs). Often existing methods are slow and do not assess statistical significance of results. We investigated the distribution of absolute maximal differential TF binding scores for general computational models that affect TF binding. We find that a modified Laplace distribution can adequately approximate the empirical distributions. A benchmark on *in vitro* and *in vivo* datasets showed that our approach improves upon an existing method in terms of performance and speed. Applications on eQTLs and on a genome-wide association study illustrate the usefulness of our statistics by highlighting cell type-specific regulators and target genes. An implementation of our approach is freely available on GitHub and as bioconda package.

INTRODUCTION

Large population studies, such as genome-wide association studies (GWASs), allow us to link genetic variants to phenotypes and lead to the discovery of novel genetic loci harboring genes that play causal roles in disease progression.[1,2] These strategies are indispensable to allow novel mechanistic studies on how pathways and cell types contribute to a disease, which may open novel avenues for therapeutic approaches in the future. However, GWASs indicate that a huge percentage of genomic variants appear in non-coding genomic regions; thus, they do not directly affect the coding region of a gene.

Transcription factors (TFs) are DNA-binding proteins that recognize short DNA patterns and thereby regulate gene expression. TF binding sites occur often enriched in regulatory elements such as promoters or enhancers. Non-coding genetic variants, such as single nucleotide variants (SNVs), localized in regulatory elements can affect gene expression by modifying transcription factor binding sites (TFBS). Several studies have reported the resulting functional consequences (reviewed for instance by F. Zhang and J.R. Lupski[3]). Therefore, methods pinpointing to such regulatory SNVs (rSNVs) are a topic of current interest.

Several methods exist that successfully highlight rSNVs based on epigenetic information such as open chromatin data, TF- and histone ChIP-seq data without taking into account which TF might be affected.[4–7]

In contrast, methods that evaluate the effect of an SNV on a TFBS rely on the ability to describe the binding behavior of a transcription factor (TF) to assess the difference induced by a non-coding SNV. The binding behavior of a TF can be described *in vitro* using high-throughput methods such as protein-binding microarrays (PBMs)[8] or SELEX,[9] or *in vivo* using ChIP-based techniques.[10,11] The identified TF binding preferences are summarized in a TF model, most prominently position weight matrices (PWMs).[12] However, there are other more complex TF models utilizing a Bayesian network or Markov models (reviewed by Valentina B.[13]). Other proposed models are for instance the binding energy model (BEM)[14] or the transcription factor flexible model (TFFM),[15] the latter being available within the JASPAR database.[16] Furthermore, there is the SLIM model,[17] which provides graphical visualization similar to the sequence logos of PWMs, and methods based on deep convolutional neural networks such as DeepBind[18] and BPNet.[19]

Computational approaches have been developed to evaluate the effect of an SNV on the binding sites of a TF. Among them is *GERV*,[20] a *k*-mer based approach that learns *de novo* TF binding based on open chromatin and TF ChIP-seq data. To evaluate the impact of an SNV, they computed the difference in the predicted read counts for the two allelic variants of an SNV. A more recently published method is *FABIAN-variants*.[21] The authors determined a differential TF binding score not only based on PWMs but also on TFFMs and allowed to

[1]Institute of Cardiovascular Regeneration, Goethe University, 60590 Frankfurt am Main, Germany
[2]Institute for Computational Genomic Medicine, Goethe University, 60590 Frankfurt am Main, Germany
[3]Institute for Computer Science, Goethe University, 60590 Frankfurt am Main, Germany
[4]German Center for Cardiovascular Research, Partner Site Rhein-Main, 60590 Frankfurt am Main, Germany
[5]German Heart Centre Munich, Department of Cardiology, School of Medicine and Health, Technical University of Munich, 80636 Munich, Germany
[6]German Centre for Cardiovascular Research, Partner Site Munich Heart Alliance, 80636 Munich, Germany
[7]Lead contact
*Correspondence: marcel.schulz@em.uni-frankfurt.de
https://doi.org/10.1016/j.isci.2024.109765

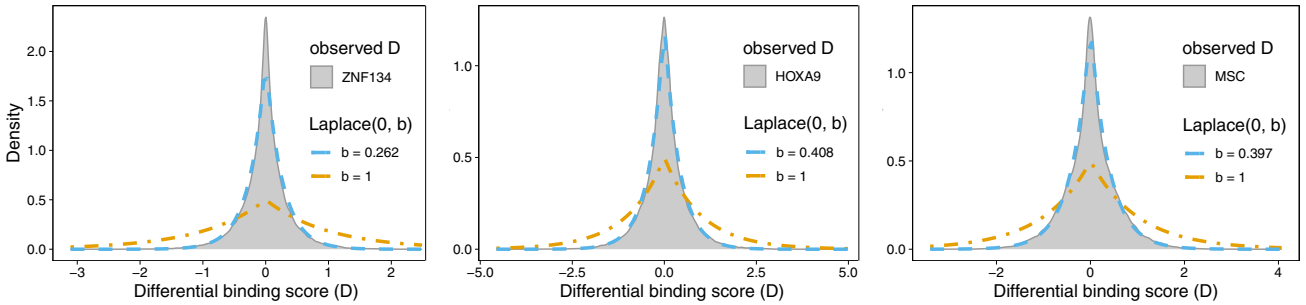


Figure 1. Differential TF binding score distributions for three TFs of different lengths

The distributions of D for the PWMs of the ZNF134 (length 22), HOXA9 (length 8), and MSC (length 10) TFs were compared to the $L(0,1)$ (orange) and $L(0,b)$ (black) distributions. The scale parameter b is estimated for each TF model separately.

take TFBS from epigenetic data into account. In comparison, QBIC-Pred[22,23] utilizes *in vitro* universal PBM data to determine the TF binding behavior with a *k*-mer based model using ordinary least squares (OLS). The authors score the effect of an SNV based on the parameters of the OSL Z score and additionally provide a p value for their score. Further, there are methods such as *sTRAP*,[24] *is-rSNP*,[25] or *atSNP*[26] that rely solely on TF models (usually PWMs) and the DNA sequence itself. To evaluate the effect of an SNV on TF binding sites, different statistical approaches have been introduced by these methods. *sTRAP* ranks TFs directly based on their differential TF binding score for the wildtype and the alternative alleles of an SNV, whereas *is-rSNP* and *atSNP* provide p values for their differential TF binding scores.

In general, we noticed that only a few methods provide a statistical significance for their introduced differential TF binding scores. However, this is necessary to determine whether a score is significantly different from the commonly assumed null hypothesis that the SNV does not affect TF binding given a TF model. Furthermore, if the TF binding score is represented by a p value, the values are directly comparable between the TFs, which is often not possible for the scores themselves. When the methods provide a p value, their statistics are dependent on their underlying TF model. For instance, *QBIC-Pred* derives their test statistics based on OLS estimation of k-mers, whereas *atSNP* assumes that the scores follow a multinomial distribution to model PWMs. *is-rSNP* is, to the best of our knowledge, the only method that allows us to compute a p value independent of the TF model. However, they determined the exact p value distribution for differential TF binding scores by assessing all possible single base changes, resulting in a quadratic algorithm that is prohibitive for large datasets.

In this work, we introduce a fast and accurate approach to determine the statistical significance of the differential TF binding score of general TF models. To do so, we examined the distribution of the maximal differential TF binding scores and found that it could be approximated well by a modified Laplace distribution. By using the modified Laplace distribution, we can derive a p value for the maximal differential TF binding score in constant time. Using experimentally validated TF-SNV pairs, we showed that our approach improved upon the previously established method *atSNP* while being an order of magnitude faster. As applications, we present the identification of cell type-specific TFs whose binding sites are perturbed by eQTLs in lymphocytes and fibroblasts. Further, we showcase how to combine our approach with publicly available regulatory elements (REMs), derived from epigenomic data, to pinpoint potential target genes affected by rSNVs in an atherosclerosis GWAS.

RESULTS

The differential TF binding score approximately followed the $L(0,b)$ distribution

Consider the sequences S^1 and S^2, each holding an allelic variant of a given SNV, and a TF model M that characterizes the binding behavior of a TF. We defined the differential TF binding score (D) between S^1 and S^2 as the log-ratio of the p-values of the TF binding scores (see STAR Methods section 'Definition of the problem').

