Reviewed Preprint
v1 • July 1, 2026
Not revised

For correspondence:
marc.robinson-rechavi@unil.ch

Competing interests: No competing interests declared

Funding: See page 21

Reviewing editor: Diethard Tautz, Max Planck Institute for Evolutionary Biology, Germany

RegEvol: detection of directional selection in regulatory sequences through phenotypic predictions and phenotype-to-fitness functions

Alexandre Laverré[1,2], Thibault Latrille[3], Marc Robinson-Rechavi[1,2]

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland • [2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland • [3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

eLife Assessment

The focus of this manuscript is a computational procedure to reveal signatures of selection on transcription factor binding sites through assessing changes in predicted binding affinity, setting out to avoid biases inherent in previous tests. The general approach could become a **valuable** resource for the community that can also be used for a broader range of questions. However, in its current implementation, the methods are **inadequate** to sufficiently support the primary claims.

https://doi.org/10.7554/eLife.111237.1.sa3

Abstract

Regulatory DNA controls when and where genes are expressed, making it a key driver of phenotypic evolution. Yet detecting selection in non-coding regions remains difficult, as most approaches rely on sequence conservation or changes in substitution rate rather than molecular effects. RegEvol bridges this gap by linking machine learning-based predictions of transcription factor binding to explicit evolutionary models. It uses the distribution of predicted mutational effects to infer fitness functions under different evolutionary scenarios including random drift, stabilising selection, and directional selection. Through maximum-likelihood estimation, it identifies the regime that best explains observed changes along a lineage from an ancestral sequence. When substitution numbers are limited, such as along short evolutionary branches, likelihood differences can be aggregated across sets of regulatory elements to increase statistical power. RegEvol corrects biases that affected previous tests based on machine learning of transcription factor binding, while remaining conservative across different levels of divergence. Applied to over 3 million Drosophila melanogaster regulatory regions, we identify 5.1% under directional selection, enriched near reproductive and immune genes. Applying the aggregation strategy to human CTCF binding across tissues reveals enrichment of directional signals in nervous and male reproductive systems. The framework is readily applicable to experimentally detected regulatory elements with alignable ancestral sequences and is flexible to future advances in understanding regulatory function, providing a powerful basis for investigating adaptation in non-coding regions.

Introduction

The evolution of *cis*-regulatory sequences plays a central role in shaping phenotypic diversity and adaptation across species. Regulatory evolution has been proposed as a significant factor in morphological and physiological change since King and Wilson's well-known suggestion that many phenotypic divergences between closely related species may be caused by regulatory changes

rather than differences in protein coding (King and Wilson 1975). This view is now supported by abundant evidence, with the majority of variants associated with phenotypic traits and human diseases falling within non-coding regulatory elements, such as enhancers and promoters (Maurano et al. 2012; F. Zhang and Lupski 2015; P. H. Lee et al. 2018). These sequences attract transcription factors (TFs) that bind to DNA by recognising specific sequence motifs, enabling precise gene expression patterns. This motif-based binding provides a direct link between genotype and phenotype, where the studied phenotype is a quantitative molecular trait. Moreover, mutations affecting cis-regulatory sequences that alter TF binding have been shown to contribute to macroscopic phenotypic diversity within and between species (Wittkopp and Kalay 2012; Albert and Kruglyak 2015), serving as the basis for sequence-based models of regulatory function (Lai et al. 2019; Sokolova et al. 2024).

Despite their functional importance, studying the evolutionary forces on regulatory sequences remains a major challenge. Unlike protein-coding sequences, where the ratio of nonsynonymous to synonymous substitutions (dN/dS) provides a framework to detect selection (Tanaka and Nei 1989; Yang 1998; Yang 2014), regulatory sequences lack a comparable metric. This is partly due to the absence of a universal "regulatory code" that would allow for a systematic interpretation of mutations' functional effects in non-coding DNA (Kim and Wysocka 2023). As a result, it remains difficult to distinguish between neutral drift, stabilising selection, and adaptive evolution in regulatory regions. The majority of commonly applied approaches to detect selection in non-coding DNA are based on substitution rate variation among phylogenies. Tools such as PhastCons (Siepel et al. 2005) and PhyloP (Pollard et al. 2010) identify conserved and accelerated regions by comparing observed substitution rates to neutral expectations. These approaches have been widely used to identify deeply conserved elements that are likely to be functionally important, and can allow for branch-specific rate shifts. However, they remain limited in several significant aspects. First, substitution rate-based methods are inherently indirect: they infer selection from sequence conservation or acceleration patterns, without modelling the functional consequences of mutations. As a result, they can be confounded by non-adaptive processes such as biased gene conversion, variation in mutation rates, or demographic history (Galtier and Duret 2007). Second, because they are based on conservation across lineages, they can be poorly suited for detecting selection in fast-evolving or recently gained regulatory elements. Indeed, most cis-regulatory elements are not conserved at large evolutionary distances (Villar et al. 2015), and can maintain conserved regulatory activity despite substantial sequence divergence (Wong et al. 2020; Phan et al. 2025). Finally, they often rely on predefined neutral proxies, such as 4-fold degenerate sites or repetitive elements. These may not accurately reflect local mutational or selective backgrounds, especially in non-model species with limited annotations. Recent studies have attempted to enhance these models by incorporating phylogenetic structure and functional annotations. For example, a probabilistic framework has been developed to model the turnover of cis-regulatory elements using ChIP-seq data and phylogenetic hidden Markov models (Dukler et al. 2020). Similarly, Yan et al. introduced PhyloAcc-GT, a Bayesian method that detects substitution rate shifts in conserved noncoding elements while accounting for gene tree discordance (H. Yan et al. 2023). These approaches represent important advances in modeling substitution rate variation, but they still rely on the underlying premise that conservation of function is coupled to conservation of sequence.

To overcome these limitations, recent efforts have focused on building genotype-to-phenotype maps that directly link regulatory sequence variation to functional output at multiple molecular levels, including chromatin accessibility, transcription factor binding affinity, and downstream gene expression. Massively parallel reporter assays have enabled the systematic measurement of the effects of thousands of mutations on enhancer activity, providing empirical fitness landscapes for regulatory elements (Patwardhan et al. 2012; Smith et al. 2013; Gallego Romero and Lea 2023). These data have been used to train machine learning models that predict TF binding, chromatin accessibility, and gene expression from DNA sequence alone (Vaishnav et al. 2022; Krieger et al. 2022). These approaches demonstrate that the functional impact of mutations in regulatory DNA can be inferred directly from sequence, providing a powerful alternative to conservation-based methods. Building on this conceptual foundation, Liu and Robinson-Rechavi

(2020) previously developed a machine learning-based method to detect positive selection on regulatory sequences by comparing observed changes in predicted TF binding affinity to a null distribution of random mutations. This approach leveraged gapped k-mer support vector machines (gkm-SVMs) trained on ChIP-seq data to predict the impact of substitutions on TF binding. It can infer whether a regulatory sequence has evolved under positive selection by comparing the predicted change in binding affinity (ASVM) to a null distribution generated by in silico mutagenesis. This method offers clear advantages over traditional rate-based approaches, as it avoids reliance on predefined neutral sites, can be applied to individual regulatory elements, and models the functional impact of mutations directly from sequence. Yet despite these strengths, several important limitations have emerged. One is ascertainment bias in ChIP-seq data, where enrichment for high-affinity binding sites can inflate false-positive rates in selection tests (Jiang and Zhang 2024). Another is that the method's accuracy declines with divergence, since its non-parametric null model becomes unreliable over larger evolutionary distances.

Here we present RegEvol, a new framework building on our previous work by combining machine learning-based predictions of mutational effects with explicit population-genetic models of selection. Rather than relying on substitution rate shifts or permutation-based tests, we directly model the fitness effects of mutations in regulatory sequences. Using a trained sequence-to-function model, our approach predicts the functional impact of all possible point mutations in ChIP-seq-defined regulatory regions. This generates genotype-to-phenotype maps that describe how sequence variation affects transcription factor binding. These predictions are then incorporated into a population-genetic framework that uses phenotype-to-fitness functions to determine the fitness effect of phenotypic changes. We evaluate the performance of this method through simulations and compare it with our previous one. Finally, we apply it to diverse empirical datasets in *Drosophila melanogaster* to study the evolutionary dynamics of regulatory sequences.

Results

1 SVM predictions capture functional signals beyond sequence conservation

We trained gapped k-mer support vector machine (gkm-SVM) models to distinguish transcription factor (TF)-bound regions identified by ChIP-seq from random genomic sequences matched for length, GC content, and repeat composition (D. Lee 2016). These models learn to recognize specific combinations of short DNA motifs (k-mers) enriched in TF-bound regions, thereby capturing the sequence context and features that define functional regulatory elements. Despite the rise of deep learning frameworks such as DeepSEA (Zhou and Troyanskaya 2015) and BPNet (Avsec et al. 2021), gkm-SVM remains among the most interpretable and TF-specific approaches, with performance comparable to deep models when trained on individual TF datasets (Tognon et al. 2025; Vorontsov et al. 2025). Moreover, gkm-SVM directly quantifies sequence-level contributions, enabling efficient in silico mutagenesis at base-pair resolution. Once trained, the model assigns an SVM score to any DNA sequence, reflecting its predicted similarity to the ChIP-seq peak class based on k-mer composition.

We first confirmed the biological relevance of these predictions by analysing how SVM scores vary along individual ChIP-seq peaks. Using a sliding window approach, we computed position-wise SVM scores across *Homo sapiens* peaks bound by CEBPA. A representative example is shown in Figure 1A, which reveals two distinct regions of elevated scores, each approximately 10 bp in length, corresponding to predicted TF binding sites. These high-scoring regions matched the canonical CEBPA motif from the JASPAR database (Castro-Mondragon et al. 2022), suggesting that the model accurately identifies functional binding sites within regulatory elements. To generalise this observation, we extracted the top 100 10-mers with the highest SVM weights from the CEBPA model and compared them to known motifs using the TomTom tool from the MEME

Suite (Gupta et al. 2007; T. L. Bailey et al. 2009). The top-scoring matches corresponded to the CEBPA motif and related forkhead TFs (Supplementary Table 1), confirming that the model captures biologically meaningful sequence features associated with TF binding.

We then quantified the predicted impact of point mutations using ASVM, defined as the change in SVM score after a single-nucleotide substitution. At each position in the sequence, we introduced all possible point mutations in silico and computed the change in predicted binding affinity. The absolute ASVM values reflect the magnitude of predicted functional disruption caused by each mutation. To assess how well these metrics reflect TF binding intensity, we computed normalised read coverage for each base within CEBPA ChIP-seq peaks, using it as a proxy for TF abundance. All peaks were centred on their summits (the position of maximum read coverage) and aligned to produce an average signal profile. We then calculated SVM scores, ASVM values, and conservation metrics (phastCons and phyloP scores) at each aligned position. Both SVM and ASVM scores exhibit a sharp peak at the summit, closely mirroring the ChIP-seq signal (Figure 1B). This indicates that mutations in these regions are predicted to most strongly affect TF binding, either directly or indirectly. In contrast, phastCons and phyloP scores lack such enrichment and have a lower dynamic range, suggesting that the conservation scores are less sensitive to fine-scale variation of TF occupancy and thus regulatory region function. Indeed, both SVM and ASVM scores are positively correlated with read coverage over individual peaks (median Pearson's correlation coefficient = 0.31 and 0.22 respectively; Fig 1C), while phastCons and phyloP scores are largely uncorrelated (median Pearson's correlation coefficient = 1.8e-2 and 2.7e-3). This confirms that SVM-based metrics capture TF binding intensity more accurately than conservation-based scores. These patterns were consistent across multiple TFs and species (Supplementary Figure 1). Importantly, the strength of the base-level SVM-ChIP correlation increases with peak quality, and exclusion of the lowest-quality peaks substantially improves these correlations (Supplementary Figure 2), indicating that modest or negative correlations primarily reflect technical variability rather than limitations of the modeling.

Finally, we examined the consistency of model predictions across species and TFs using mammalian liver ChIP-seq datasets (Supplementary Figure 3-4). Clustering based on k-mer scores revealed that peaks grouped primarily by TF identity rather than species, indicating that the models capture conserved binding preferences. An exception was observed for HNF4 and FOXA1, which clustered by species rather than individual TF. These two transcription factors are known to co-occupy regulatory regions in the genome, acting cooperatively and sequentially to activate liver-specific gene expression (Horisawa et al. 2020), and thus there is little sequence signal differentiating their individual binding from each other.

2 Defining genotype-to-fitness maps for TF-ChIP-seq peaks

To leverage this functional prediction by SVMs and investigate the evolutionary forces acting on ChIP-seq peaks, we developed RegEvol. This framework quantifies selection on regulatory sequences by integrating predictive models of binding affinity with evolutionary theory. This approach is illustrated in Figure 2 and detailed mathematical formalism is described in Supplementary Materials.

The first step involves constructing a genotype-to-phenotype map for each ChIP-seq peak, describing how sequence variation affects transcription factor binding affinity (i.e., molecular phenotype). This is achieved by performing in silico mutagenesis on all possible point mutations within the peak and computing the change in predicted binding affinity for each mutation using ASVM. The resulting ASVM values define a Distribution of Phenotypic Effects (DPE), which captures the range of potential binding changes accessible from the ancestral sequence in one point mutation (Figure 2.A, left). Each DPE is specific to a given individual peak and provides a quantitative representation of its local genotype-to-phenotype landscape. Substitutions that occurred along the focal lineage are identified by comparing the reference sequence to its inferred ancestral state, using whole-genome alignments (Figure 2.A, right). The corresponding ASVM is

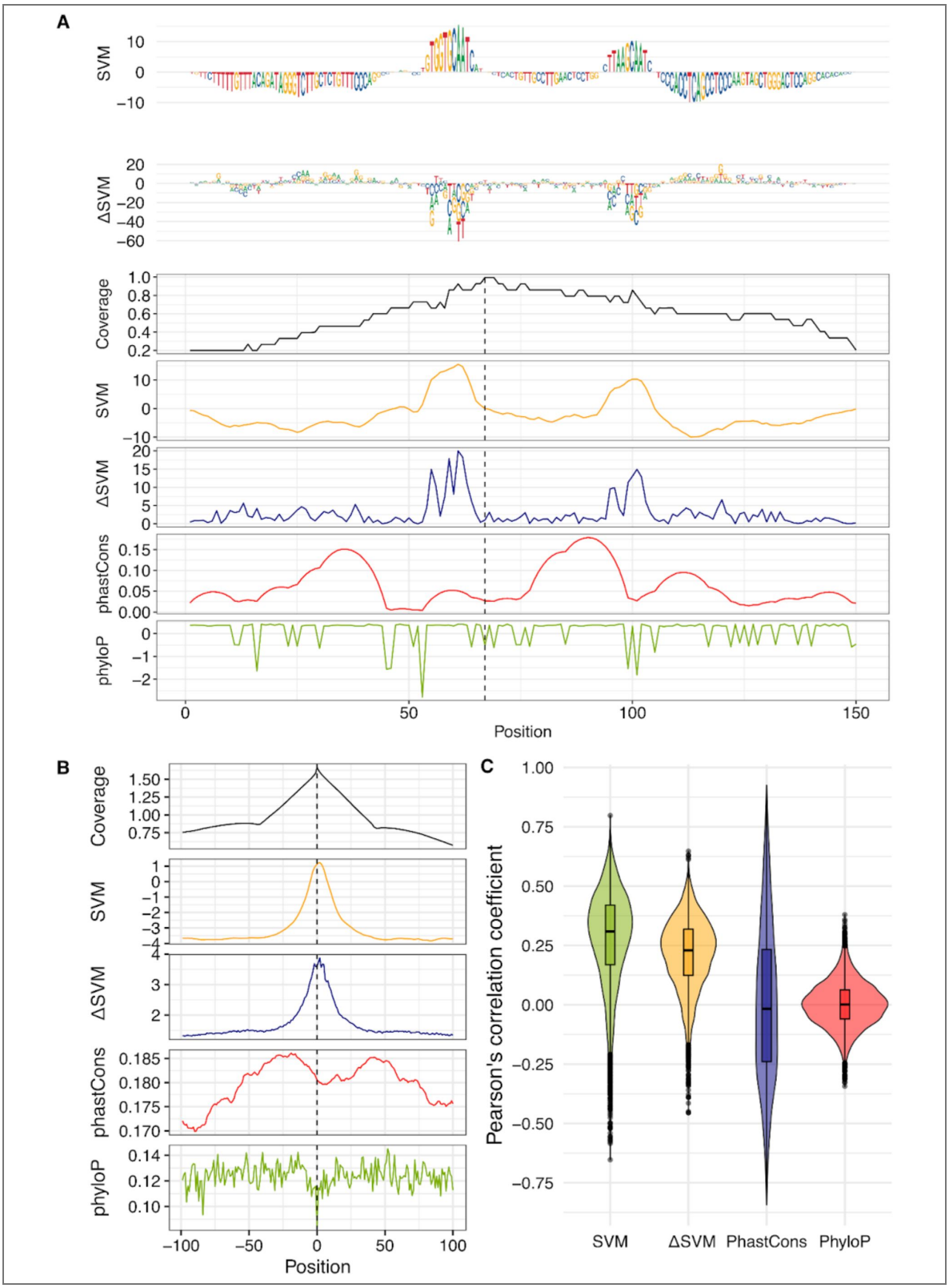

Figure 1. SVM-based predictions capture functional signals of transcription factor binding.

A) SVM scores computed across a CEBPA ChIP-seq peak in *Homo sapiens*. Each letter represents a nucleotide, with font size proportional to the SVM score. Two regions of elevated scores match the canonical CEBPA binding motifs (JASPAR). The summit of the peak is shown by a dotted line. **B)** Average profiles of normalised read coverage (red), SVM prediction scores (yellow), absolute ASVM values (green), phastCons score (blue), and phyloP score (purple), centred on the summits of CEBPA ChIP-seq peaks. **C)** Distribution of Pearson correlation coefficients between each metric (SVM, ASVM, phastCons, phyloP) and the base-level read coverage profile across individual CEBPA peaks.

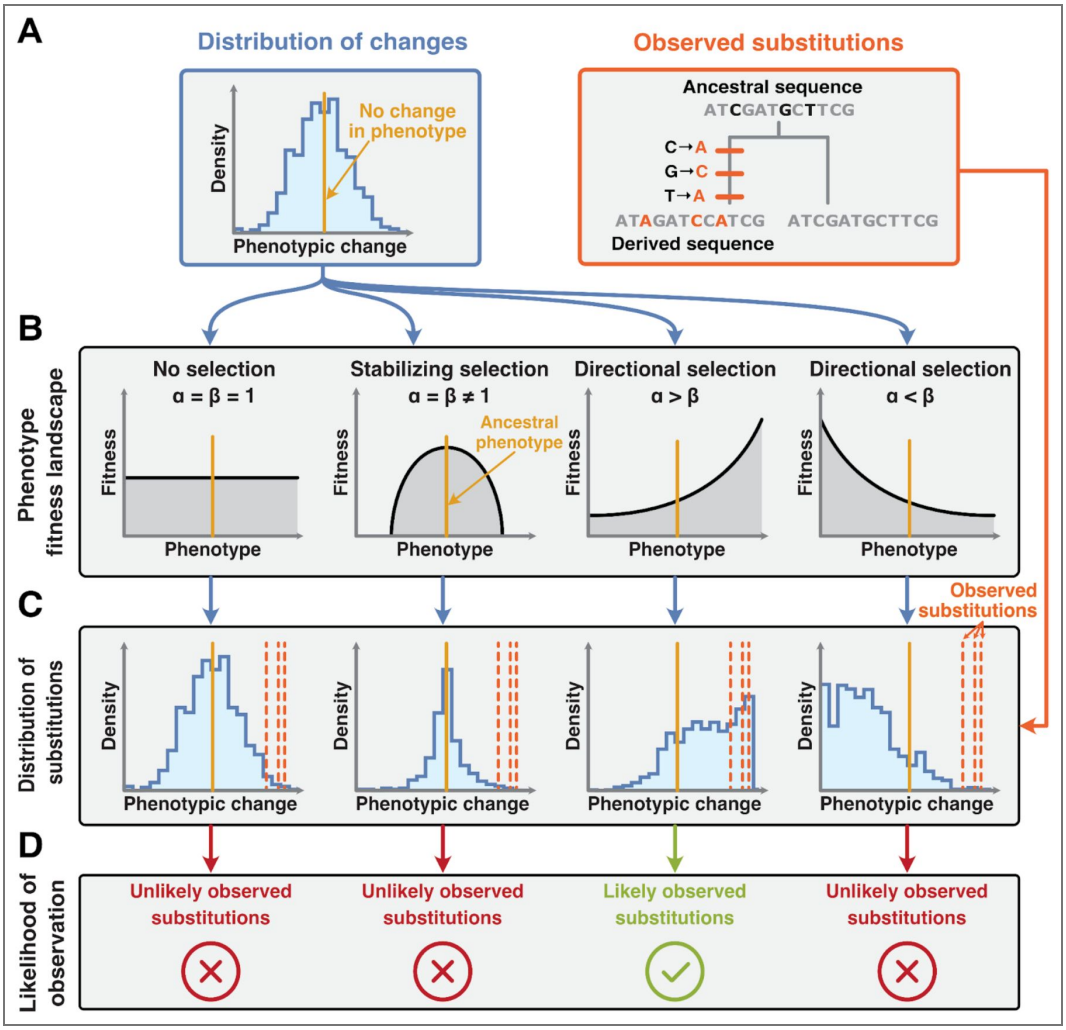

Figure 2. Outline of RegEvol: from genotype to fitness.

A) Distribution of phenotypic effects (DPE) for all possible point mutations within a ChIP-seq peak (left). The observed substitutions relative to the ancestral sequence represent a subset of this distribution and are used to infer the peak's evolutionary regime (right). **B)** Definition of nested selective regimes using Beta distributions parameterized by a and 0, which define the shape of the underlying fitness landscape. **C)** Expected distributions of all possible substitutions under each phenotypic fitness landscape. Vertical orange lines indicate the observed substitutions. **D)** Likelihood of the observed substitutions under each selective regime. In this example, because the substitutions are biased toward positive phenotypic change, the model of positive directional selection is the most likely.

computed for each observed substitution, representing the phenotypic effect of the mutations that were actually fixed in the lineage. These observed substitutions can be viewed as a subsample of the whole mutational landscape defined by the DPE.

Three nested evolutionary scenarios are defined to model how selection may have shaped these substitutions, each corresponding to a different phenotype-to-fitness map (Figure 2.B). These maps transform the predicted phenotypic effects into fitness consequences, effectively modelling a distribution of fitness effects (DFE) specific to each regulatory element. Rather than assuming a fixed DFE, we derive it from a mechanistic model of molecular function, consistent with the population genetic framework proposed by Eyre-Walker and Keightley (2007). These models are parameterised using Beta distributions.

The neutral model assumes that there is no selection on binding affinity, with mutations having a probability of fixation that does not depend on ASVM. This is captured by a flat fitness landscape, corresponding to a uniform Beta distribution with a = 0 = 1. Importantly, variation in mutation and substitution rates across the genome is accounted for in all evolutionary scenarios (Materials & Methods). The stabilising selection model assumes that the ancestral binding affinity is optimal, disfavoring mutations that deviate from this optimum. This is modelled by a symmetric Beta distribution centred at ASVM = 0, with a = 0 # 1. The shape of the distribution reflects the strength of selection, with higher values of a and 0 indicating stronger selection. The directional selection model allows for asymmetric fitness landscapes, favouring mutations that increase or decrease binding affinity. This is modelled by an asymmetric Beta distribution with a # 0, allowing the fitness peak to shift away from the ancestral state. Each fitness function is combined with the DPE to generate an expected distribution of fixed mutations under the corresponding selection regime. The observed substitutions are then compared to these expectations to assess which model best explains the data.

To infer the most likely evolutionary scenario for each peak, model parameters are estimated by maximising the likelihood of the observed substitutions, given the DPE (Figure 2.C). The selection coefficient for each mutation is computed as the log-ratio of fitness between the derived and ancestral states. Fixation probabilities are calculated accordingly, and the likelihood of the observed substitutions is computed under each model. Likelihood ratio tests are then used to compare models and determine the best-fitting scenario for each peak. The model with the highest likelihood, adjusted for complexity, is selected as the most probable evolutionary explanation: neutral evolution, stabilising selection, or directional selection (Figure 2.D).

3 Detection of Directional selection on simulated peaks evolution

To evaluate the performance of our likelihood-based method, we simulated the evolution of ChIP-seq peaks under three evolutionary scenarios: random sampling (i.e., drift), skewed towards mutations with low effect (i.e., stabilising selection), and skewed towards a directional change (i.e., directional selection). For each scenario, we generated synthetic sequences by sampling substitutions according to their mutation probability, predicted phenotypic effect (ASVM), and fixation probability under a specified selection model. Selection strength was controlled by adjusting the parameters of the Beta distribution used to define the phenotype-to-fitness map.

RegEvol produced well-calibrated p-values and controlled false discovery rates across scenario comparisons, demonstrating that likelihood-based inference accurately reflects statistical confidence (Supplementary Figure 5). Across all scenarios, the method showed high accuracy in identifying the correct evolutionary model (Figure 3.A). The method showed strong power to detect directional selection, with true positive rates (TPR) of 0.90 for Positive Directional Selection and 0.77 for Negative Directional Selection under moderate selection strength (**a**=10) Notably, the test was conservative: false positives (i.e cases where directional selection was inferred under neutrality or stabilising selection) remained at 1e-4 at a FDR threshold of 1%.