Even though our statistical approach is designed for general TF models, we had to decide on a concrete TF model to investigate the distribution of the differential TF binding scores. We represent the TF models with PWMs, which are widely used and easily accessible for hundreds of human TFs, and for which other methods exist for comparison. We computed D for 200,000 SNVs randomly sampled from the dbSNP database for PWMs of different lengths. In Figure 1, the resulting distributions of the differential TF binding scores for three TFs are visualized. For comparison, the Laplace(0,1) ($L(0,1)$) distribution is also displayed. Even though we argued that theoretically, D should be $L(0,1)$ distributed, we observed that our experiments suggest otherwise. The sequences of S^1 and S^2 differ by only one letter at the position of the SNV. Hence, the TF binding scores for S^1 and S^2 do not change much, especially if the SNV does not affect the binding site. Using a Chi-square test, we verified that the p-values of the TF binding scores of S^1 and S^2 are not independent of each other (p-value ≤ 0.05 for all 817 PWMs of our motif set, see STAR Methods, section 'Evaluation of the independence of the TF binding score of the wildtype and alternative allele'). Consequently, we observed that D is close to 0 more often than one would expect for independent p-values $L(0,1)$ distributed scores. However, we empirically observed that D can be approximated by a $L(0,b)$ distribution with a scale parameter b fitted for each TF model M (see Figure 1, black curves). As we shall see in a moment, our findings in the next section support this fact, aligning with our derivation in the section evaluation of our approach on experimentally validated TF-SNV pairs.

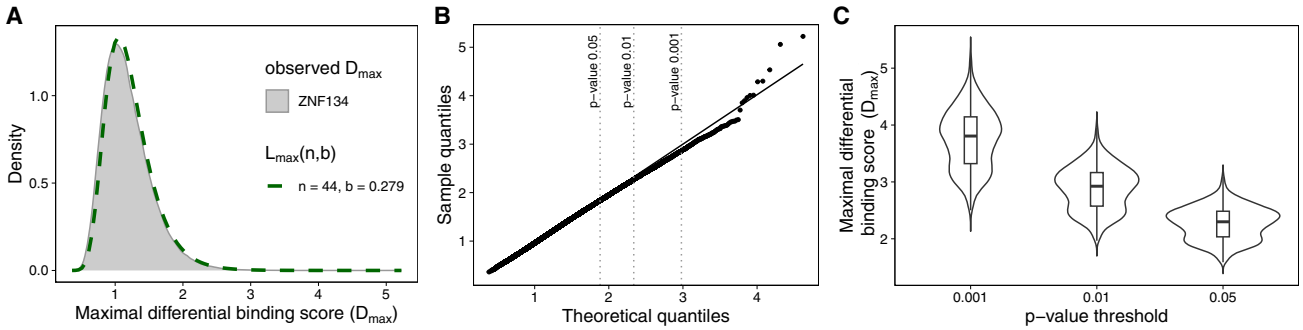


Figure 2. Distribution of D_{max} values

(A) Distributions of observed D_{max} values for ZNF134 in comparison to the $L_{max}(n, b)$ distribution; parameter values for n and b (fitted) are shown in the plot.

(B) A Quantil-Quantil plot for the data shown in B, is visualized, where the y axis represents the quantiles from the $L_{max}(n, b)$ distribution and the x axis the quantiles from the observed D_{max} values. The dotted vertical lines mark the p-value thresholds 0.05, 0.01, and 0.001.

(C) Violin plot for observed D_{max} values for 818 PWMs for different p-value thresholds.

The maximal differential TF binding score distribution differed among multiple TFs

In the previous section, we assumed that we are interested in D for a specific sequence position; thus, the sequence S and the TF model had the same length. However, this assumption is unrealistic. Usually, the sequence is longer than the TF model. As explained in the STAR Methods section 'Definition of the problem', we computed D for all subsequences overlapping the SNV to identify where the binding affinity of a TF is most affected and derived the absolute maximal differential TF binding score (D_{max}).

Since the resulting distribution of D_{max} is dependent on the TF model used, it is not possible to compare D_{max} between different TFs directly. However, the usual application of such a statistic is to evaluate and compare the effects of several hundred TFs on an SNV set. To allow this comparison, we aimed to compute a p-value for D_{max}. We first tried to fit known distributions to the observed D_{max} values (using the R package gamlss[27]), but this approach did not result in the same distribution for multiple TFs. As an alternative, we derived the $L_{max}(n, b)$ distribution (see STAR Methods section 'Derivation of the distribution of the maximal differential TF binding scores'). The parameter n depends on the length of the TF model and is given by the number of sequence windows tested by the model and the scale parameter b is estimated for each TF separately using the MLE from Equation 9.

In Figure 2A, the comparison between the distribution of observed D_{max} values for the TF ZNF134 and $L_{max}(n, b)$ shows exemplary that D_{max} can be adequately approximated by $L_{max}(n, b)$. Additionally, in the corresponding Quantil-Quantil (QQ) plot (see Figure 2B), one can observe that for commonly used p-value thresholds, the $L_{max}(n, b)$ distribution accurately approximates the observed D_{max} values. Furthermore, we evaluated whether observed D_{max} values for randomly sampled SNVs follow our modified Laplace distribution, applying a Kolmogorov-Smirnov test separately for each TF motif. For 635 out of 817 TF motifs the H_0 hypothesis, that the observed D_{max} values follow the $L_{max}(n, b)$ distribution, could not be rejected (p-value >0.05). We concluded, that for the majority of the considered TF motifs our modified Laplace distribution can adequately approximate the empirical distributions. To better understand which characteristic of a PWM has an impact on the approximation quality, we evaluated the p-values from the Kolmogorov-Smirnov test in relation to the TF-family, the TF motif length and the overall entropy. We noticed that for TF-families such as *AP-2* or Brachyury-related factors for which the H_0 hypothesis was generally rejected, the TF motifs had high entropy bases at central motif positions (Figures S2A and S3). Additionally, we observed that neither the motif length nor the overall entropy is associated to how well the $L_{max}(n, b)$ distribution approximates the distribution of the observed D_{max} values (Spearman's correlation coefficient: TF motif length: R = −0.1; entropy: R = −0.039 Figures S2B and S2C).

Using the CDF of the $L_{max}(n, b)$ distribution, we can compute a p-value for D_{max} values, which are then comparable between multiple TFs. Figure 2C visualizes for different p-value thresholds the corresponding D_{max} values for the 817 TF motifs used in our analyses. The D_{max} values for a fixed p-value threshold differ from each other. For instance, for the p-value threshold 0.001 D_{max} values between 2.5 and 5.2 were observed, supporting the need for a p-value computation to compare the results between different TFs.

Evaluation of our approach on experimentally validated TF-SNV pairs

To analyze the performance of our approach, we collected TF-SNV pairs from data sources with experimental evidence that the TF is affected by the SNV. We gathered ASB events, which were defined using TF ChIP-seq data and data collected from SNP-SELEX experiments, an *in vitro* measurement of the TF-DNA interaction strength for each allele of the SNV (see STAR Methods section 'Collecting allele-specific binding events' and section 'Collecting SNP-SELEX data').

To evaluate how well a method can distinguish experimentally validated TF-SNV pairs from those that are not validated, we defined a classification task, in which the positive class contains the collected TF-SNV pairs. The negative class was defined as all possible combinations of considered TFs and SNVs, excluding those from the positive class. The ASB dataset consists of 368 positively labeled TF-SNV pairs and 4,036 negative, and the SNP-SELEX data of 1,814 positive and 58,162 negative, respectively.

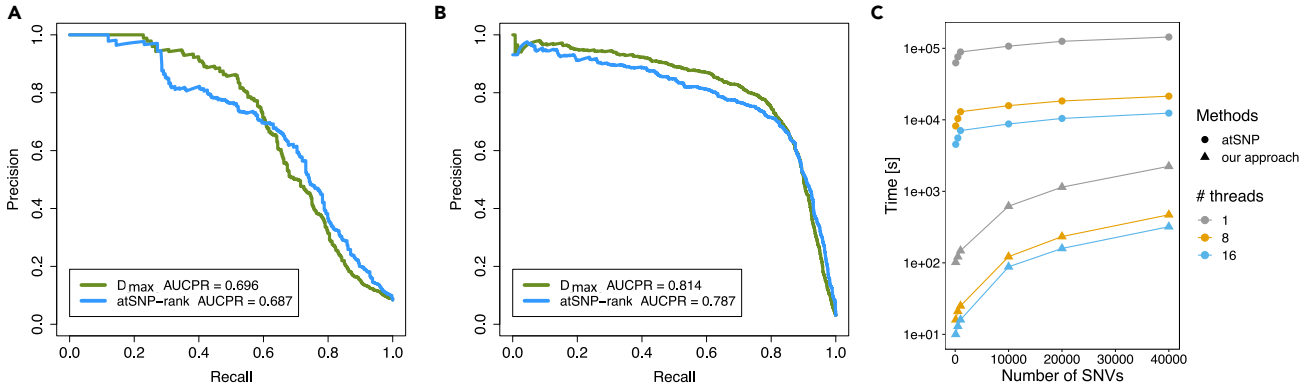


Figure 3. Comparison between our approach and *atSNP*

Precision-recall curve for the ASB events (A) and the SNP-SELEX dataset (B) for the D_{max} *p*-value and the rank-based *p*-value of *atSNP*.