We quantified the factors influencing detection power by simulating the evolution of sequences under directional selection from *D. melanogaster* CTCF peaks (Figure 3.B-E), and contrasted RegEvol to our previous permutation test (Liu and Robinson-Rechavi 2020). RegEvol

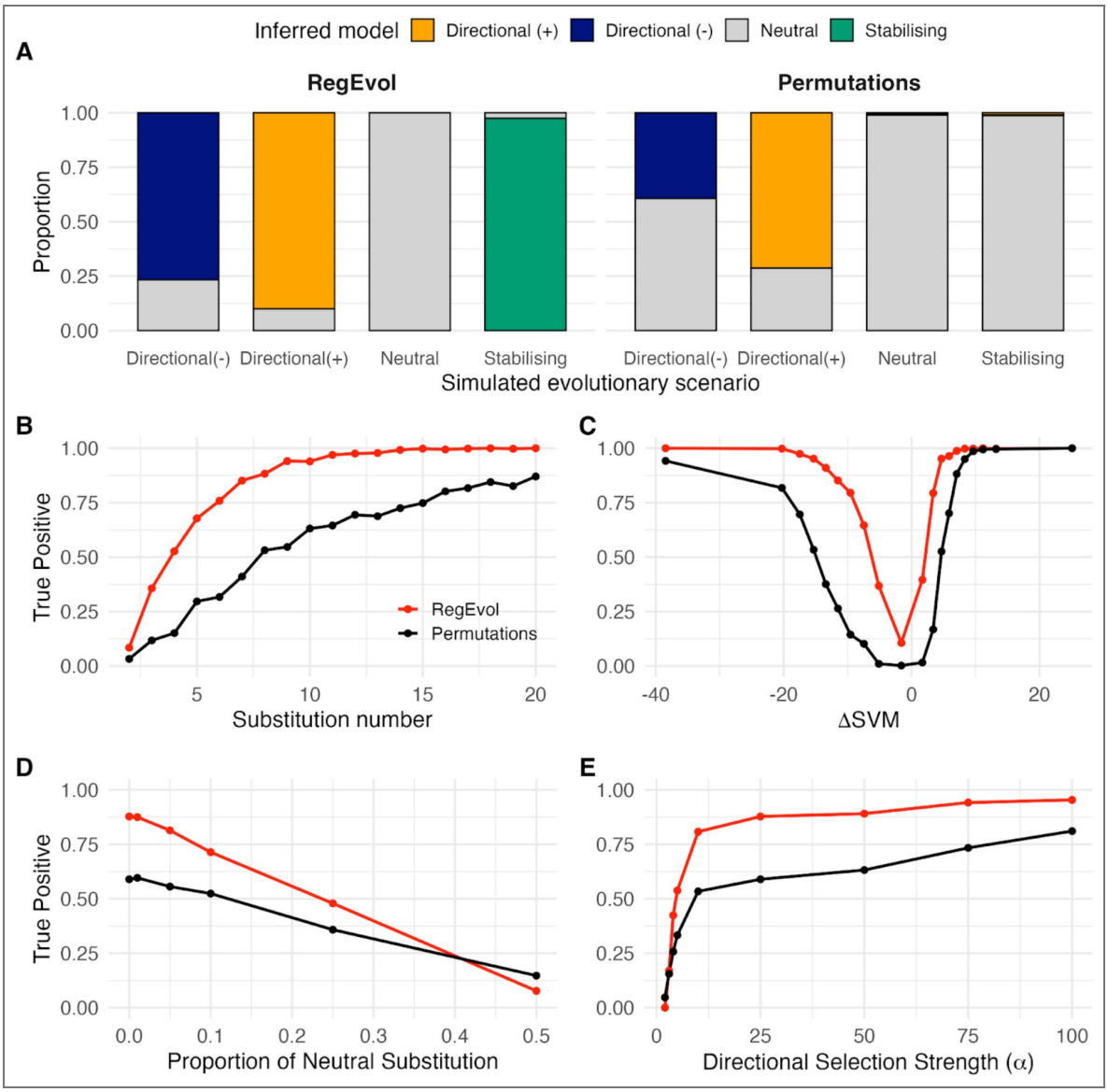

Figure 3. Detection of directional selection on simulated evolution of drosophila CTCF peaks.

A) Proportion of peaks simulated under Directional, Stabilising and Random scenarios and detected by RegEvol and the Permutations Test as evolving under positive directional selection (orange), negative directional selection (blue), stabilising selection (green) and random drift (grey). **B-E)** True Positive rate of peaks simulated under directional selection (toward positive (+) or negative (-) changes) detected by RegEvol (red) or the Permutation Test (black), as a function of **B)** number of substitutions per peak, **C)** ASVM (quantile 5%), **d)** proportion of random substitutions, and **E)** selection strength (**a** parameter). *N* = 10,000 simulated peaks for each evolutionary scenario; FDR < 0.01.

outperformed the permutation test across all ranges of conditions, with both higher sensitivity and specificity. As expected, detection rates increased sharply with substitution counts (Figure 3.B). Peaks with fewer than five substitutions rarely reached significance in the permutation test, showing a TPR below 0.3, and remained below 0.75 for RegEvol. The TPR for RegEvol plateaued near its maximum after about ten substitutions, whereas the permutation test increased more gradually and never reached full detection, even at 20 substitutions per peak. With only two substitutions per peak, RegEvol's TPR remained low even under strong directional selection (TPR=0.24 at a=100) but reached its maximum plateau at four substitutions (Supplementary Figure 6). This reflects the conservative nature of our likelihood framework, which requires consistent phenotypic shifts across multiple substitutions to confidently reject the null model.

We observed an asymmetry in detection power for directional selection: peaks evolving toward decreased binding affinity (i.e., negative ASVM) were less frequently detected (Figure 3.A and C). This reflects the typical asymmetry of the DPE across species and transcription factors, with a higher proportion of mutations showing negative effects (Supplementary Figure 7; mean ASVM in DPE for Drosophila CTCF peaks=-0.33, sd=1.44). Moreover, as expected, detection rates were low for simulated peaks with the smallest ASVM (*i.e.*, those causing little to no change in predicted binding affinity). However, while the permutation test performs poorly in this regime, RegEvol remains capable of detecting consistent directional shifts even when individual substitutions have modest effects. This is because the likelihood framework integrates directional consistency across multiple substitutions, rather than relying solely on the sum of their effects. This distinction highlights the strength of RegEvol in identifying subtle but systematic evolutionary trends.

To test the robustness of our method, we introduced a proportion of randomly fixed substitutions (i.e., mutations sampled without regard to their ASVM) to simulate evolutionary noise. This reflects realistic scenarios where drift is strong or multiple factors beyond the focal TF impact the evolution of regulatory elements. As expected, detection power declined with increasing proportions of such substitutions, dropping below 0.5 when one-quarter were random (Figure 3.D). RegEvol outperformed the permutation test until half of the substitutions were random, where the two methods have TPR below 0.25. A similar pattern was observed when varying directional selection strength: both methods improved with stronger selection and plateaued at comparable levels, with RegEvol maintaining superior performance (Figure 3.E).

4 Specificity and Ascertainment Bias

We also evaluated the sensitivity of both methods to extreme-effect substitutions, sequence divergence, and potential ascertainment bias. The permutation test, by design, is highly responsive to large ASVM values. While this can be advantageous for detecting isolated, high-impact substitutions, it also makes the test more prone to false positives when such mutations are fixed by chance. In simulations where substitutions were drawn independently of their ASVM (Figure 4.A), the permutation test showed an overall low false positive rate (FPR=0.01) but a sharp increase for extreme ASVM values, particularly positive ones (quantile 1% FPR=0.15; quantile 99% FPR=0.39). A similar pattern was observed in simulations under stabilising selection, where substitutions are constrained to low ASVM values (Figure 4.B, quantile 99% FPR=0.26). In contrast, RegEvol remained stable and conservative across all simulations, as it integrates information across all substitutions and emphasizes consistent directional trends rather than outliers (maximum FPR across quantile=1e-4).

Distinguishing stabilising selection from random drift is inherently difficult, as most substitutions have low predicted effects on TF binding affinity (Supplementary Figure 7). We therefore assessed RegEvol's performance in detecting stabilising selection across a range of simulation parameters (Supplementary Figure 8). Moreover, because of the asymmetry of the DPE toward negative changes, the distribution of ASVM in random permutations becomes increasingly negative with the number of substitutions. Consequently, under stabilising selection, the false discovery rate of the permutation test positively correlates with sequence divergence and can

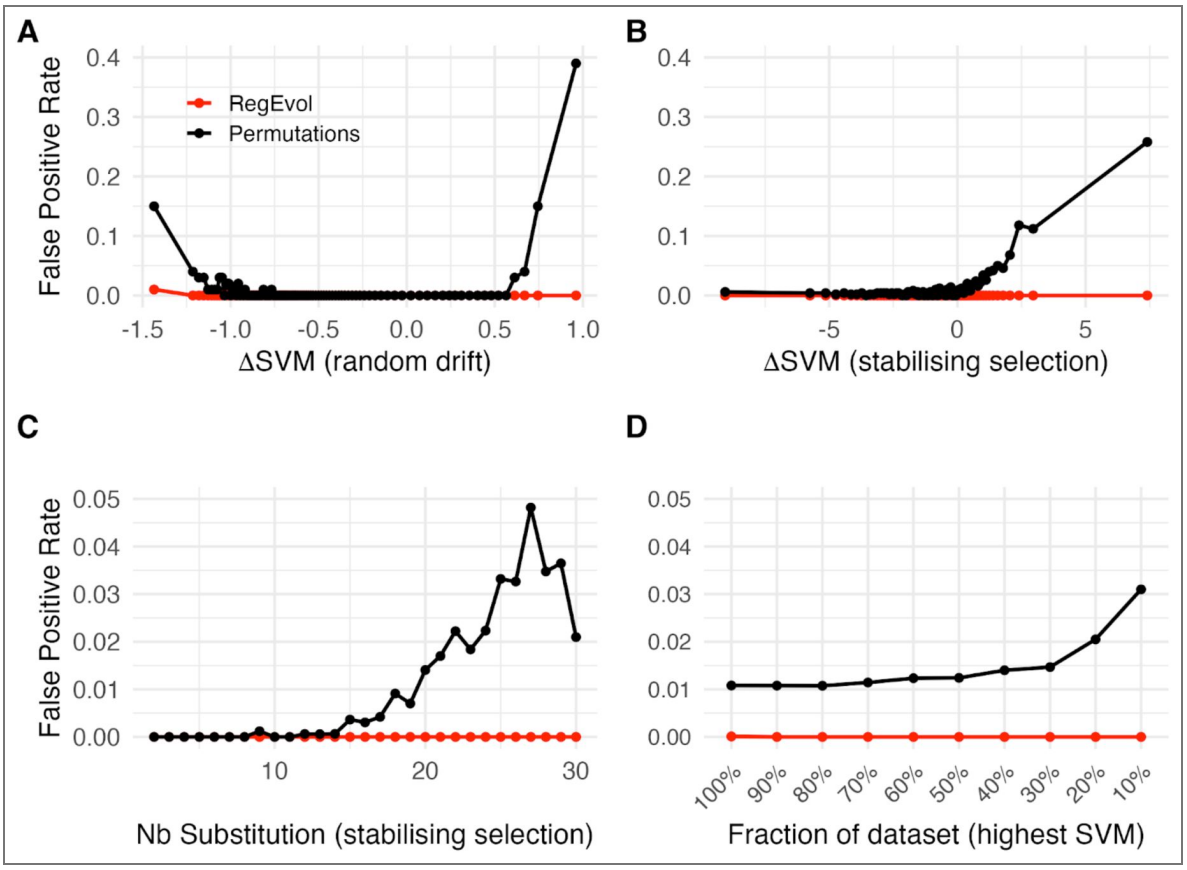

Figure 4. Sensitivity to extreme substitutions and robustness to ascertainment bias.

Proportion of simulated *Drosophila* CTCF peaks detected under directional selection using the Permutation Test (black lines) and RegEvol (red lines). Substitutions were either drawn randomly (**A, D**) or under stabilising selection (**B, C**; mean(a=P)=45). **A-B)** False positive rate across 1% quantiles of ASVM; **C)** false positive rate as a function of the number of substitutions; **D)** fraction of datasets stratified by the highest derived SVM scores. *N* = 10,000 simulated peaks; FDR < 0.01.

reach a FPR of 0.05 with high number of substitutions (Figure 4.C, Spearman's rho=0.96, p=7.4e-16). By integrating the ASVM distribution as a DFE for each peak and explicitly modeling stabilising selection around ASVM = 0, RegEvol effectively controls for this bias (FPR=0).

An ascertainment bias on ChIP-seq peaks has been suggested to affect permutation tests, as higher SVM score peaks are both more likely to be called and to show high positive ASVM (Jiang and Zhang 2024). To test this, we subsampled peaks by derived ASVM in random simulations and assessed the proportion detected under directional selection. RegEvol maintained a FPR close to 0 across peak categories, while the permutation test showed a positive association with derived ASVM reaching FPR above 0.03 (Figure 4.D). Applying the same subsampling to the previously analysed human CEBPA dataset (Liu and Robinson-Rechavi 2020), the correlation for RegEvol remained low, with only the highest SVM category showing a slight increase, potentially reflecting a true biological signal (Supplementary Figure 9). These results demonstrate that RegEvol is robust to peak strength, supporting its reliability in real datasets where ChIP-seq signal intensity may influence peak calls.

To evaluate practical limitations of RegEvol, we examined how detection power is affected by the number of substitutions per regulatory element and multiple testing correction (Supplementary Figures 10-11). In empirical datasets, many peaks do not have enough substitutions, particularly in closely related species. For example, in mammalian datasets, the median number of substitutions per peak can be low depending on the divergence since the closest ancestral state (human median substitution number = 2.8), which reduces statistical power and leads to a scarcity of low p-values after multiple testing correction (Supplementary Figure 11). In Drosophila, higher divergence results in more substitutions per peak, providing a more stable detection power across regulatory elements (Supplementary Figure 10). Given these observations and the low true positive rate for peaks with few substitutions even under strong simulated directional selection (Supplementary Figure 6), we restricted our analyses to peaks with at least four substitutions in vertebrates datasets (Supplementary Figure 12). This corresponds to an average of 1.7% sequence divergence (median length=291bp; Supplementary Figure 13-14).

5 Analysis of *Drosophila melanogaster* ChIP-seq peaks

To illustrate the utility of RegEvol, we applied it to a large-scale dataset of transcription factor binding sites in *Drosophila melanogaster*. Using ChIP-seq peaks from the modERN consortium (Kudron et al. 2024), we analysed regulatory elements across multiple developmental stages and transcription factors (Supplementary Data Table 2). For each peak, we computed the DPE using gkm-SVM models trained on the corresponding TF and inferred the most likely selection regime since the divergence from *D. simulans*, based on ancestral sequences reconstructed from whole-genome alignments.

We inferred a total proportion of 5.1% of fruit fly peaks under directional selection (Figure 5A), out of 2.8M tested. This is not due to high permissiveness of the test, since on mammalian ChIP-seq peaks we find much lower proportions of directional selection (Supplementary Figure 12). This is consistent with reports of a higher proportion of amino acid substitutions fixed by directional selection, and generally higher efficiency of selection, in the fruit fly than in mammals (Eyre-Walker 2006; Sella et al. 2009; Lin et al. 2025). Among the fly peaks, there is considerable variation between TFs and conditions, from almost none to 50% of peaks inferred under directional selection. With the low developmental resolution of those data, there is no strong trend of differences over ontogeny but there is less signal of directional selection in prepupal and pupal stages (Supplementary Figure 15).

As expected, the signal for directional selection increases with the ASVM (Figure 5B): peaks with larger predicted phenotypic effects are more often inferred to have evolved under directional selection. We observe much more positive than negative directional selection, a pattern likely biological rather than methodological, since it is absent from simulations. We detect few peaks under stabilising selection, which can be explained by the framing of our test relative to the nature of purifying selection: stabilising selection tends to eliminate mutations, leading to few

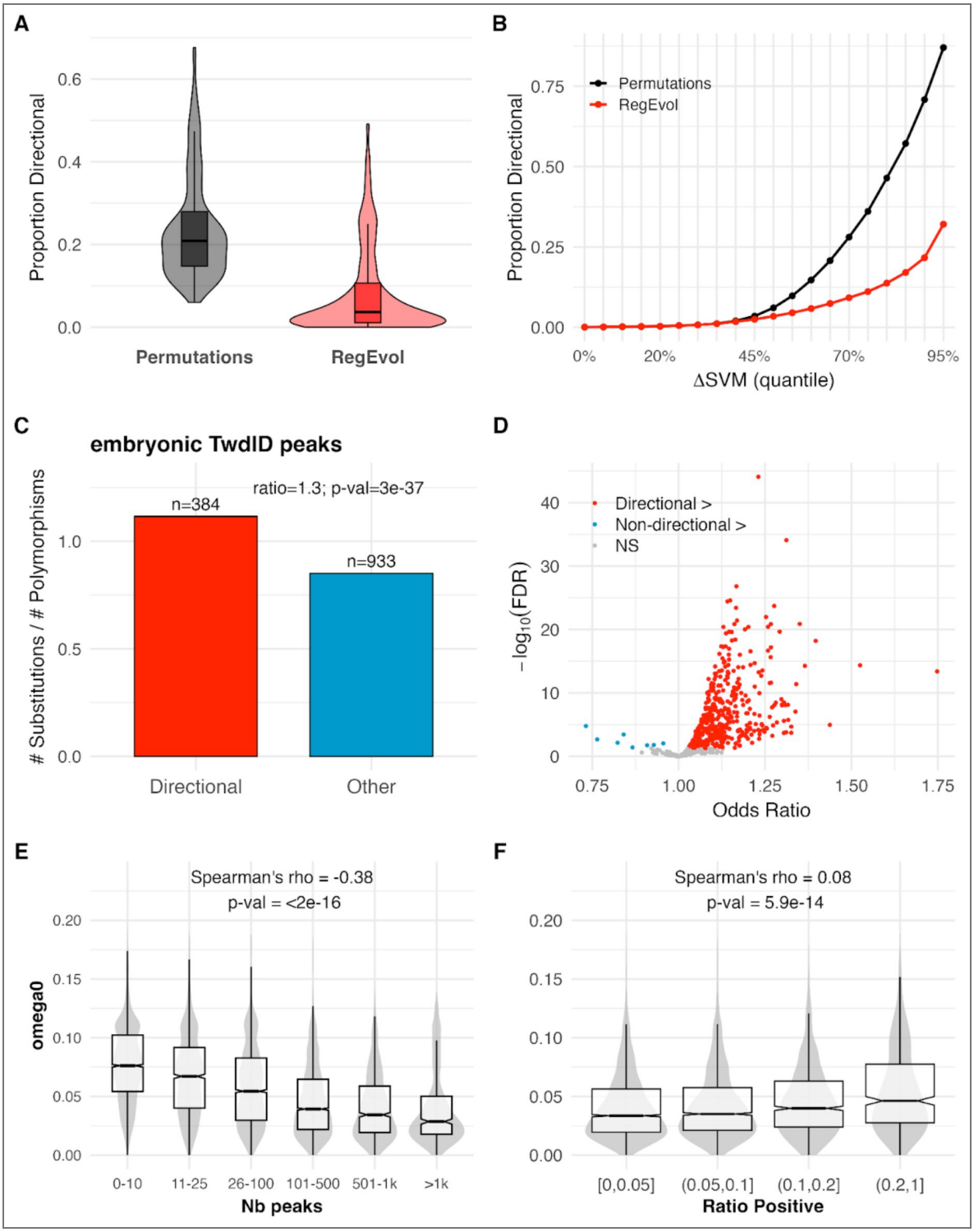

Figure 5. Application of RegEvol to Drosophila melanogaster transcription factor binding sites.

A. Proportion of peaks inferred to be under directional selection using the original permutation test (black; $p < 0.01$) and RegEvol (red; FDR < 0.05). **B.** Proportion of peaks under directional selection across the 5% quantiles of derived ΔSVM values for all peaks. **C.** Ratio of substitutions to single-nucleotide polymorphisms (SNPs) in embryonic *TwdID* peaks inferred by RegEvol under directional selection (red; $N = 385$) or not (blue; $N = 877$). The odds ratio and Fisher's exact test between the two categories are indicated. **D.** Odds ratios between directional and non-directional peaks across datasets (restricted to experiments with >10 directional peaks). Experiments where directional peaks show higher ratios (red; odds ratio > 1), lower ratios (blue; odds ratio < 1), or no significant difference (grey; FDR > 0.05) are shown. **E-F.** Relationship between purifying selection on protein-coding genes (wo) and the total number of associated peaks (**E**) or the proportion of their peaks inferred under directional selection (**F**). Spearman's correlation coefficients were computed across all genes.

substitutions and a low power of the likelihood test. The role of stabilising selection in our model is essentially to avoid false positives in the test of directional selection; however, our approach is poorly adapted to inferring the proportion of peaks evolving under stabilising selection *per se*.

A key prediction of recent directional selection is a local reduction in polymorphism due to selective sweeps, together with increased inter-species divergence from fixation of mutations (McDonald and Kreitman 1991; Fu and Akey 2013). Accordingly, we expect a higher substitution-to-SNP ratio in peaks inferred to have evolved under directional selection. This expectation is confirmed in *D. melanogaster*, with almost all experiments supporting this pattern (74.8% experiment, odds ratio > 1 and Fisher's exact test FDR < 0.05) (Figure 5 C, D). Because our model gains power with increasing numbers of substitutions, this pattern could in principle reflect a bias. To evaluate this, we compared substitutions and SNPs separately and found that peaks under directional selection have both significantly more substitutions and fewer SNPs (Wilcoxon test p-values < 5e10"^{4}; Supplementary Figure 16), consistent with recent selective sweeps rather than a power artefact. This provides independent validation, as SNP frequencies are not used by the model, and shows that our k-mer-based DFE inference aligns with classical expectations of natural selection.

While it is difficult to reliably associate TF ChIP-seq peaks with the genes they regulate, the closest genes provide a good approximation, especially in compact genomes such as that of *D. melanogaster* (Kudron et al. 2024). As expected, protein-coding genes under stronger purifying selection (wo) are associated with more TF peaks (Figure 5E), thus presumably more complex regulation and more pleiotropy. There is also an association, although weaker, between selection on protein-coding genes and on regulatory sequences (Spearman's rho=0.08, p-values=8e-16, Figure 5F): proteins under weaker purifying selection are associated with a higher proportion of peaks under directional selection. We do not find an association between adaptive selection on protein-coding genes (w2) and the proportion of directional selection on their associated peaks.

Genes associated with peaks under directional selection are enriched for morphogenesis and developmental processes (Supplementary Figure 17). However, genes linked to a larger total number of peaks are also overrepresented in these functional categories. When considering the ratio of peaks under directional selection, no significant functional enrichment was observed (Supplementary Figure 17). The lack of functional enrichment may reflect a limitation of this approach. By aggregating TF peaks from 740 different experiments and contexts, the resulting gene set may be too large and functionally heterogeneous for conventional functional tests. To explore whether signals of directional selection could be detected at a broader scale, we applied TopAnat (Komljenovic et al. 2018a; Bastian et al. 2021), which identifies anatomical structures where a gene set is preferentially expressed. Genes with a high proportion of selected TF peaks (>0.2) were enriched in reproductive tissues (testis, seminal fluid glands, ovary) and immune-associated organs (fat body, Malpighian tubule) (Supplementary Figure 18).

6 Tissue-level aggregation reveals directional selection in human CTCF

Applying RegEvol to mammalian ChIP-seq peaks individually yielded a low proportion of elements classified under directional selection (Supplementary Figure 12). This pattern is inherent to lineage-specific evolutionary analyses conducted over short evolutionary distances, where the limited number of substitutions per regulatory element constrains statistical power. To address this limitation, we performed a tissue-level aggregation analysis inspired by the SUMSTAT framework of Daub et al. (2017). For each tissue, we computed cumulative SUM scores from the differences between the likelihood of the Directional model and the Neutral model across active peaks (Material and Methods). Analyses were controlled for sample size through resampling to ensure comparability across tissues. We applied this strategy to human CTCF ChIP-seq peaks detected across multiple tissues (ENCODE Project Consortium 2012).

SUM scores displayed a broad plateau across most tissues, followed by a clear upward shift culminating in a distinct cluster of high-ranking tissues (Figure 6). The highest-ranked tissues were predominantly central nervous system-associated cell types together with male reproductive tissues, whereas peripheral nervous tissue and the majority of other somatic tissues remained within the lower plateau. Formally grouping tissues by biological system revealed that both the Nervous system and the Male reproductive system exhibited higher cumulative SUM scores compared to all other systems (Wilcoxon rank-sum tests between systems significant, $p < 2.2e-16$; Figure 6). Effect size estimates confirmed a strong positive shift, i.e. excess of positive selection signal, for the Nervous system (Cliff's 5 = 0.55) and a smaller positive shift for the Male reproductive system (Cliff's 5 = 0.23). No other systems had positive shifts (all Cliff 5 and Wilcoxon p-values in Sup. Table XX), including immune-associated tissues.

Discussion

Understanding how regulatory sequences evolve is essential for linking genotype to phenotype, yet most evolutionary analyses of non-coding DNA still rely on substitution rates or sequence conservation (Adam Siepel et al. 2005; Pollard et al. 2010). These approaches are powerful for identifying constraints but often lack sensitivity to detect selection acting on specific molecular functions (Ward and Kellis 2012; Villar et al. 2015), especially directional selection. RegEvol addresses this limitation by testing how predicted mutational effects on transcription factor (TF) binding have been shaped by selection. By linking quantitative predictions of regulatory activity to patterns of sequence divergence, RegEvol provides a functionally informed and mechanistically interpretable view of regulatory evolution. In contrast to motif-centric models, the predictive component of RegEvol is trained on entire ChIP-seq peaks, capturing local sequence context and potential co-factors that influence TF binding (J. Yan et al. 2021). This shift from rate-based to phenotype-based inference enables the detection of subtle selection signals that may not produce strong shifts in sequence conservation but shape regulatory activity.

RegEvol builds on a previous permutation-based framework (Liu and Robinson-Rechavi 2020), which evaluated selection by comparing the summed ASVM of observed substitutions to a null distribution derived from random permutations. While this approach provided a useful baseline, it was sensitive to extreme substitution effects, lacked robust multiple-test correction, and was susceptible to ascertainment bias (Jiang and Zhang 2024). RegEvol addresses these limitations by modelling three nested selective regimes (neutral, stabilising, and directional), through fitness functions that link predicted TF-binding effects to substitutions patterns. Using a maximum-likelihood framework, our method identifies the selective regime that best explains the observed substitutions. Moreover, because inference is analytical rather than simulation-based, p-values remain well calibrated and support rigorous multiple-test correction, overcoming a key weakness of the permutation approach (Jiang and Zhang 2024). A key strength of RegEvol is its nested modeling of selective regimes, allowing direct likelihood-based comparison of directional and stabilising selection. This is important because the expectation under stabilising selection is centred on no phenotypic change (ASVM = 0), whereas the permutation-based null is typically slightly negatively shifted. This asymmetry has little impact when few substitutions are available but leads to a marked inflation of false positives with increasing divergence (Figure 4). Including a stabilising regime corrects this bias and improves robustness by providing an intermediate model between neutrality and directional selection.