(C) Runtime analysis of our approach and *atSNP*. The lineplot shows the runtimes of both methods (y axis, log10-scale) for randomly sampled SNV sets of different sizes (x axis) for different numbers of threads.

Using these datasets we do not only want to evaluate our approach, we also want to compare our method to the previously published method *atSNP*.[26] To the best of our knowledge, *atSNP* is the fastest PWM-based method currently available that also provides a *p*-value for the differential TF binding score. *is-rSNP* is too slow to be applied to the large datasets considered here.[25] We applied our approach and *atSNP* for each of the two datasets separately and evaluated the performance of the D_{max} *p*-value in comparison to *atSNP*'s rank-based *p*-value, which is recommended by the authors. The rank-based *p*-value indicates whether the log-odds ratio of the *p*-values of the TF binding score for the wildtype and the alternative allele significantly differs from what one would expect by chance. Additionally, they provided a diff-based *p*-value that directly evaluates the changes in the TF binding score directly. However, the authors of *atSNP* mentioned that the diff-based *p*-value is not reliable, and our experiments confirmed a poor performance (data not shown).

Figure 3 Aand 3B show the resulting precision-recall curves and the AUCPR for the ASB events (Figure 3A) and the SNP-SELEX data (Figure 3B). Even though the negatively labeled TF-SNV sets are several times larger than the positive sets, a reasonable AUCPR is reached for both datasets, indicating the quality of our method and the rank-based *p*-value of *atSNP*. The AUCPR of the D_{max} *p*-value is improved in comparison to the rank-based *p*-value of *atSNP*. For the ASB events, the AUCPR of the D_{max} *p*-value is 0.9% higher than that of the rank-based *p*-value. For the SNP-SELEX dataset, the improvement of the D_{max} *p*-value is 2.7% in comparison to the rank-based *p*-value. Additionally, for both datasets, the difference of the area under the ROC curves between the D_{max} *p*-value compared to the rank-based *p*-value is significant (p-value ≤ 0.05, ASB data: p-value = 0.00541, SNP-SELEX data: p value $9.457e^{-5}$) according to the method of DeLong et al. (see STAR Methods section 'Method performance evaluation').

We compared the runtime of our approach and *atSNP* for 6 randomly sampled SNV sets of sizes between 100 and 40,000 SNVs for 817 TF motifs using 1, 8, and 16 threads. As shown in Figure 3C, our method is between 623 (500 SNVs on 1 threads) and 38 (40,000 SNVs on 16 threads) times faster than *atSNP*. To determine a reasonable D_{max} *p*-value cutoff, we computed the *F*1-score for the ASB and SNP-SELEX data. The resulting D_{max} *p*-value cutoff of 0.01 is used in the following analyses.

Identification of TFs with altered binding sites induced by genetic variants mediating gene expression

Identifying cell type-specific regulators with modified binding behavior induced by genetic variants associated to genes might be helpful for revealing regulatory pathways or molecular mechanisms involved. Therefore, we aimed to identify TFs more often affected by a set of SNVs than one would expect from random SNV data. Given the speed of our approach, such analyses can be performed in a reasonable amount of time even for large SNV sets.

For example, we analyzed 14,722 eQTLs associated with lymphocytes and 45,917 eQTLs associated with fibroblasts. For each eQTL, we computed the D_{max} *p*-value for 817 human TFs. We counted how often each binding site was significantly (D_{max} *p*-value ≤ 0.01) affected by each TF across all the eQTLs. To identify TFs more often affected by the eQTLs than expected, we computed an odds ratio between the TF counts of the eQTLs and TF counts on 1,000 SNV sets of the same size as the eQTL data (see STAR Methods section 'eQTL analysis').

If the odds ratio of a TF is >2, we assumed that the TF is enriched since it occurs 2 times more often than expected. For fibroblasts, 57 TFs were identified, and for lymphocytes, 66 TFs were identified (full list per cell type is given in our ZENODO data repository (https://doi.org/10.5281/zenodo.7588272)). Figure 4A visualizes the odds ratios of the analyzed TFs of fibroblasts against lymphocytes, and the enriched TFs with an absolute difference in the odds ratio >1 are labeled. For several of the cell type-specific TFs, we found evidence in the literature. For instance, it has been shown that in skin fibroblasts, EGR3 can upregulate genes associated with tissue remodeling and wound healing.[28] The expression of the TF SNAI1 in cancer-associated fibroblasts is directly associated with chemoresistance via the mediation of the extracellular matrix.[29] For lymphocytes, we identified several highly expressed TFs from the Ets-related TF family, among others, with additional evidence from the literature. For instance, in mice, it has been shown that the TFs ELK1 and ELK4 function redundantly to restrict the

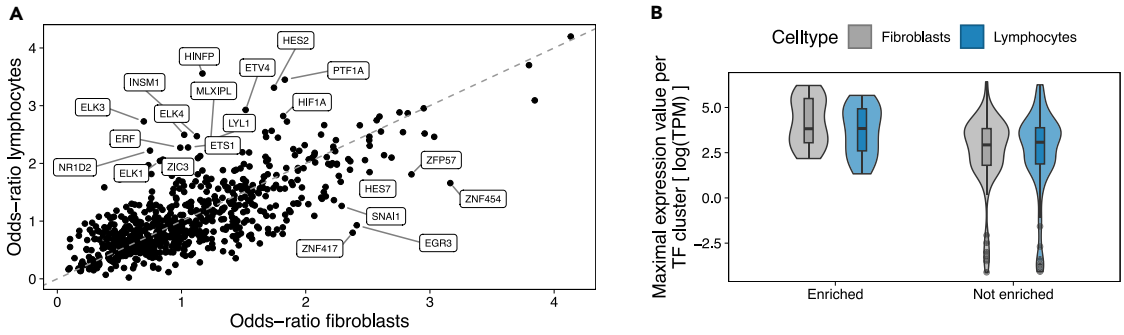


Figure 4. Identification of TFs with altered binding sites by genetic variants mediating gene expression in lymphocytes and fibroblasts

(A) Scatterplot showing the computed odds ratio of a TF for eQTLs from fibroblasts (x axis) against lymphocyte eQTLs (y axis). The TFs are labeled if the odds ratio >2 and the absolute difference in the odds ratio between the two cell types are >1.

(B) Violin-boxplot illustrating the differences in terms of expression values between the TFs that are more often affected by the eQTLs (enriched odds ratio) than expected (not enriched odds ratio) for two different cell types (coloring).

generation of innate-like CD8^{+} T-cells.[30] Recently, Tsiomita et al. showed that ERF is a potential regulator of T lymphocyte maturation.[31] Further, for innate lymphoid cells (ILC), which are a population of lymphocytes, it can be shown that ELK3 is regulated by the circRNA circTmem241, and that the knockdown of ELK3 significantly decreases the number of ILCs.[32]

We cannot provide literature evidence for all cell type-specific TFs; therefore, we wanted to evaluate the expression values of enriched TFs in comparison to TFs not enriched (Figure 4B). Since motifs from the same TF family are often highly similar, resulting in redundancy in our motif collection, we determined the expression value of the highest expressed TF per cluster. According to a one-sided Wilcoxon rank-sum test, the differences in gene expression between the two groups were significant (fibroblasts: $p = 0.0016$, lymphocytes $p = 0.028$). Thus, many of the TFs that show strong enrichment in altered binding behavior for eQTLs of the two cell types, are likely to mediate gene expression differences and are interesting candidates for further cell type-specific investigations.

Identification of candidate target genes affected by the rSNVs in an atherosclerosis GWAS

Even if no eQTL data are available, one might be interested in identifying target genes that are regulated by the TFs with modified TFBSs caused by rSNVs. By combining our approach with publicly available regulatory elements (REMs), we were able to identify candidate target genes. To illustrate this application, we have applied our approach to a set of 4,326 lead and proxy SNVs from a GWAS for the disease atherosclerosis. To associate the resulting rSNVs to target genes, we used 2.4 million REMs linked to target genes downloaded from the EpiRegio webserver[33,34] (see STAR Methods section 'Application atherosclerosis GWAS', Figure S4).