RegEvol remains robust across a wide range of divergence levels. Simulations show that it is conservative at short timescales, minimizing false positives, and gains power while maintaining specificity as substitutions accumulate. Even when the overall signal is modest, RegEvol detects directional selection when substitution effects are consistent. By modelling the full distribution of predicted effects rather than focusing on extreme values, the method emphasizes consistency over magnitude, reducing false positives driven by a few large-effect mutations and limiting ascertainment biases. This yields conservative inference, a desirable property when testing for positive selection (Jordan and Goldman 2012; Venkat et al. 2018; Wisotsky et al. 2020; Soni et al. 2023). A small number of strong-effect substitutions may therefore be insufficient to

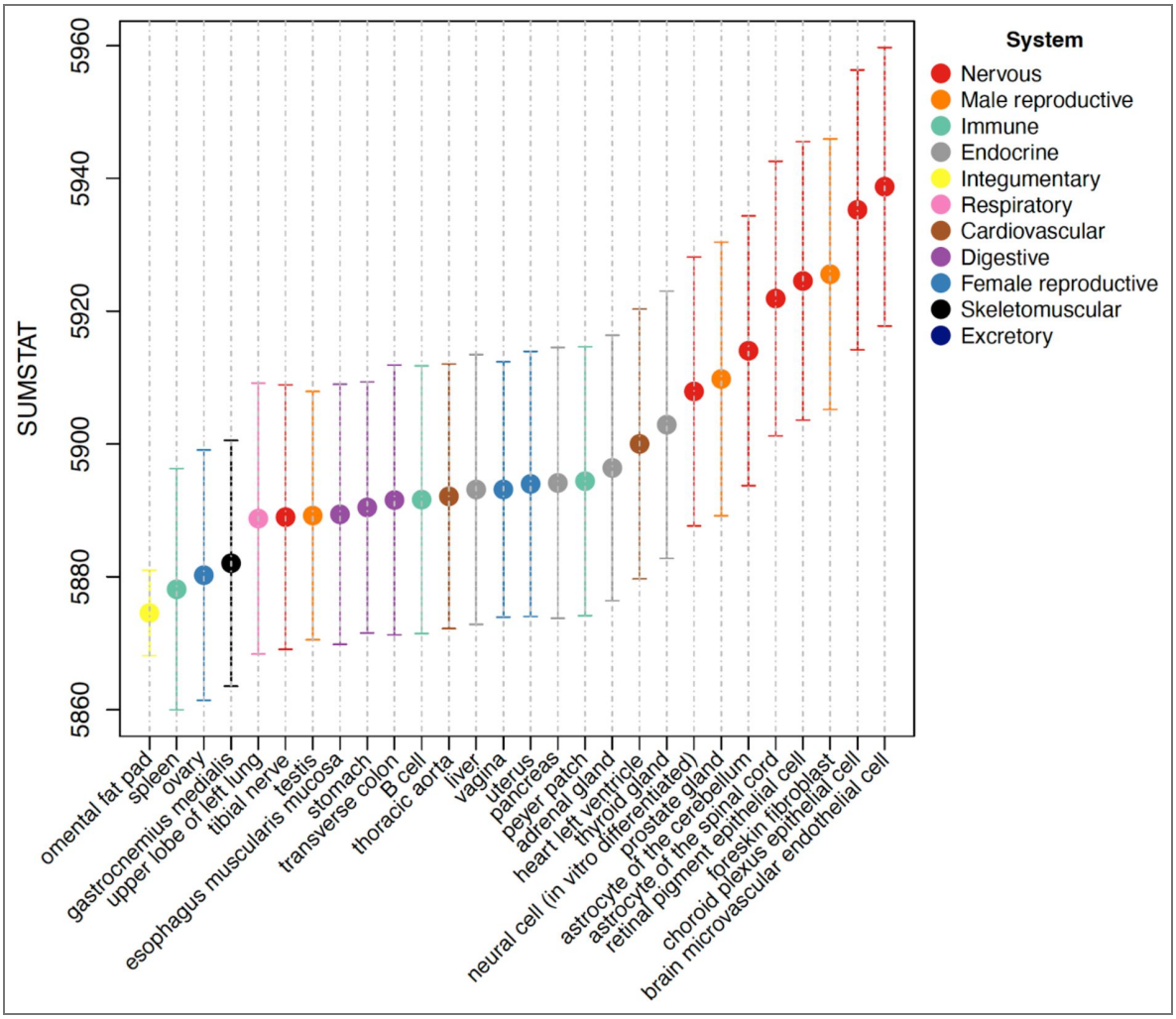

Figure 6. Tissue-level aggregation of directional selection on human CTCF peaks.

Median SUMSTAT scores across 10,000 resamplings of 5,000 human CTCF peaks per tissue. Bars indicate the standard deviation across resamplings. SUMSTAT scores were calculated following the framework from Daub et al. (2017) by taking the fourth-root of the per-peak Alog-likelihood (Directional minus Neutral model) and summing across all peaks within each tissue. Tissues are sorted by median SUMSTAT and colored according to organ system.

support directional selection, reflecting a trade-off between robustness and sensitivity. In contrast, a permutation-based approach can interpret extreme mutations as evidence of selection, potentially increasing power in rare-event scenarios but at the cost of elevated false-positive rates.

By decoupling substitution rate from phenotypic effect, RegEvol provides a function informed view of regulatory evolution, analogous to how coding sequence evolution is interpreted through the nature of amino acid changes rather than their frequency (Halpern and Bruno 1998; Rodrigue et al. 2021). While coding sequences are constrained by protein structure and pleiotropy, regulatory elements evolve under different pressures, including redundancy, turnover, and modularity, which enhance evolvability but complicate evolutionary analysis (Wittkopp and Kalay 2012; Kim and Wysocka 2023). Application to *Drosophila melanogaster* ChIP-seq data illustrates this framework empirically: 5.1% of peaks were inferred under directional selection, consistent with previous reports of adaptive evolution in fly transcription factor binding sites (He et al. 2011) and with the higher efficiency of selection in flies compared to mammals (Eyre-Walker 2006; Sella et al. 2009; Lin et al. 2025). Peaks under directional selection showed elevated substitution-to-SNP ratios, consistent with recent selective sweeps (McDonald and Kreitman 1991; Fu and Akey 2013). These peaks were also associated with genes enriched in reproductive and immune tissues, systems that are major targets of sexual selection and host-pathogen interactions and are known to evolve rapidly (Buchon et al. 2014; Vaibhvi et al. 2022), supporting the biological relevance of the inference. This framework thus provides a link between regulatory sequence evolution and downstream phenotypic effects, offering new perspectives on the co-evolution of regulatory elements and associated genes, and more broadly connecting regulatory variation to organismal adaptation.

Reduced power under low substitution numbers is an inherent limitation of lineage-specific tests. In some regions, this scarcity of substitutions may result from strong purifying selection, in which case conservation-based methods complement rather than compete with functionally informed approaches such as RegEvol. In the mammalian dataset, the limited number of substitutions per regulatory element constrained per-peak inference, motivating the use of a tissue-level aggregation strategy. We applied this approach to human CTCF binding sites, summing per-peak likelihood differences between directional and neutral models across biologically coherent tissue groups, following principles similar to SUMSTAT analyses of coding sequences (Daub et al. 2017). Aggregating likelihoods preserves the contribution of individual peaks and avoids composite sequences, revealing coherent enrichment in Nervous and Male reproductive tissues, consistent with previous evidence of accelerated and adaptive evolution in these systems (Dorus et al. 2004; Khaitovich et al. 2006; Liu and Robinson-Rechavi 2020). While we focused on tissue-level aggregation here, the same framework could be extended to other biologically meaningful sets, such as pathways, cell types, or regulatory modules, allowing RegEvol to uncover directional trends that are otherwise undetectable at the level of individual elements.

Improving the genotype-to-phenotype map remains essential to extend RegEvol's reach. While gkm-SVM models remain interpretable and TF-specific, capturing the additive effects of individual substitutions, they have limited ability to model complex regulatory features such as motif syntax, long-range dependencies, or non-additive interactions (de Boer and Taipale 2024). Recent deep learning models trained on high-resolution functional genomics data have shown the ability to learn these features directly from sequence, providing more accurate, context-aware predictions of mutational impact (Avsec et al. 2021; Benegas et al. 2023; Karollus et al. 2024). Because RegEvol is agnostic to the underlying predictive model, it can incorporate these and future advances, producing more informative phenotypic effect distributions and improving sensitivity in complex regulatory landscapes. Moreover, regulatory elements often integrate inputs from multiple TFs, chromatin states, and cellular contexts, and the functional impact of a mutation may vary across these layers (Hu and Tee 2017; Kurafeiski et al. 2018). Probabilistic frameworks, such as Bayesian or mixed-effect models, could capture heterogeneous and overlapping selective pressures within a single element, capturing this complexity more realistically than a single fitness landscape per element.

Together, these developments position RegEvol as a flexible and extensible framework for studying the evolution of regulatory sequences in a functionally informed manner. By integrating predictive models of molecular phenotypes, which link genotype to quantitative regulatory effects, with well-established evolutionary equations that map phenotypes to fitness, RegEvol provides a principled approach to detect directional selection, complementing traditional conservation-based analyses. As functional genomics, predictive modeling, and comparative data continue to advance, including the ability to transfer predictive models across species, RegEvol could support more systematic, multi-lineage analyses. This would help bridge the gap between regulatory sequence variation and phenotypic evolution and offer new insights into the molecular mechanisms underlying adaptation across evolutionary timescales.

Materials and Methods

Catalogue of Transcription Factor Binding Sites

We used pre-processed ChIP-seq peak data for *Drosophila melanogaster* TFBS from the modENCODE (model organism Encyclopedia of DNA Elements) and modERN (model organism Encyclopedia of Regulatory Networks) consortia (Kudron et al. 2024). It provides a harmonised dataset across 740 experiments covering 645 TFs and five developmental stages (Supplementary Data Table 2).

ChiP-seq peaks detection and coverage quantification

We collected publicly available ChIP-seq data for five transcription factors (TF) across ten species: *Homo sapiens*, *Macaca macaca*, *Mus musculus*, *Mus spretus*, *Mus caroli*, *Rattus norvegicus*, *Canis lupus familiaris*, *Felis catus*, *Oryctolagus cuniculus*, *Gallus gallus*, *Drosophila melanogaster* (Ballester et al. 2014; D. Schmidt et al. 2010; Dominic Schmidt et al. 2012; Rensch et al. 2016; Stefflova et al. 2013) (Supplementary Data Table 1). To ensure consistency across datasets, we re-processed all raw sequencing data using the NextFlow ChIP-seq pipeline v2.0 (Ewels et al. 2022), specifying Bowtie2 (Langmead and Salzberg 2012) as the aligner and using the narrow_peak option for peak calling with MACS2 (Y Zhang et al. 2008). For downstream analyses, we retained only peaks located on complete chromosomes, excluding unplaced scaffolds.

Read coverage was computed at each base within ChIP-seq peaks using bamCoverage, normalised by counts per million (Ramirez et al. 2014). For each experiment, we calculated the consensus peak coverage by summing normalised read counts across replicates. Peak summits were defined as the position with the highest normalised read coverage within each peak. In cases where multiple positions shared the maximum value, we selected the position with the highest mean coverage across a 10 bp window centred on the candidate site.

Whole genome alignment and ancestral state

We retrieved publicly available whole-genome alignments generated with Progressive Cactus. For *Drosophila melanogaster*, we used a 20-species alignment (Peng and Zhao 2024), and for vertebrates, we used the 241-species alignment from the Zoonomia Consortium (Armstrong et al. 2020). These alignments include both reference genomes and inferred ancestral genomes at each internal node of the species tree.

For each analysed species, we used the reference genome as the query and extracted alignments with the closest related species and their most recent common ancestor. Aligned sequences were retrieved using the hal2maf tool from the Comparative Genomics Toolkit (Hickey et al. 2013), with the noDupes option enabled to exclude putative duplicated regions. We extracted the alignment of each set of ChIP-seq peaks using their genomic coordinates as input for the mafsInRegion tool from UCSC (Nassar et al. 2023).

gkm-SVM model training and validation

To model the relationship between DNA sequence and TF binding affinity, we trained gapped k-mer support vector machine (gkm-SVM; Ghandi et al. 2014) models for each TF and tissue. The positive training set consisted of genomic regions identified as ChIP-seq peaks in the reference genome. The negative set was generated by randomly sampling genomic sequences matched for length, GC content, and repeat content using the genNullSeqs function from the gkmSVM R package (Ghandi et al. 2016). When pre-built BSgenome objects were not available for the genome assemblies analysed, we constructed them following the BSgenome R package guidelines (Pages 2024), using annotations from UCSC and GenBank (Supplementary Data Table 1).