Among the genes affected by at least one rSNV were the *ABO* and *CELSR2* loci, which were previously associated with coronary artery disease (CAD) according to a GWAS.[35] CAD can be induced by atherosclerosis occurring in the large vessels supplying oxygen to the myocardium. Both loci were further implicated to play a role in lipid metabolism, with *ABO*, for example, being associated with total cholesterol[36] and *CELSR2* representing a candidate gene at the chromosome 1p13 CAD/cholesterol locus.[37] Phenome-wide association results are depicted in Figure S4. The association with both, CAD and cholesterol levels, renders an involvement of hypercholesterolemia as the likely responsible intermediate phenotype. Furthermore, the binding activity of the TF OSR1 is predicted to be affected by the rSNV rs629301, which is linked to the gene *CELSR2*. The *OSR1* gene itself was associated with the traditional CAD risk factor blood pressure.[38] The predicted functional connection between *OSR1* and *CELSR2* might therefore indicate an interaction between two traditional risk factors via genomic variant.

DISCUSSION

Throughout the manuscript, we presented a new statistical approach to identify regulatory SNVs affecting the binding sites of TFs. We aimed to provide a method that allows us to compute statistical significance for general TF models. We compared our new approach to the previous method *atSNP* in terms of performance and runtime. We demonstrated that our new approach is at least as accurate as *atSNP* is. By comparing the runtimes of both methods, we show that our approach is extremely fast for large sets of SNVs and hundreds of TFs.

To test our approach, we had to specify a TF model, and we decided to use PWMs, which are commonly used and available for hundreds of human TFs. The selection of experimentally validated TF-SNV pairs is affected by this decision. We excluded those TF-SNV pairs that did not overlap with a predicted TFBS. Given the limitations of PWMs, it would not have been possible to predict these pairs correctly neither for our statistical approach nor for *atSNP*. Thus, excluding these pairs allows us to precisely evaluate how well both methods detect differential TF binding. Thus our results do not allow an assessment, of which type of TF model works best on a given dataset.

Our approach does not directly take into account cell type- or tissue-specific information. However, a useful approach is to exclude motifs from TFs not expressed in the cell type or tissue of interest to reduce the number of false-positive predicted rSNVs. Further, one can easily combine the predicted rSNVs with other epigenomic data, as shown in the application for the atherosclerosis GWAS or in a recently published study, where we identified non-coding disease genes.[39]

Additionally, we want to emphasize that we combined several TFs within one dataset to compare the methods, resulting in a highly imbalanced dataset. In our opinion, this approach is a more realistic evaluation setting than evaluating the TFs one by one, as is often done.

Another advantage of our approach is that it has no significant additional runtime compared to widespread score-based approaches that do not assess significance. The scale parameter b needs to be precomputed only once for the TF motifs used. On our github repository (https://github.com/SchulzLab/SNEEP), we provide our approach implemented in C++ as an easy-to-install bioconda package, also including the precomputed scales for the 817 TF motifs used for the presented analyses.

We believe that our approach will be helpful for identifying novel rSNVs and thereby contribute to the understanding of molecular mechanisms leading to various traits and diseases.

Limitations of the study

In this study, we decided to validate our statistical approach using PWMs. However, we believe, that our approach can in principle be applied to any other TF model other than PWMs once a p value for the TF binding score is computed. This can be done using Monte Carlo sampling if not otherwise available. Thus, an interesting research direction is to explore whether the $L_{max}(n, b)$ distribution fits the observed D_{max} values of other TF models such as TFFMs or SLIM models. All TF models have advantages and disadvantages; thus, the prediction quality for a TF could be improved by combining the results of several TF models such as PWMs, TFFM, or SLIM model with each other (e.g., smallest p value over all tested TF models or Fisher's meta p value aggregation method). However, further research is needed to evaluate how well our statistical approach works with other TF models than PWMs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Definition of the problem
  - Estimating the distribution of differential TF binding scores
  - Derivation of the distribution of the maximal differential TF binding scores
  - Computation of TF binding scores with position weight matrices
  - Derive exact TF binding score distribution
  - Evaluation of the independence of the TF binding score of the wildtype and alternative allele
  - Fitting the scale parameter b
  - Statistical evaluation of model fit
  - Collecting allele-specific binding events
  - Collecting SNP-SELEX data
  - Method performance evaluation
  - eQTL analysis
  - Application atherosclerosis GWAS
  - Details about the used commands to run *atSNP* and our approach

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109765.

ACKNOWLEDGMENTS

We thank the GTEx Portal and the NHGRI-EBI GWAS catalog for providing the data used in the applications and Fatemeh Behjati Ardakani for proofreading the manuscript. This work has been supported by the DZHK (German Centre for Cardiovascular Research, IDs: 81Z0200101 and 81X2200151), the Cardio-Pulmonary Institute (CPI) [EXC 2026] ID: 390649896, the DFG SFB (TRR 267) Noncoding RNAs in the cardiovascular system (Z03, project ID 403584255), DFG SFB1531 (S03, project ID 456687919) and the HESSIAN Center for AI (hessian.AI).

AUTHOR CONTRIBUTIONS

Conceptualization, N.B. and M.H.S.; methodology, N.B. and M.H.S.; software, N.B. and L.R.; validation, N.B.; investigation, N.B., T.K., and M.H.S.; writing – original draft, N.B., M.H.S., and T.K.; writing – review and editing, N.B. and M.H.S.; funding acquisition, M.H.S.; supervision, M.H.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 14, 2023
Revised: January 30, 2024
Accepted: April 15, 2024
Published: April 18, 2024

REFERENCES

1. Aragam, K.G., Jiang, T., Goel, A., Kanoni, S., Wolford, B.N., Atri, D.S., Weeks, E.M., Wang, M., Hindy, G., Zhou, W., et al. (2022). Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. Nat. Genet. *54*, 1803–1815. https://doi.org/10.1038/s41588-022-01233-6.
2. Ma, Q., Shams, H., Didonna, A., Baranzini, S.E., Cree, B.A.C., Hauser, S.L., Henry, R.G., and Oksenberg, J.R. (2023). Integration of epigenetic and genetic profiles identifies multiple sclerosis disease-critical cell types and genes. Commun. Biol. *6*, 342. https://doi.org/10.1038/s42003-023-04713-5.
3. Zhang, F., and Lupski, J.R. (2015). Non-coding genetic variants in human disease: Figure 1. Hum. Mol. Genet. *24*, R102–R110. https://doi.org/10.1093/hmg/ddv259.
4. Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961. https://doi.org/10.1038/ng.3331.
5. Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999. https://doi.org/10.1101/gr.200535.115.
6. Tiffany, A., Yang, L., Gazal, S., et al. (2019). IMPACT: Genomic annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors. Am. J. Hum. Genet. *104*, 879–895. https://doi.org/10.1016/j.ajhg.2019.03.012.
7. Chen, L., Wang, Y., and Zhao, F. (2022). Exploiting deep transfer learning for the prediction of functional non-coding variants using genomic sequence. Bioinformatics *38*, 3164–3172. https://doi.org/10.1093/bioinformatics/btac214.
8. Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435. https://doi.org/10.1038/nbt1246.
9. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873. https://doi.org/10.1101/gr.100552.109.
10. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The human transcription factors. Cell *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.
11. He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol. *33*, 395–401. https://doi.org/10.1038/nbt.3121.
12. Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16–23. https://doi.org/10.1093/bioinformatics/16.1.16.
13. Valentina, B. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. Front. Genet. *7*. https://doi.org/10.3389/fgene.2016.00024.
14. Yue, Z., Ruan, S., Pandey, M., and Stormo, G.D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. Genetics *191*, 781–790. https://doi.org/10.1534/genetics.112.138685.
15. Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput. Biol. *9*, e1003214. https://doi.org/10.1371/journal.pcbi.1003214.
16. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. https://doi.org/10.1093/nar/gkab1113.
17. Keilwagen, J., and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. Nucleic Acids Res. *43*, e119. https://doi.org/10.1093/nar/gkv577.
18. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838. https://doi.org/10.1038/nbt.3300.
19. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6.
20. Zeng, H., Hashimoto, T., Kang, D.D., and Gifford, D.K. (2016). GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. Bioinformatics *32*, 490–496. https://doi.org/10.1093/bioinformatics/btv565.
21. Steinhaus, R., Robinson, P.N., and Seelow, D. (2022). FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. Nucleic Acids Res. *50*, W322–W329. https://doi.org/10.1093/nar/gkac393.
22. Martin, V., Zhao, J., Afek, A., Mielko, Z., and Gordân, R. (2019). QBiC-pred: quantitative predictions of transcription factor binding changes due to sequence variants. Nucleic Acids Res. *47*, W127–W135. https://doi.org/10.1093/nar/gkz363.
23. Zhao, J., Li, D., Seo, J., et al. (2017). Quantifying the impact of non-coding variants on transcription factor-DNA binding. In Lecture Notes in Computer Science (Springer International Publishing), pp. 336–352.
24. Manke, T., Heinig, M., and Vingron, M. (2010). Quantifying the effect of sequence variation on regulatory interactions. Hum. Mutat. *31*, 477–483. https://doi.org/10.1002/humu.21209.
25. Macintyre, G., Bailey, J., Haviv, I., and Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics *26*, i524–i530. https://doi.org/10.1093/bioinformatics/btq378.
26. Zuo, C., Shin, S., and Keleş, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics *31*, 3353–3355. https://doi.org/10.1093/bioinformatics/btv328.
27. Rigby, R.A., and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape (with discussion). J. Roy. Stat. Soc.: Series C (Applied Statistics) *54*, 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x.
28. Fang, F., Shangguan, A.J., Kelly, K., Wei, J., Gruner, K., Ye, B., Wang, W., Bhattacharyya, S., Hinchcliff, M.E., Tourtellotte, W.G., and Varga, J. (2013). Early growth response 3 (egr-3) is induced by transforming growth factor-β and regulates fibrogenic responses. Am. J. Pathol. *183*, 1197–1208. https://doi.org/10.1016/j.ajpath.2013.06.016.
29. Galindo-Pumariño, C., Collado, M., Castillo, M.E., Barquín, J., Romio, E., Larriba, M.J., Muñoz de Mier, G.J., Carrato, A., de la Pinta, C., and Pena, C. (2022). SNAI1-expressing fibroblasts and derived-extracellular matrix as mediators of drug resistance in colorectal cancer patients. Toxicol. Appl. Pharmacol. *450*, 116171. https://doi.org/10.1016/j.taap.2022.116171.
30. Maurice, D., Costello, P., Sargent, M., and Treisman, R. (2018). ERK signaling controls innate-like CD8+ t cell differentiation via the ELK4 (SAP-1) and ELK1 transcription factors. J. Immunol. *201*, 1681–1691. https://doi.org/10.4049/jimmunol.1800704.
31. Tsiomita, S., Liveri, E.M., Vardaka, P., Vogiatzi, A., Skiadaresis, A., Saridis, G., Tsigkas, I., Michaelidis, T.M., Mavrothalassitis, G., and Thyphronitis, G. (2022). ETS2 repressor factor (ERF) is involved in t lymphocyte maturation acting as regulator of thymocyte lineage commitment. J. Leukoc. Biol. *112*, 641–657. https://doi.org/10.1002/jlb.1a0720-439r.
32. Liu, N., He, J., Fan, D., Gu, Y., Wang, J., Li, H., Zhu, X., Du, Y., Tian, Y., Liu, B., and Fan, Z. (2022). Circular RNA circTmem241 drives group III innate lymphoid cell differentiation

via initiation of elk3 transcription. Nat. Commun. *13*, 4711. https://doi.org/10.1038/s41467-022-32322-z.

33. Baumgarten, N., Hecker, D., Karunanithi, S., Schmidt, F., List, M., and Schulz, M.H. (2020). EpiRegio: analysis and retrieval of regulatory elements linked to genes. Nucleic Acids Res. *48*, W193–W199. https://doi.org/10.1093/nar/gkaa382.
34. Schmidt, F., Marx, A., Baumgarten, N., Hebel, M., Wegner, M., Kaulich, M., Leisegang, M.S., Brandes, R.P., Göke, J., Vreeken, J., and Schulz, M.H. (2021). Integrative analysis of epigenetics data identifies gene-specific regulatory elements. Nucleic Acids Res. *49*, 10397–10418. https://doi.org/10.1093/nar/gkab798.
35. CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nat. Genet. *45*, 25–33. https://doi.org/10.1038/ng.2480.
36. Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283. https://doi.org/10.1038/ng.2797.
37. Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. *40*, 161–169. https://doi.org/10.1038/ng.76.
38. Kato, N., Loh, M., Takeuchi, F., Verweij, N., Wang, X., Zhang, W., Kelly, T.N., Saleheen, D., Lehne, B., Leach, I.M., et al. (2015). Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. Nat. Genet. *47*, 1282–1293. https://doi.org/10.1038/ng.3405.
39. Zhu, C., Baumgarten, N., Wu, M., Wang, Y., Das, A.P., Kaur, J., Ardakani, F.B., Duong, T.T., Pham, M.D., Duda, M., et al. (2023). CVD-associated SNPs with regulatory potential reveal novel non-coding disease genes. Hum. Genom. *17*, 69. https://doi.org/10.1186/s40246-023-00513-4.
40. Larry Wasserman (2010). All of Statistics : A Concise Course in Statistical Inference (Springer).
41. Kotz, S., and Kozubowski, T.J.; Krzysztof Podgórski (2001). The Laplace Distribution and Generalizations (Birkhäuser Boston).
42. Beckstette, M., Homann, R., Giegerich, R., and Kurtz, S. (2006). Fast index based algorithms and software for matching position specific scoring matrices. BMC Bioinf. *7*, 389. https://doi.org/10.1186/1471-2105-7-389.
43. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
44. Shi, W., Fornes, O., Mathelier, A., and Wasserman, W.W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res. *44*, 10106–10116. https://doi.org/10.1093/nar/gkw691.
45. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.
46. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.
47. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987. https://doi.org/10.1093/nar/gkt1249.
48. Pape, U.J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between position frequency matrices with an application to clustering. Bioinformatics *24*, 350–357. https://doi.org/10.1093/bioinformatics/btm610.
49. Yan, J., Qiu, Y., Ribeiro dos Santos, A.M., Yin, Y., Li, Y.E., Vinckier, N., Nariai, N., Benaglio, P., Raman, A., Li, X., et al. (2021). Systematic analysis of binding of transcription factors to noncoding variants. Nature *591*, 147–151. https://doi.org/10.1038/s41586-021-03211-0.
50. Boytsov, A., Abramov, S., Makeev, V.J., and Kulakovskiy, I.V. (2022). Positional weight matrices have sufficient prediction power for analysis of noncoding variants. F1000Res. *11*, 33. https://doi.org/10.12688/f1000research.75471.3.
51. Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in r. Bioinformatics *31*, 2595–2597. https://doi.org/10.1093/bioinformatics/btv153.
52. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics *44*, 837. https://doi.org/10.2307/2531595.
53. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.
54. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.
55. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center LDACC—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx eGTEx groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al.. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.
56. Brown, A.A., Viñuela, A., Delaneau, O., Spector, T.D., Small, K.S., and Dermitzakis, E.T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. Nat. Genet. *49*, 1747–1751. https://doi.org/10.1038/ng.3979.
57. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., and Kastenmüller, G. (2015). SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics *31*, 1334–1336. https://doi.org/10.1093/bioinformatics/btu779.
58. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
59. Hadley, W. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data used in this study	This paper	https://doi.org/10.5281/zenodo.7588272
eQTL data	GTEx portal[55]	https://gtexportal.org/home/downloads/adult-gtex/qtl
Expression data for fibroblasts and lymphoctes	GTEx portal[55]	https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression
Pre-processed SNPs from dbSNP database	This paper	https://doi.org/10.5281/zenodo.4892591
Software and algorithms		
Our statistical approach	This paper	https://doi.org/10.5281/zenodo.10830009
atSNP	Zuo et al.[26]	https://github.com/keleslab/atSNP
Similarity measurement and clustering approach for PWMs	Pape et al.[48]	http://mosta.molgen.mpg.de/index.html
SNiPA	Arnold et al.[57]	https://www.snipa.org/snipa3/
PRROC (v.1.3.1)	Grau et al.[51]	https://cran.r-project.org/web/packages/PRROC/index.html
pROC (v. 1.18.0)	Robin et al.[53]	https://CRAN.R-project.org/package=pROC
Stats (v. 4,1,2)	R Foundation	https://cran.r-project.org/package=STAT
ggplot2 (v.3.4.0)	Wickham[59]	https://CRAN.R-project.org/package=ggplot2
gamlss	Rigby et al.[27]	https://CRAN.R-project.org/package=gamlss
R (v 4.1.2)	R Foundation	https://www.r-project.org
bedtools	Quinlan et al.[58]	https://bedtools.readthedocs.io/en/latest/index.html
Python (v. 3.8.10)	Python Software Foundation	https://www.python.org
G++, the GNU C++ Compiler (v. 9.4.0)	GCC team	https://gcc.gnu.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Marcel H. Schulz (marcel.schulz@em.uni-frankfurt.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Data: All data used to validate our approach is publicly available as Zenodo (https://doi.org/10.5281/zenodo.7588272). DOI is listed in the key resources table.
- Code: All original code has been deposited at Zenodo (https://zenodo.org/records/10830009) and is publicly available as of the date of publication. DOI is listed in the key resources table. Furthermore, details on how to run atSNP and our statistical approach are given in the STAR Methods Section details about the used commands to run atsnp and our approach. Additionally, we provide our statistical approach in our GitHub repository (https://github.com/SchulzLab/SNEEP) and as a bioconda package. Details on how to install and use our software are given at ReadTheDocs (https://sneep.readthedocs.io/en/latest/index.html)
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Definition of the problem

In the following, we explain how to evaluate the effect of a SNV on a TFBS. In the first problem definition, we explain how to compute the differential TF binding score for a fixed TF model position. The second definition describes a more general case in which the differential TF binding is computed for all sequences overlapping the SNV.