Each model was trained using the gkmtrain function from the LS-GKM software (D. Lee 2016), with default parameters except for a k-mer length of 10 bp. Model performance was evaluated using 5-fold cross-validation using the -x 5 option. After training, we used the gkmpredict function to compute SVM weights for all possible 10-mers. Experiments with fewer than 1,000 peaks or with an associated model AUC below 0.8 were excluded from downstream analyses (Supplementary Figure 19).

SVM Score Computation for Peaks and Positions

We computed SVM scores for each ChIP-seq peak at two levels. At the peak level, the total SVM score was calculated as the sum of SVM weights from a 10 bp sliding window with a 1 bp step size. At the positional level, the SVM score for a specific nucleotide position was defined as the sum of SVM weights for all overlapping windows that include that position. These two levels of scoring were used for downstream analyses of both peak-level and site-specific effects.

In Silico Mutagenesis and ASVM Computation

For each ChIP-seq peak, we performed in silico mutagenesis by introducing all possible single-nucleotide mutations. The change in predicted binding affinity for each mutation was computed as ASVM, defined as the difference in SVM score between the mutated and ancestral sequence. This yielded a Distribution of Phenotypic Effects (DPE) for each peak, representing the genotype-to-phenotype landscape.

Observed substitutions were identified by comparing the reference sequence to its inferred ancestral state. Indels were removed using TrimAl with the “nogaps” option (Capella-Gutierrez et al. 2009). SVM scores were computed for both the reference and ancestral sequences using the same sliding window approach, and ASVM values were calculated for each observed substitution.

Substitution matrix inference

To account for variation in mutation and substitution rates across the genome, we estimated substitution matrices for each chromosome. Whole-genome alignments were split by chromosome using the mafSplit tool, and exonic regions were masked using GFF annotations. We then ran PhyML (Guindon et al. 2010) on each chromosome alignment using a Generalised Time-Reversible model with the params=r option to optimise substitution rate parameters. These matrices were used to weight mutation probabilities in the evolutionary model.

Evolutionary Model Inference via Likelihood Optimisation

To infer the selection regime acting on each ChIP-seq peak, we modeled the evolutionary process using three nested phenotype-to-fitness maps parameterized by Beta distributions: (i) a neutral model with a flat fitness landscape (a = p = 1), (ii) a stabilising selection model with maximal fitness at the ancestral phenotype (a = p # 1), and (iii) a directional selection model allowing asymmetric fitness landscapes (a # P).

For each possible single-nucleotide mutation within a peak, we computed the predicted phenotypic effect (ASVM), the corresponding selection coefficient (as the log-ratio of fitness between derived and ancestral phenotypes), and the fixation probability under an origin-fixation

model. These were combined with mutation probabilities derived from chromosome-specific substitution matrices to calculate a substitution probability for every in silico mutation.

The likelihood of the observed substitutions was then computed by summing the log-probabilities of the corresponding mutations under each model. Model parameters (a, P) were optimised using the scipy.optimize.minimize function (Virtanen et al. 2020). To compare the performance of each model and determine the most likely evolutionary scenario for each peak, we performed likelihood ratio tests using a threshold value of 1%. A detailed mathematical formalism of this framework is provided in Supplementary Materials.

Improvement on the Permutation Test

We refined the previously published Permutation Test to improve accuracy and interpretability (Liu and Robinson-Rechavi 2020). Mutation and substitution rates are now incorporated using chromosome-specific substitution matrices. Substitution positions are sampled without replacement to match the observed count, avoiding undersampling artefacts. We also implemented a two-sided test to detect both increases and decreases in binding affinity, and applied false discovery rate (FDR) correction to account for multiple testing.

Transcription factor binding motif recovery

To assess whether the gkm-SVM models captured known TF binding preferences, we extracted the SVM weights for all possible 10-mers from the trained models for human and mouse datasets. We selected the top 100 scoring 10-mers for each model and aligned them using Clustal Omega with default parameters (Sievers et al. 2011). The resulting alignments were submitted to the TomTom motif comparison tool (Gupta et al. 2007) in the MEME Suite website (Timothy L. Bailey et al. 2015) to identify matches against known TF motifs in the JASPAR CORE 2022 database (Castro-Mondragon et al. 2022).

Sequence conservation metrics

To compare model-based predictions with evolutionary conservation, we retrieved per-base conservation scores from the UCSC Genome Browser (Perez et al. 2025). Specifically, we used phastCons (Siepel et al. 2005) and phyloP scores (Pollard et al. 2010) from the 17-way vertebrate alignment for human, the 60-way glire subset for mouse, and the 27-way alignment for *Drosophila melanogaster*. For each ChIP-seq peak, we computed the conservation score at every nucleotide position and summarised the signal by calculating the average score across the peak.

Simulation of sequence evolution

To assess the accuracy and robustness of our selection inference framework, we simulated the evolution of regulatory sequences under controlled evolutionary scenarios. From randomly sampled drosophila CTCF ChIP-seq peak, we computed the predicted phenotypic effect (ASVM values) for all possible single-nucleotide mutations, along with mutation probabilities derived from chromosome-specific substitution matrices.

Substitutions in these peaks were sampled probabilistically according to their mutation rate, predicted phenotypic effect, and fixation probability under a specified evolutionary model. These models included neutral evolution, stabilising selection around the ancestral phenotype, and directional selection favouring increased or decreased binding affinity. Each model was parameterised using Beta distributions to define the shape and strength of the phenotype-to-fitness map.

Simulations were performed across a range of conditions, varying the number of substitutions per sequence, the strength of selection and the proportion of neutral mutations. This allowed us to generate synthetic datasets reflecting different evolutionary regimes, which were then used to benchmark the performance of RegEvol and the Permutations Test.

Drosophila divergence and polymorphism

Polymorphism data from the *Drosophila melanogaster* Genetic Reference Panel (DGRP) (Mackay et al. 2012) were retrieved and lifted over from the dm3 to the dm6 genome assembly using LiftOver (Hinrichs 2006). Overlap with TF peaks was determined using bedtools intersect. Fixed substitutions between *D. melanogaster* and *D. simulans* were counted within each peak, and a substitution-to-SNP ratio was calculated for every peak. Fisher's exact tests were then applied for each experiment to compare peaks inferred to have evolved under directional selection against all other peaks.

Gene Ontology and Anatomical Enrichments

Associations between ChIP-seq peaks and genes were obtained from the modERN consortium based on genomic proximity (Kudron et al. 2024). Gene Ontology enrichment analysis was performed using the *clusterProfiler* R package (version 4.16) with gene annotations from *org.Dm.eg.db* (version 3.21.0) for *Drosophila melanogaster* (Wu et al. 2021). Enrichment was computed against a background of all genes with at least one associated peak. We tested genes with at least one peak under directional selection, ranked gene lists based on either the total number of peaks or the number of peaks under directional selection, and genes with a directional-to-total peak ratio above 0.25 compared to all genes with at least two associated peaks.

Anatomical enrichment was carried out using TopAnat in the BgeeDB R package (version 2.34.0) with Bgee release 15.2 (Komljenovic et al. 2018; Bastian et al. 2025). TopAnat identifies anatomical entities where a gene set is preferentially expressed based on expression calls integrated by Bgee across anatomical structures. The analysis was restricted to RNA-seq expression calls and limited to *Drosophila melanogaster* fully formed stage (UBERON:0000066). Only the gene set with a directional-to-total peak ratio above 0.25 and the corresponding background with at least two peaks was tested. Enrichment was assessed using the *weight01* algorithm (Alexa et al. 2006), and anatomical entities with FDR < 0.01 were considered significant.

Drosophila coding sequences evolution

To assess whether genes associated with TF peaks also show signatures of coding sequence evolution, we retrieved w_0 and w estimates from Selectome (Moretti et al. 2014). Selectome applies branch-site codon models (Davydov et al. 2019) to orthologous gene alignments to estimate the strength of purifying selection (wo) and positive selection (W2) on specific phylogenetic branches. These values were extracted for *D. melanogaster* genes associated with TF peaks and used to test for correlations with the number of peaks and the directional-to-total peak ratio.

Tissue-level aggregation analysis of human CTCF binding sites

We analysed ChIP-seq peak data for CTCF in *Homo sapiens* identified by the ENCODE consortium and pre-processed in a previous study (ENCODE Project Consortium 2012; Liu and Robinson-Rechavi 2020). This dataset contains merged CTCF peaks across multiple tissues and cell types, allowing peaks to be assigned to broad physiological systems. RegEvol was applied to each peak to estimate the likelihood of each selective model.

To increase statistical power and test for coordinated signals of directional selection across biologically related elements, we implemented a tissue-level aggregation analysis inspired by the SUMSTAT framework of Daub et al. (2017). For each peak *i*, we computed the log-likelihood difference ($\Delta \log \mathscr{L}_i$) between two scenarios: first the maximum-likelihood ($\mathscr{L}$) of the observed data $\mathcal{O}_i$ derived under a scenario of directional selection, with $\bar{\alpha}$ $\bar{\beta}$ are the two parameters of directional selection estimated at the maximum (see Supplementary Materials for detailed equations on the computation for the likelihood), and second the likelihood for a scenario of no selection $\mathscr{L}[\mathcal{O}_i]$, as:

$$\Delta \log \mathscr{L}_i = \log\left(\mathscr{L}\left[\mathcal{O}_i \mid \bar{\alpha}, \bar{\beta}\right]\right) - \log\left(\mathscr{L}\left[\mathcal{O}_i\right]\right)$$

For each tissue, we performed 10,000 resamplings of 5,000 peaks to control for differences in sample size across tissues. For each resample, a cumulative SUM score was calculated as:

$$\mathrm{SUMSTAT} = \sum_i \sqrt[4]{\Delta \log \mathscr{L}_i}$$

Median and standard deviation of these SUM scores were then recorded for each tissue. Statistical significance of tissue-specific enrichment was assessed by comparing the distribution of SUM scores from each tissue to all other tissues using a Wilcoxon rank-sum test. Effect sizes were estimated using Cliff's delta. This approach allows weak but concordant signals of directional selection to accumulate and be detected at the level of biologically coherent groups.

Computation and Statistics

All data analyses were performed using R (version 4.5.0), Python (version 3.10), and Snakemake (version 7.24.0) for workflow management. The analyses were conducted on the high-performance computing infrastructure of the DCSR at the University of Lausanne. Multiple testing correction was applied using the Benjamini-Hochberg procedure to control the false discovery rate (FDR) across genomic regions. To assess potential ascertainment bias, we stratified ChIP-seq peaks by their SVM scores and evaluated detection rates across quantiles.

Data availability

The current manuscript is a computational study, so no data have been generated for this manuscript. Source code is available at *https://github.com/mrrlab/RegEvol*. Datasets used are available at *https://doi.org/10.5281/zenodo.17867859*.

Additional information

Author Contributions

M.R-R and A.L. originally conceived the project, A.L. and T.L. designed and implemented the method and computational framework. A.L. conducted all analyses and wrote the initial draft of the manuscript. All authors discussed the results and contributed to the final version. M.R-R supervised the project.

Funding

This work was supported by the Swiss National Science Foundation grant 207853 to Marc Robinson-Rechavi and grant 219757 to Nicolas Salamin.