The differential TF binding problem

Let M be a general TF model of length m, $S = \{s_1, \ldots, s_m\}$ be a DNA sequence with $s_i \in \Sigma$ over the alphabet $\Sigma = \{A,C,G,T\}$, and be of length m; assume that there is a SNV with two allelic variants called wildtype and alternative allele at position i with $i \in \{1, \ldots, m\}$ in S. Hence, we considered two variants of the sequence S, S^1 containing the wildtype allele and S^2 containing the alternative allele. For each of these sequence variants, we computed a TF binding score describing the TF binding affinity to the sequence according to the model M. From the distribution of all binding scores for a TF under the model M, we can compute the probability $P(a \leq t, M)$, the p-value of observing a binding score a smaller than a certain threshold t. We can use this to determine the corresponding p-value of the TF model M for each variant, given as $p(S^1, M)$ and $p(S^2, M)$, respectively. The exact computation of the TF binding score and the p-value depend on the used TF model (see STAR Methods section 'Computation of TF binding scores with position weight matrices'). To evaluate the effect of a SNV on a TFBS, we computed a log-ratio similar to that of Manke et al.[24] and called this the *differential TF binding score (D)*:

$$D(S^1, S^2, M) = \log\left(\frac{p(S^1, M)}{p(S^2, M)}\right). \quad \text{(Equation 1)}$$

A positive D indicates that the exchange from the wildtype allele to the alternative allele increases the binding affinity of the TF given the TF model M and may lead to a gain of a binding site. However, a negative D decreases the binding affinity, and therefore, the binding site might be lost.

The maximal differential TF binding problem

Usually, the position at wich the binding affinity is most affected by the SNV is not known. As a consequence, we evaluated all sequences overlapping with the SNV; hence, we defined a window of size $2m - 1$ centered around the SNV. We slide the TF model over the $2m - 1$ long sequence and compute D for each of the subsequences of length m overlapping the SNV. To identify the optimal D, we re-trained the maximal absolute value.

Therefore, our definition changes: we are given a DNA sequence $S = \{s_1, \ldots, s_{2m-1}\}$ and a SNV with two allelic variants centered in the middle of the sequence at position m. We considered two variants of the sequence S: S^1 containing the wildtype allele and S^2 containing the alternative allele. Further, a k-mer, which is a subsequence of S of length m, is defined as $k_i = \{s_i, \ldots, s_{i+m-1}\}$ with $i \in \{1, \ldots, m\}$. For each sequence variant, we defined the k-mers overlapping the SNVs. The k-mers of S^1 are denoted as k_i^1, and the k-mers for S^2 are denoted as k_i^2 with $i \in \{1,\ldots,m\}$. We maximized over the absolute D values of all k-mers overlapping the SNV and identified the TF position at which the binding site was most affected. The *absolute maximal TF binding score* $D_{\max}$ was defined as follows:

$$D_{\max}(S^1, S^2, M) = \max_{i \in \{1,\ldots,m\}} \left(\left|D\left(k_i^1, k_i^2, M\right)\right|\right). \quad \text{(Equation 2)}$$

Estimating the distribution of differential TF binding scores

In this section, we investigate whether the distribution of D follows a known distribution. To do so, we describe the p-values of the TF binding scores for the sequences S^1 and S^2 of length m as random variables W and Z, respectively. Since the TF binding score itself is continuous, the p-values are uniformly distributed in [0,1] under the null hypothesis.[40] Further, we assume that the random variables W and Z are independent of each other. Hence, the following theorem can be applied:

Theorem 1. If two independent random variables W, Z, are uniformly distributed in the range [0,1], then $\log\left(\frac{W}{Z}\right)$ is $Laplace(0, 1)$ distributed[41].

In theory, we conclude that D can be represented as a random variable $X = \log\left(\frac{W}{Z}\right)$, which is $Laplace(0, 1)$ ($L(0, 1)$) distributed. However, our experiments showed that the p-values of the binding scores for S^1 and S^2 might not be independent of each other, at least not when PWMs are used as TF model. Nevertheless, D can be adequately approximated by a $L(0, b)$ distribution, with a scale parameter b different from 1 (see Results section The differential TF binding score approximately followed the $L(0, b)$ distribution). The scale parameter is given as $b = \sqrt{\frac{\sigma^2}{2}}$, where σ^2 is the variance in the observed differential TF binding scores of a TF model M for a set of SNVs.

Derivation of the distribution of the maximal differential TF binding scores

When computing $D_{\max}$, we can represent the differential TF binding scores of the k-mers as n independent and identical $L(0, b)$ distributed random variables X_i with $i \in \{1, \ldots, n\}$, where n is the overall number of k-mers. $D_{\max}$ can be described as the absolute maximum over all random variables X_i, so $Y = \max_{i \in \{1,\ldots,n\}}(|X_i|)$. To efficiently compute a p-value for $D_{\max}$, we are interested in identifying the cumulative distribution function (CDF) of Y.

To do so, we split the following into three parts: (1) we determine the probability distribution function (PDF) and CDF of the absolute values of n Laplace(0, b) distributed variables. (2) We derive the CDF of the n maximal $L(0, b)$ distributed random variables, and (3) we combine both parts to obtain the CDF of Y.

(1) **Computation of the PDF and CDF of the absolute values of n $L(0, b)$ distributed random variables.** The general PDF of the $L(0, b)$ distribution with the scale $b > 0$ is defined as

$$f(x) = \frac{1}{2b} \cdot e^{\frac{-|x|}{b}}. \quad \text{(Equation 3)}$$

To obtain the density for the absolute value $|x|$, one needs to add up the densities for the positive and negative values:

$$f_{|x|}(x) = f(x) + f(-x) = \frac{1}{2b} \cdot e^{\frac{-|x|}{b}} + \frac{1}{2b} \cdot e^{\frac{-|-x|}{b}} = \frac{1}{b} \cdot e^{\frac{-|x|}{b}}, \quad \text{(Equation 4)}$$

with $x \in \mathbb{R}^+$. The corresponding CDF is given by integrating $f_{|x|}(x)$ from 0 to x:

$$F_{|x|}(x) = \int_0^x f_{|x|}(x)dy = \int_0^x \frac{1}{b} \cdot e^{\frac{-|y|}{b}} dy = \left[-e^{\frac{-|y|}{b}} \right]_0^x = 1 - e^{\frac{-|x|}{b}} \quad \text{(Equation 5)}$$

(2) **Derivation of the CDF of n maximal $L(0, b)$ distributed random variables** The CDF of a random variable V is defined as $F(x) = P(V \leq x)$. We derive the CDF of the maximal X_i with $i \in \{1, \ldots, n\}$ using the definition of CDFs and the independence of the n $L(0, b)$ distributed random variables:

$$\begin{aligned} F_{\max}(x) &= P(\max(X_i) \leq x) \\ &= P[(X_1 \leq x) \cap (X_2 \leq x) \cap \ldots \cap P(X_n \leq x)] \\ &= \prod_{i=1}^{n} P(X_i \leq x) = \prod_{i=1}^{n} F(x) = F(x)^n \end{aligned} \quad \text{(Equation 6)}$$

(3) **Derivation of the CDF for an absolute maximal $L(0, b)$ distributed random variable** To finally obtain the CDF of $Y = \max_{i \in \{1,\ldots,m\}}(|X_i|)$, we can plug in the CDF for the absolute $L(0, b)$ distributed random variables (Equation 5) in Equation 6, resulting in:

$$F_{\max|x|}(x) = F_{|x|}(x)^n = \left(1 - e^{\frac{-|x|}{b}}\right)^n. \quad \text{(Equation 7)}$$

The corresponding PDF is the derivative of Equation 7:

$$f_{\max|x|}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(F_{\max|x|}(x)\right)^n = \frac{\mathrm{d}}{\mathrm{d}x}\left(1 - e^{\frac{-|x|}{b}}\right)^n = \frac{n}{b} \cdot e^{\frac{-|x|}{b}} \cdot \left(1 - e^{\frac{-|x|}{b}}\right)^{n-1}. \quad \text{(Equation 8)}$$

In summary, we are able to mathematically describe the distribution of $D_{\max}$ as follows:

Theorem 2. *Let $D_{\max}$ be defined as $Y = \max_{i \in \{1,\ldots,n\}}(|X_i|)$, where the X_i's are independent and identical $L(0, b)$ distributed random variables; then, Y follows a modified Laplace distribution $L_{\max}(n, b) = f_{\max|x|}(x)$.*

The distribution of $D_{\max}$ depends on the parameter n and the scale parameter b. Both parameters need to be determined for each TF model separately. Here, n denotes the number of k-mers overlapping the SNV. Since we also consider the reverse complement, n is given by 2 times the length of the TF model.