Funding

Funder	Grant reference number	Author
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (SNF)	207853	Alexandre Laverré Marc Robinson-Rechavi
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (SNF)	219757	Alexandre Laverré

Author ORCID iDs

Marc Robinson-Rechavi: https://orcid.org/0000-0002-3437-3329

Additional files

Supplementary Figures

Supplementary Materials

Supplementary Table 1

Supplementary Table 2

Supplementary Table 3

References

1. **Albert Frank W,** Kruglyak Leonid (2015) The Role of Regulatory Variation in Complex Traits and Disease. *Nature Reviews Genetics* **16**:197-212 https://doi.org/10.1038/nrg3891 | PubMed
2. **Alexa Adrian,** Rahnenfuhrer Jorg, Lengauer Thomas (2006) Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure. *Bioinformatics* **22**:1600-1607 https://doi.org/10.1093/bioinformatics/btl140 | PubMed
3. **Armstrong Joel,** Hickey Glenn, Diekhans Mark, et al. (2020) Progressive Cactus Is a Multiple-Genome Aligner for the Thousand-Genome Era. *Nature* **587**:246-51 https://doi.org/10.1038/s41586-020-2871-y | PubMed
4. **Avsec Ziga,** Agarwal Vikram, Visentin Daniel, et al. (2021) Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions. *Nature Methods* **18**:10 https://doi.org/10.1038/s41592-021-01252-x | PubMed
5. **Bailey TL,** Boden M, Buske FA, et al. (2009) MEME SUITE: Tools for Motif Discovery and Searching. *Nucleic Acids Research* **37**:W202-8 https://doi.org/10.1093/nar/gkp335 | PubMed
6. **Bailey Timothy L,** Johnson James, Grant Charles E, Noble William S (2015) The MEME Suite. *Nucleic Acids Research* **43**:W39-49 https://doi.org/10.1093/nar/gkv416 | PubMed
7. **Ballester Benoit,** Medina-Rivera Alejandra, Schmidt Dominic, et al. (2014) Multi-Species, Multi-Transcription Factor Binding Highlights Conserved Control of Tissue-Specific Biological Pathways. *eLife* **3**:e02626 https://doi.org/10.7554/eLife.02626 | PubMed
8. **Bastian Frederic B,** Cammarata Alessandro Brandulas, Carsanaro Sara, et al. (2025) Bgee in 2024: Focus on Curated Single-Cell RNA-Seq Datasets, and Query Tools. *Nucleic Acids Research* **53**:D878-85 https://doi.org/10.1093/nar/gkae1118 | PubMed
9. **Bastian Frederic B,** Roux Julien, Niknejad Anne, et al. (2021) The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals. *Nucleic Acids Research* **49**:D831-47 https://doi.org/10.1093/nar/gkaa793 | PubMed
10. **Benegas Gonzalo,** Batra Sanjit Singh, Song Yun S (2023) DNA Language Models Are Powerful Predictors of Genome-Wide Variant Effects. *Proceedings of the National Academy of Sciences* **120**:e2311219120 https://doi.org/10.1073/pnas.2311219120 | PubMed
11. **Boer Carl G. de,** Taipale Jussi (2024) Hold out the Genome: A Roadmap to Solving the Cis-Regulatory Code. *Nature* **625**:41-50 https://doi.org/10.1038/s41586-023-06661-w | PubMed
12. **Buchon Nicolas,** Silverman Neal, Cherry Sara (2014) Immunity in Drosophila Melanogaster — from Microbial Recognition to Whole-Organism Physiology. *Nature Reviews Immunology* **14**:796-810 https://doi.org/10.1038/nri3763 | PubMed
13. **Capella-Gutierrez Salvador,** Silla-Martinez Jose M, Gabaldon Toni (2009) trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* **25**:1972-73 https://doi.org/10.1093/bioinformatics/btp348 | PubMed
14. **Castro-Mondragon Jaime A,** Riudavets-Puig Rafael, Rauluseviciute Ieva, et al. (2022) JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles. *Nucleic Acids Research* **50**:D165-73 https://doi.org/10.1093/nar/gkab1113 | PubMed
15. **Daub JT,** Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M (2017) Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. *Molecular Biology and Evolution* **34**:1391-402 https://doi.org/10.1093/molbev/msx083 | PubMed
16. **Davydov Iakov I,** Salamin Nicolas, Robinson-Rechavi Marc (2019) Large-Scale Comparative Analysis of Codon Models Accounting for Protein and Nucleotide Selection. *Molecular Biology and Evolution* **36**:1316-32 https://doi.org/10.1093/molbev/msz048 | PubMed

17. **Dorus Steve,** Vallender Eric J, Evans Patrick D, et al. (2004) Accelerated Evolution of Nervous System Genes in the Origin of Homo Sapiens. *Cell* **119**:1027-40 https://doi.org/10.1016/j.cell.2004.11.040 | PubMed

18. **Dukler Noah,** Huang Yi-Fei, Siepel Adam (2020) Phylogenetic Modeling of Regulatory Element Turnover Based on Epigenomic Data. *Molecular Biology and Evolution* **37**:2137-52 https://doi.org/10.1093/molbev/msaa073 | PubMed

19. **ENCODE Project Consortium** (2012) An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**:57-74 https://doi.org/10.1038/nature11247 | PubMed

20. **Ewels Philip,** Peltzer Alexander, Fillinger Sven, et al. (2022) The Nf-Core Framework for Community-Curated Bioinformatics Pipelines. Zenodo. version: v 2.4.1 https://doi.org/10.5281/zenodo.7139814

21. **Eyre-Walker Adam** (2006) The Genomic Rate of Adaptive Evolution. *Trends in Ecology & Evolution* **21**:569-75 https://doi.org/10.1016/j.tree.2006.06.015

22. **Eyre-Walker Adam,** Keightley Peter D (2007) The Distribution of Fitness Effects of New Mutations. *Nature Reviews Genetics* **8**:610-18 https://doi.org/10.1038/nrg2146 | PubMed

23. **Fu Wenqing,** Akey Joshua M (2013) Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics* **14**:467-89 https://doi.org/10.1146/annurev-genom-091212-153509 | PubMed

24. **Gallego Romero Irene,** Lea Amanda J (2023) Leveraging Massively Parallel Reporter Assays for Evolutionary Questions. *Genome Biology* **24**:26 https://doi.org/10.1186/s13059-023-02856-6 | PubMed

25. **Galtier Nicolas,** Duret Laurent (2007) Adaptation or Biased Gene Conversion? Extending the Null Hypothesis of Molecular Evolution. *Trends in Genetics* **23**:273-77 https://doi.org/10.1016/j.tig.2007.03.011 | PubMed

26. **Ghandi Mahmoud,** Lee Dongwon, Mohammad-Noori Morteza, Beer Michael A (2014) Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features. *PLOS Computational Biology* **10**:e1003711 https://doi.org/10.1371/journal.pcbi.1003711 | PubMed

27. **Ghandi Mahmoud,** Mohammad-Noori Morteza, Ghareghani Narges, Lee Dongwon, Garraway Levi, Beer Michael A (2016) gkmSVM: An R Package for Gapped-Kmer SVM. *Bioinformatics* **32**:2205-7 https://doi.org/10.1093/bioinformatics/btw203 | PubMed

28. **Guindon Stephane,** Dufayard Jean-Francois, Lefort Vincent, Anisimova Maria, Hordijk Wim, Gascuel Olivier (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**:307-21 https://doi.org/10.1093/sysbio/syq010 | PubMed

29. **Gupta Shobhit,** Stamatoyannopoulos John A, Bailey Timothy L, Noble William Stafford (2007) Quantifying Similarity between Motifs. *Genome Biology* **8**:R24 https://doi.org/10.1186/gb-2007-8-2-r24 | PubMed

30. **Halpern AL,** Bruno WJ (1998) Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Residue Frequencies. *Molecular Biology and Evolution* **15**:910-17 https://doi.org/10.1093/oxfordjournals.molbev.a025995 | PubMed

31. **He Bin Z,** Holloway Alisha K, Maerkl Sebastian J, Kreitman Martin (2011) Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. *PLOS Genetics* **7**:e1002053 https://doi.org/10.1371/journal.pgen.1002053 | PubMed

32. **Hickey Glenn,** Paten Benedict, Earl Dent, Zerbino Daniel, Haussler David (2013) HAL: A Hierarchical Format for Storing and Analyzing Multiple Genome Alignments. *Bioinformatics* **29**:1341-42 https://doi.org/10.1093/bioinformatics/btt128 | PubMed

33. **Hinrichs AS** (2006) The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research* **34**:D590-98 https://doi.org/10.1093/nar/gkj144 | PubMed

34. **Horisawa Kenichi,** Udono Miyako, Ueno Kazuko, et al. (2020) The Dynamics of Transcriptional Activation by Hepatic Reprogramming Factors. *Molecular Cell* **79**:660-676.e8. https://doi.org/10.1016/j.molcel.2020.07.012

35. **Hu Zhenhua,** Tee Wee-Wei (2017) Enhancers and Chromatin Structures: Regulatory Hubs in Gene Expression and Diseases. *Bioscience Reports* **37**:BSR20160183 https://doi.org/10.1042/BSR20160183 | PubMed

36. **Jiang Daohan,** Zhang Jianzhi (2024) Ascertainment Bias in the Genomic Test of Positive Selection on Regulatory Sequences. *Molecular Biology and Evolution* **41**:msad284 https://doi.org/10.1093/molbev/msad284 | PubMed

37. **Jordan Gregory,** Goldman Nick (2012) The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. *Molecular Biology and Evolution* **29**:1125-39 https://doi.org/10.1093/molbev/msr272 | PubMed

38. **Karollus Alexander,** Hingerl Johannes, Gankin Dennis, Grosshauser Martin, Klemon Kristian, Gagneur Julien (2024) Species-Aware DNA Language Models Capture Regulatory Elements and Their Evolution. *Genome Biology* **25**:83 https://doi.org/10.1186/s13059-024-03221-x | PubMed

39. **Khaitovich Philipp,** Enard Wolfgang, Lachmann Michael, Paabo Svante (2006) Evolution of Primate Gene Expression. *Nature Reviews Genetics* **7**:693-702 https://doi.org/10.1038/nrg1940 | PubMed

40. **Kim Seungsoo,** Wysocka Joanna (2023) Deciphering the Multi-Scale, Quantitative *Cis*-Regulatory Code. *Molecular Cell* **83**:373-92 https://doi.org/10.1016/j.molcel.2022.12.032 | PubMed

41. **King Mary-Claire,** Wilson Allan C (1975) Evolution at Two Levels in Humans and Chimpanzees. *Science* **188**:107-16 https://doi.org/10.1126/science.1090005 | PubMed

42. **Komljenovic Andrea,** Roux Julien, Wollbrett Julien, Robinson-Rechavi Marc, Bastian Frederic B (2018a) BgeeDB, an R Package for Retrieval of Curated Expression Datasets and for Gene List Expression Localization Enrichment Tests. *F1000Research* **5**:2748 https://doi.org/10.12688/f1000research.9973.2 | PubMed

43. **Komljenovic Andrea,** Roux Julien, Wollbrett Julien, Robinson-Rechavi Marc, Bastian Frederic B (2018b) BgeeDB, an R Package for Retrieval of Curated Expression Datasets and for Gene List Expression Localization Enrichment Tests. *F1000Research* **5**:2748 https://doi.org/10.12688/f1000research.9973.2 | PubMed

44. **Krieger Gat,** Lupo Offir, Wittkopp Patricia, Barkai Naama (2022) Evolution of Transcription Factor Binding through Sequence Variations and Turnover of Binding Sites. *Genome Research* **32**:1099-111 https://doi.org/10.1101/gr.276715.122 | PubMed

45. **Kudron Michelle,** Gevirtzman Louis, Victorsen Alec, et al. (2024) Binding Profiles for 961 Drosophila and C. Elegans Transcription Factors Reveal Tissue-Specific Regulatory Relationships. *Genome Research* **34**:2319-34 https://doi.org/10.1101/gr.279037.124 | PubMed

46. **Kurafeiski Jasmin D,** Pinto Paulo, Bornberg-Bauer Erich (2018) Evolutionary Potential of Cis-Regulatory Mutations to Cause Rapid Changes in Transcription Factor Binding. *Genome Biology and Evolution* **11**:406-14 https://doi.org/10.1093/gbe/evy269 | PubMed

47. **Lai Xuelei,** Stigliani Arnaud, Vachon Gilles, et al. (2019) Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Molecular Plant* **12**:743-63 https://doi.org/10.1016/j.molp.2018.10.010

48. **Langmead Ben,** Salzberg Steven L (2012) Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods* **9**:357-59 https://doi.org/10.1038/nmeth.1923 | PubMed

49. **Lee Dongwon** (2016) LS-GKM: A New Gkm-SVM for Large-Scale Datasets. *Bioinformatics* **32**:2196-98 https://doi.org/10.1093/bioinformatics/btw142 | PubMed

50. **Lee Phil H,** Lee Christian, Li Xihao, Wee Brian, Dwivedi Tushar, Daly Mark (2018) Principles and Methods of In-Silico Prioritization of Non-Coding Regulatory Variants. *Human Genetics* **137**:15-30 https://doi.org/10.1007/s00439-017-1861-0 | PubMed

51. **Lin Meixi,** Chakraborty Sneha, Carlos Eduardo G Amorim, et al. (2025) The Distribution of Fitness Effects Varies Phylogenetically across Animals. *bioRxiv* https://doi.org/10.1101/2025.05.13.653358 | PubMed

52. **Liu Jialin,** Robinson-Rechavi Marc (2020) Robust Inference of Positive Selection on Regulatory Sequences in the Human Brain. *Science Advances* **6**:eabc9863 https://doi.org/10.1126/sciadv.abc9863 | PubMed

53. **Mackay Trudy FC,** Richards Stephen, Stone Eric A, et al. (2012) The Drosophila Melanogaster Genetic Reference Panel. *Nature* **482**:173-78 https://doi.org/10.1038/nature10811 | PubMed

54. **Maurano Matthew T,** Humbert Richard, Rynes Eric, et al. (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**:7 https://doi.org/10.1126/science.1222794 | PubMed

55. **McDonald JH,** Kreitman M (1991) Adaptive Protein Evolution at the Adh Locus in Drosophila. *Nature* **351**:652-54 https://doi.org/10.1038/351652a0 | PubMed

56. **Moretti Sebastien,** Laurenczy Balazs, Gharib Walid H, et al. (2014) Selectome Update: Quality Control and Computational Improvements to a Database of Positive Selection. *Nucleic Acids Research* **42**:D917-21 https://doi.org/10.1093/nar/gkt1065 | PubMed

57. **Nassar Luis R,** Barber Galt P, Benet-Pages Anna, et al. (2023) The UCSC Genome Browser Database: 2023 Update. *Nucleic Acids Research* **51**:D1188-95 https://doi.org/10.1093/nar/gkac1072 | PubMed

58. **Pages Herve** (2024) BSgenome: Software Infrastructure for Efficient Representation of Full Genomes and Their SNPs. R Package. version: v 1.72.0 https://rdrr.io/bioc/BSgenome/