To determine the scale parameter b for j observed maximal differential TF binding scores x_i with $i \in \{1, \ldots, j\}$ of a TF model M, we set up a maximum log-likelihood estimator (MLE) of the PDF of $L_{\max}$

$$L(x_i|b) = \log\left(\prod_{i=1}^{j} f_{\max|x|}(x_i)\right) = \log(k) - \log(b) - \sum_{i=1}^{j}\left(\frac{|x_i|}{b} + \log\left(1 - e^{\frac{-|x_i|}{b}}\right)^{n-1}\right) \quad \text{(Equation 9)}$$

and computed the corresponding derivative with respect to b. Since the resulting equation is not analytically solvable, we used Newton's method to numerically approximate b (STAR Methods section "Fitting the scale parameter b").

Using the CDF of the $L_{\max}(n, b)$ distributed maximal differential TF binding scores, we are able to compute a p-value for $D_{\max}$ as $1 - F_{\max|x|}(x)$.

Computation of TF binding scores with position weight matrices

Throughout this manuscript, we illustrate our statistical approach using position weight matrices (PWMs) as an example of a TF model. We rely on PWMs, because they are still commonly used and available for hundreds of human TFs. In addition, we wanted to provide a fair comparison to the previously established method *atSNP*, which is based on PWMs. Often, TFs can have varying binding motifs, that cannot be represented by a single PWM. When this is the case, we consider all known PWMs of this TF. A PWM M describing a TF motif of length m, is

an $4 \times m$ matrix, holding for each base the log-likelihood at position i with $i \in \{1,\ldots,m\}$. The TF binding score for a sequence S given a PWM M is computed as $\sum_{i=1}^{|S|} P(\sigma, s_i)$, where $|S| = m$ and σ is the base in S at position i. Using the dynamic programming approach from Beckstette et al.,[42] we computed the exact TF binding score distribution for a PWM M. As a result, we can determine the p-value for every possible TF binding score obtainable by the PWM M. If the p-value of a TF binding score is smaller than a given threshold t, we assume that the TF, represented by the PWM, is able to bind to the sequence.

Derive exact TF binding score distribution

The downloaded TF models from the JASPAR, HOCOMOCO and Kellis ENCODE motif database are originally position count matrices. We convert them to Position Weight Matrices (PWMs), thereby we added an epsilon of 0.001 to every entry to avoid 0 entries. To apply the dynamic programming approach of Beckstette et al.,[42] we shifted the resulting log-likelihoods in such a way that all values are >0 and rounded them with an accuracy of 0.001. We precomputed the exact TF binding score distribution for all given PWMs (for more details see *Section Calculation of exact PSSM score distributions* in the paper of Beckstette et al.). An implementation of their approach can be found on our github repository (https://github.com/SchulzLab/SNEEP, file *src/pvalue_copy.hpp*).

Evaluation of the independence of the TF binding score of the wildtype and alternative allele

To evaluate whether the random variables Z and W, that represent the p-values of the TF binding scores for the wildtype and alternative allele, are statistically independent in the case of PWM TF models, we performed a Chi-Square test. p-values of the TF binding score for all k-mers for all PWMs of our motif set for 100 randomly sampled SNVs were derived. Then, for each PWM it was evaluated whether the p-values of the TF binding score of the wildtype and alternative allele, respectively, were independent of one another.

Fitting the scale parameter b

In all analyses we conduct within this work we used as TF motif set either the collection of non-redundant human PWMs combined from JASPAR (version 2022), Hocomoco and Kellis ENCODE motif database or subsets of it. We removed flanking bases of the TF motifs with an entropy higher than 1.9, since we observed that TF motifs with flanking bases that exhibit a high entropy have a negative effect on the fit of the distribution to the observed D_{max}. To apply our method, we pre-computed the scale parameter b for each TF motif. To do so, we randomly sampled 200.000 SNVs from the dbSNP database (build id 154). For each TF, we computed D_{max} for all SNVs. These values are plugged in the MLE of the PDF of L_{max} and numerically solved using Newton's method to approximate b (done with python library scipy.optimize[43]). We want to describe the tail of the distribution of D_{max} as accurate as possible. Therefore, we minimized the mean squared error (MSE) for the tail of the distribution of D_{max} (25% of all values) by decreasing / increasing the estimated scale parameter b by 0.01 as long as the MSE decreased.

Statistical evaluation of model fit

To compute how well the $L_{max}(n, b)$ distribution approximates the empirical distributions, for each TF motif a Kolmogorov-Smirnov test was applied. We randomly sampled 250 SNVs from the dbSNP database, and computed D_{max} for all TF motifs. With the Kolmogorov-Smirnov test, we evaluated whether the distribution of the observed D_{max} values is identical to our $L_{max}(n, b)$ distribution. For a more robust result, the procedure was repeated 100 times and the average p-value was derived.

Collecting allele-specific binding events

We collected 1,760 Allele Specific Binding (ASB) events from Shi et al.[44] identified in the human cell line GM12878 using 14 TF ChiP-seq datasets. For heterozygous binding sites of a TF, an ASB event was defined if the number of mapped ChIP-seq reads for one allele was significantly greater than that for the other allele. The authors noted that only 19.3% of the ASB events overlapped with a predicted TFBS of the TF for which the ChIP-seq experiment was designed for. Hence, we wanted to consider only the ASB events overlapping with a TFBS of the used TF motif. Therefore, we gathered the 14 TF motifs from the JASPAR database (version 2022)[16] and computed the TFBS per TF using Fimo (version meme-5.2.0, default p-value cutoff).[45] Since redundant motifs would lead to false positives later in our analysis, we clustered a combined TF motif set of the JASPAR, Hocomoco[46] and Kellis ENCODE motif database[47] using the similarity measurement and clustering approach from Pape et al.[48] We checked which of the 14 TFs belonged to the same cluster and retrained only the TF with the highest number of SNVs per TF motif cluster. In doing so, we removed two TFs, resulting in 368 SNVs for 12 TFs (see Figure S1 and Table S1, used data is given in our ZENODO repository (https://doi.org/10.5281/zenodo.7588272)).

Collecting SNP-SELEX data

The SNP-SELEX data were downloaded from the web portal GVAT database.[49] We gathered the SNVs called *original batch*. In total, Jian et al. studied 1,612,172 TF-SNV pairs for 271 different TFs. For each TF, we collected all SNVs that had biological evidenceof a differential binding event. Therefore, we filtered the SNVs for those that had an oligonucleotide binding score p-value <0.05 and a preferential binding score p-value <0.01, as proposed by the authors, resulting in 9,840 SNVs for 129 TFs. We obtained the TF motifs from Boytsov et al.,[50] which provides optimized PWM motifs for the SNP-SELEX dataset. As for the ASB dataset, we excluded for each TF the SNVs without a TFBS for at least one of the two alleles. If less than 5% of the SNVs associated with a TF had a TFBS, we excluded all the TF-SNVs pairs from the analysis. Next,

we checked which of the TFs belonged to the same TF motif cluster based on the clustering used for the ASB SNVs. For each TF motif cluster, we selected the TF with the most SNVs, resulting in a total of 33 TFs and 1,494 SNVs (see Figure S1 and Table S2, used data is given in our ZENODO repository (https://doi.org/10.5281/zenodo.7588272)).

Method performance evaluation

We applied our method and *atSNP* to the SNV-TF pairs collected for the ASB events (see STAR Methods section 'Collecting allele-specific binding events') and SNP-SELEX data (see STAR Methods section 'Collecting SNP-SELEX data'). In STAR Methods Section details about the used commands to run *atSNP* and our approach, the commands used are listed and the required input data is given in our ZENODO repository (https://doi.org/10.5281/zenodo.7588272). To evaluate the performance of each method, we computed a precision-recall curve and the area under the precision recall curve (AUCPR) (see Figures 3A and 3B) using the R package *PRROC*.[51] To test whether the area under the receiver operating characteristic (ROC) curve was significantly different between the two approaches, we applied the method of DeLong et al.[52] using the R package pROC.[53]

We compared the runtimes of our approach and those of *atSNP* for 100, 500, 1,000, 10,000, 20,000 and 40,000 SNVs randomly sampled from the dbSNP database (build id 154).[54] As both methods provide a parallel mode, we compared the runtimes for 1, 8 and 16 threads (see Figure 3C).

eQTL analysis

We collected the eQTLs for *EBV-transformed lymphocytes* and *cultured fibroblasts* from the GTEx portal[55] (version 8; dbGaP Accession phs000424.v8.p2) on December 21, 2022. The fine-mapped eQTLs were extracted from the file *GTEx_v8_finemapping_CaVEMaN.txt.gz*,[56] resulting in 14,722 eQTLs for lymphocytes and 45,917 eQTLs for fibroblasts. Further, we downloaded the gene transcripts per million (TPM) values of lymphocytes and fibroblasts from the GTEx portal (*gene_tpm_2017-06-05_v8_cells_cultured_fibroblasts.gct.gz* and *gene_tpm_2017-06-05_v8_cells_ebv- transformed_lymphocytes.gct.gz*). We computed the $D_{\max}$ p-value for each eQTL separately for each dataset. As a motif set, we used a combined TF motif set from the JASPAR (version 2022), Hocomoco and Kellis ENCODE motif database (given in our ZENODO repository (https://doi.org/10.5281/zenodo.7588272)). Based on the dbSNP database (build id 154), we randomly sampled 1,000 SNV sets, that contained the same number of unique SNVs as the corresponding eQTL dataset. Additionally, for each randomly sampled SNV set, we computed the $D_{\max}$ p-value. Next, we counted how often the binding sites of a TF were significantly affected across all the eQTLs ($D_{\max}$ p-value ≤ 0.01). We denote this count as *TFcount*. To identify cell type-specific TFs, it is necessary to normalize the *TFcount* by those observed for randomly sampled data, since every motif has a different probability of occurring by chance depending on the properties of the TF motif itself. Similarly, for each randomly sampled SNV set, we counted how often the binding site of a TF was significantly affected. We took the mean over all randomly sampled SNV sets, denoted by *bgCount*. To identify TFs that are more often affected by the eQTLs of the current cell type than expected, we computed an odds-ratio for each TF, which is defined as $\text{odds} - \text{ratio(TF)} = \frac{\alpha\ /\ (1-\alpha)}{\beta\ /\ (1-\beta)}$, where $\alpha = \frac{TFcount}{\#SNVs}$, and $\#SNVs$ are the numbers of unique SNVs per cell type and $\beta = \frac{bgCount}{\#SNVs}$, respectively.

Since we cannot distinguish TFs with binding motifs of high similarity, we wanted to evaluate the results at the level of TF motif clusters (see STAR Methods section 'Collecting allele-specific binding events'). If a TF within a cluster has an odds-ratio ≥ 2, we assume that it is enriched, as it is two times more often affected by the SNVs of the cell type of interest than is expected by chance. For each cluster, we considered the maximal TPM value over all TFs in the cluster as gene expression level. To determine whether the difference in expression between the TFs with enriched odds-ratio and the TFs without an enriched odds-ratio was significant, we applied a one-sided Wilcoxon rank sum test using the *wilcox.test* functionality in R.

Application atherosclerosis GWAS

The atherosclerosis GWAS was downloaded from the GWAS catalog (EFO_0003914), including the child traits *carotid atherosclerosis* (EFO_0003914) and *peripheral arterial disease* (EFO_0009783) on 02/04/2021. We excluded all indels, duplicated SNVs and all SNVs from GWASs not based on a European cohort. For the remaining 255 out of 261 SNVs, we determined the SNVs in linkage disequilibrium (LD) by applying SNiPA[57] using as a population the European cohort and an LD threshold of 0.8. We gathered 4,326 unique proxy SNVs (given in our ZENODO data repository (https://doi.org/10.5281/zenodo.7588272)). We computed the $D_{\max}$ p-value for each collected SNV associated with atherosclerosis. As input motifs, we used 817 non-redundant human motifs from the JASPAR (version 2022), HOCOMOCO and Kellis ENCODE motif database. If the $D_{\max}$ p-value was ≤ 0.01, we assumed that the binding site of a TF was significantly affected by the considered SNV. To link the regulatory SNVs to target genes, we used 2.4 million regulatory elements (REMs) associated to target genes downloaded from the EpiRegio database.[33,34] We overlapped the identified rSNVs with these REMs using bedtools' *intersect* functionality.[58]

Details about the used commands to run *atSNP* and our approach

For a better reproducibility of our results, we provide all executed commands and the input files given in our ZENODO data repository (https://doi.org/10.5281/zenodo.7588272). For all analyses we used as genome version hg38, beside for the SNP-SELEX data where we used hg19. The dataset specific input files are indicated with < >. The SNV and motifs data for the different data set for our approach and *atSNP* are provided in their required format, the content is the same.

Used commands to apply our approach

In the following the used commands to run our approach are listed. All script and more details how to run them can be found in our github repository: https://github.com/SchulzLab/SNEEP.

In a first step, we estimate the scale parameter *b* for the used PWMs once using

```
bash estimateScalePerMotif.sh 200000 <motifs> <outputDir> <motifNames> 1.9
```

As *motifs* we used the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt*, *motifNames* lists the names of all TF motif for which we wish to compute *b*. The result is a file providing for each considered TF model the estimated scale parameter *b*, called *estimatedScalesPerMotif_1.9.txt* in the following and available at our github repository in the directory *necessaryInputFiles/*.

To compute the results for the ASB events, the SNP-SELEX data and the runtime analyses we used the following command:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <outputDir> -n <numberCores>
-p 1.0 -c 1.0 -j 0 <motifFile> <input-snps> <path-to-genome-file>
estimatedScalesPerMotif_1.9.txt.
```

The following combinations of input files for *motifFile* and *input-snvs* where used:

- ASB events: snps_ASB.txt, motifs_ASB.txt
- SNP-SELEX data: snps_SNP_SELEX.txt, motifs_SNP_SELEX.txt
- the randomly sampled SNPs can be found in the files sampledSNPs100.txt, sampledSNPs500.txt, sampledSNPs1000.txt, sampledSNPs10000.txt, sampledSNPs20000.txt, sampledSNPs40000.txt, motifs: *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt*

For the eQTL analyses we performed a background sampling which can be done automatically using:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <outputDir> -n 16 -p 0.5
-c 0.01 -j 1000 -l 10 -k dbSNPs_sorted.txt
combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar2022.txt
<input-snps> <path-to-genome-file> estimatedScalesPerMotif_1.9.txt
```

The *input-snvs* we downloaded from the GTEx Portal (see STAR Methods Section eQTL analysis), the file *dbSNPs_sorted.txt* is a sorted an filtered version of the dbSNP database and part of our Github repository and the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt* can be found in our ZENODO data repository.

To link rSNVs to regulatory elements as we did it for the atherosclerosis GWAS, we executed the command:

```
time ./src/differentialBindingAffinity_multipleSNPs -o <outputDir> -n 10 -p 0.5
-c 0.01 -r REM.txt -g ensemblID_GeneName.txt -j 0
combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar2022.txt
<input-snps.txt> <path-to-genome-file> estimatedScalesPerMotif_1.9.txt
```

The files *interactionsREMs.txt* and *ensemblID_GeneName.txt* can be found in our github repository and in our ZENODO data repository the SNVs of the GWAS atheroclersis are stored in the file snps_atherosclerosis.txt and the motifs are provided in the file *combined_Jaspar_Hocomoco_Kellis_human_transfac_jaspar_2022.txt*.

Commands used to run atSNP

To run *atSNP* (version 1.14.0) the following commands where executed in R:

```
pwms <- LoadMotifLibrary(<motifFile>, tag = ''MOTIF'', skiprows = 2, skipcols = 0, transpose = FALSE, field = 2,
sep = ''_'', pseudocount = 0)
snps <- LoadSNPData(filename = <snpFile>, genome.lib = <genomeVersion>)
scores <- ComputeMotifScore(pwms, snps, ncores = <numberCores>)
diffBind <- ComputePValues(motif.lib = pwms, snp.info = snps, motif.scores = scores$motif.scores, ncores =
<numberCores>)
```

The used motif input files and the SNV files are the following:

- ASB events: snps_ASB_atSNP.txt, motifs_ASB_atSNP.txt
- SNP-SELEX data: snps_SNP_SELEX_atSNP.txt, motifs_SNP_SELEX_atSNP.txt
- randomly sampled data: snps: sampledSNPs100_atSNP.txt, sampledSNPs500_atSNP.txt, sampledSNPs1000_atSNP.txt, sampledSNPs10000_atSNP.txt, sampledSNPs20000_atSNP.txt, sampledSNPs40000_atSNP.txt, motifs: motifs_JASPAR_HOCOMOCO_Kellis_1.9.meme

As *genomeVersion* we used "BSgenome.Hsapiens.UCSC.hg38" for the ASB events and "BSgenome.Hsapiens.UCSC.hg19" used for the SNP-SELEX data.