59. **Patwardhan Rupali P,** Hiatt Joseph B, Witten Daniela M, et al. (2012) Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo. *Nature Biotechnology* **30**:265-70 https://doi.org/10.1038/nbt.2136 | PubMed

60. **Peng Junhui,** Zhao Li (2024) The Origin and Structural Evolution of de Novo Genes in Drosophila. *Nature Communications* **15**:810 https://doi.org/10.1038/s41467-024-45028-1 | PubMed

61. **Perez Gerardo,** Barber Galt P, Benet-Pages Anna, et al. (2025) The UCSC Genome Browser Database: 2025 Update. *Nucleic Acids Research* **53**:D1243-49 https://doi.org/10.1093/nar/gkae974 | PubMed

62. **Phan Mai HQ,** Zehnder Tobias, Puntieri Fiona, et al. (2025) Conservation of Regulatory Elements with Highly Diverged Sequences across Large Evolutionary Distances. *Nature Genetics* **57**:1524-34 https://doi.org/10.1038/s41588-025-02202-5 | PubMed

63. **Pollard KS,** Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of Nonneutral Substitution Rates on Mammalian Phylogenies. *Genome Research* **20**:110-21 https://doi.org/10.1101/gr.097857.109 | PubMed

64. **Ramirez Fidel,** Dundar Friederike, Diehl Sarah, Gruning Bjorn A, Manke Thomas (2014) deepTools: A Flexible Platform for Exploring Deep-Sequencing Data. *Nucleic Acids Research* **42**:W187-91 https://doi.org/10.1093/nar/gku365 | PubMed

65. **Rensch Thomas,** Villar Diego, Horvath Julie, Odom Duncan T, Flicek Paul (2016) Mitochondrial Heteroplasmy in Vertebrates Using ChIP-Sequencing Data. *Genome Biology* **17**:139 https://doi.org/10.1186/s13059-016-0996-y | PubMed

66. **Rodrigue Nicolas,** Latrille Thibault, Lartillot Nicolas (2021) A Bayesian Mutation-Selection Framework for Detecting Site-Specific Adaptive Evolution in Protein-Coding Genes. *Molecular Biology and Evolution* **38**:1199-208 https://doi.org/10.1093/molbev/msaa265 | PubMed

67. **Schmidt D,** Wilson MD, Ballester B, et al. (2010) Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**:1036-40 https://doi.org/10.1126/science.1186176 | PubMed

68. **Schmidt Dominic,** Schwalie Petra C, Wilson Michael D, et al. (2012) Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell* **148**:335-48 https://doi.org/10.1016/j.cell.2011.11.058

69. **Sella Guy,** Petrov Dmitri A, Przeworski Molly, Andolfatto Peter (2009) Pervasive Natural Selection in the Drosophila Genome?. *PLOS Genetics* **5**:e1000495 https://doi.org/10.1371/journal.pgen.1000495 | PubMed

70. **Siepel A,** Bejerano G, Pedersen JS, et al. (2005) Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes. *Genome Research* **15**:1034-50 https://doi.org/10.1101/gr.3715005 | PubMed

71. **Siepel Adam,** Bejerano Gill, Pedersen Jakob S, et al. (2005) Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes. *Genome Research* **15**:1034-50 https://doi.org/10.1101/gr.3715005 | PubMed

72. **Sievers Fabian,** Wilm Andreas, Dineen David, et al. (2011) Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology* **7**:539 https://doi.org/10.1038/msb.2011.75 | PubMed

73. **Smith Justin D,** McManus Kimberly F, Fraser Hunter B (2013) A Novel Test for Selection on Cis-Regulatory Elements Reveals Positive and Negative Selection Acting on Mammalian Transcriptional Enhancers. *Molecular Biology and Evolution* **30**:2509-18 https://doi.org/10.1093/molbev/mst134 | PubMed

74. **Sokolova Ksenia,** Chen Kathleen M, Hao Yun, Zhou Jian, Troyanskaya Olga G (2024) Deep Learning Sequence Models for Transcriptional Regulation. *Annual Review of Genomics and Human Genetics* **25**:105-22 https://doi.org/10.1146/annurev-genom-021623-024727 | PubMed

75. **Soni Vivak,** Johri Parul, Jensen Jeffrey D (2023) Evaluating Power to Detect Recurrent Selective Sweeps under Increasingly Realistic Evolutionary Null Models. *Evolution* **77**:2113-27 https://doi.org/10.1093/evolut/qpad120 | PubMed

76. **Stefflova Klara,** Thybert David, Wilson Michael D, et al. (2013) Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell* **154**:530-40 https://doi.org/10.1016/j.cell.2013.07.007

77. **Tanaka T,** Nei M (1989) Positive Darwinian Selection Observed at the Variable-Region Genes of Immunoglobulins. *Molecular Biology and Evolution* **6**:447-59 https://doi.org/10.1093/oxfordjournals.molbev.a040569 | PubMed

78. **Tognon Manuel,** Kumbara Alisa, Betti Andrea, Ruggeri Lorenzo, Giugno Rosalba (2025) Benchmarking Transcription Factor Binding Site Prediction Models: A Comparative Analysis on Synthetic and Biological Data. *Briefings in Bioinformatics* **26**:bbaf363 https://doi.org/10.1093/bib/bbaf363 | PubMed

79. **Vaibhvi Vaibhvi,** Kunzel Sven, Roeder Thomas (2022) Hemocytes and Fat Body Cells, the Only Professional Immune Cell Types in Drosophila, Show Strikingly Different Responses to Systemic Infections. *Frontiers in Immunology* **13** https://doi.org/10.3389/fimmu.2022.1040510 | PubMed

80. **Vaishnav Eeshit Dhaval,** de Boer Carl G, Molinet Jennifer, et al. (2022) The Evolution, Evolvability and Engineering of Gene Regulatory DNA. *Nature* **603**:7901 https://doi.org/10.1038/s41586-022-04506-6 | PubMed

81. **Venkat Aarti,** Hahn Matthew W, Thornton Joseph W (2018) Multinucleotide Mutations Cause False Inferences of Lineage-Specific Positive Selection. *Nature Ecology & Evolution* **2**:1280-88 https://doi.org/10.1038/s41559-018-0584-5 | PubMed

82. **Villar Diego,** Berthelot Camille, Aldridge Sarah, et al. (2015) Enhancer Evolution across 20 Mammalian Species. *Cell* **160**:554-66 https://doi.org/10.1016/j.cell.2015.01.006 | PubMed

83. **Virtanen Pauli,** Gommers Ralf, Oliphant Travis E, et al. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**:261-72 https://doi.org/10.1038/s41592-019-0686-2 | PubMed

84. **Vorontsov Ilya E,** Kozin Ivan, Abramov Sergey, et al. (2025) Cross-Platform Motif Discovery and Benchmarking to Explore Binding Specificities of Poorly Studied Human Transcription Factors. *Communications Biology* **8**:1545 https://doi.org/10.1038/s42003-025-08909-9 | PubMed

85. **Ward Lucas D,** Kellis Manolis (2012) Interpreting Noncoding Genetic Variation in Complex Traits and Human Disease. *Nature Biotechnology* **30**:1095-106 https://doi.org/10.1038/nbt.2422 | PubMed

86. **Wisotsky Sadie R,** Pond Sergei L Kosakovsky, Shank Stephen D, Muse Spencer V (2020) Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Molecular Biology and Evolution* **37**:2430-39 https://doi.org/10.1093/molbev/msaa037 | PubMed

87. **Wittkopp Patricia J,** Kalay Gizem (2012) Cis-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence. *Nature Reviews Genetics* **13**:59-69 https://doi.org/10.1038/nrg3095 | PubMed

88. **Wong Emily S,** Zheng Dawei, Tan Siew Z, et al. (2020) Deep Conservation of the Enhancer Regulatory Code in Animals. *Science* **370**:eaax8137 https://doi.org/10.1126/science.aax8137 | PubMed

89. **Wu Tianzhi,** Hu Erqiang, Xu Shuangbin, et al. (2021) clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* **2**:100141 https://doi.org/10.1016/j.xinn.2021.100141 | PubMed

90. **Yan Han,** Hu Zhirui, Thomas Gregg WC, Edwards Scott V, Sackton Timothy B, Liu Jun S (2023) PhyloAcc-GT: A Bayesian Method for Inferring Patterns of Substitution Rate Shifts on Targeted Lineages Accounting for Gene Tree Discordance. *Molecular Biology and Evolution* **40**:msad195 https://doi.org/10.1093/molbev/msad195 | PubMed

91. **Yan Jian,** Qiu Yunjiang, dos Santos Andre M Ribeiro, et al. (2021) Systematic Analysis of Binding of Transcription Factors to Noncoding Variants. *Nature* **591**:147-51 https://doi.org/10.1038/s41586-021-03211-0 | PubMed

92. **Yang Z** (1998) Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution. *Molecular Biology and Evolution* **15**:568-73 https://doi.org/10.1093/oxfordjournals.molbev.a025957 | PubMed

93. **Yang Ziheng** (2014) *Molecular Evolution: A Statistical Approach* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602605.001.0001

94. **Zhang Feng,** Lupski James R (2015) Non-Coding Genetic Variants in Human Disease. *Human Molecular Genetics* **24**:R102-10 https://doi.org/10.1093/hmg/ddv259 | PubMed

95. **Zhang Yong,** Liu Tao, Meyer Clifford A, et al. (2008) Model-Based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137 https://doi.org/10.1186/gb-2008-9-9-r137 | PubMed

96. **Zhou Jian,** Troyanskaya Olga G (2015) Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model. *Nature Methods* **12**:10 https://doi.org/10.1038/nmeth.3547 | PubMed

Ballester B, et al (2014) Combinatorial transcription factor binding evolution in five placental mammals. NCBI BioProject. ID PRJEB1571 https://www.ncbi.nlm.nih.gov/bioproject/PRJEB1571

Schmidt D, et al (2010) CEBPA binding in five vertebrates. NCBI BioProject. ID PRJEB2006 https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2006

Schmidt D, et al (2012) CTCF binding evolution in mammals. NCBI BioProject. ID PRJEB2329 https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2329

Myers AN, et al (2021) Genome-wide mapping of regulatory regions in seven domestic cat tissues [ChIP-seq] (domestic cat). NCBI BioProject. ID PRJNA758414 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA758414

Rensch T, et al (2016) Mitochondrial heteroplasmy in vertebrates using ChIP-sequencing data. NCBI BioProject. ID PRJEB11182 https://www.ncbi.nlm.nih.gov/bioproject/PRJEB11182

Stefflova K, et al (2013) Transcription Factor binding evolution in five mouse species. NCBI BioProject. ID PRJEB1247 https://www.ncbi.nlm.nih.gov/bioproject/PRJEB1247

Ni X, et al (2012) Genome-wide comparative ChIP-seq data of CTCF and RNA-seq data in Drosophila white prepupa on Illumina Genome Analyzer. NCBI BioProject. ID PRJNA132691 https://www.ncbi.nlm.nih.gov/bioproject/PRJNA132691

Peng J, Zhao L (2024) Whole genome alignments of 20 Drosophila species by progressive cactus. Figshare. https://doi.org/10.6084/m9.figshare.19989395

Peer reviews

Reviewer #1 (Public review):

Summary:

In this manuscript, the authors present a method to detect natural selection on transcription factor binding sites (TFBSs), which is an upgraded version of a previously published method (Liu and Robinson-Rechavi, 2020). This upgraded version of the test implements more explicit models of evolution and is shown to outperform its predecessor in terms of both power and false positive rate. I think this method can be a valuable resource for the community and can be helpful not only to studies of TFBSs but also broader evolutionary questions related to genotype-phenotype maps or fitness landscapes.

Major comments:

(1) Questions related to Figure 1

Figure 1, along with the first section of the Results, shows that the SVM score and its sensitivity to mutations are generally correlated with the strength of ChIP-seq signals. It is not very clear to me, however, what the motivation is behind this part of the paper. It seems that the model used to predict binding strength is a pre-existing one, and it is unclear what is new in this section. Was the prediction model retrained using different data? Was its validity confirmed using new data? I would appreciate some more elaboration on how these results differ from what was presented in the previous study of Liu and Robinson-Rechavi (2020).

The existence of weak or negative correlations between SVM and coverage, which reportedly reflects low-quality peaks, seems applicable not only to this paper, but also to previous ones, so I would like to have it confirmed whether the question and the authors' answers apply to previous studies as well.

It is reported that SVM scores capture TF binding signals better than conservation-based statistics do. My intuitive interpretation is that both ChIP-seq peaks and SVM scores are supposed to reflect binding strength, whereas conservation is supposed to reflect selection (i.e., different definitions of "function" as mentioned above). It is not explicitly explained in the Results, however, what the difference indicates, leaving only an impression that the SVM score is "better" than the conservation statistics.

In summary, I think further elaboration on the above problems would make the flow of thought of this paper easier to follow.

(2) Lack of directional selection for low binding affinity

In the analysis of Drosophila melanogaster ChIP-seq peaks, there were more cases of directional selection for higher binding affinity than directional selection for lower binding affinity. The authors suggested that this observation is "likely biological" because the same pattern was not seen in simulations (line 412-413). I wonder if this could have resulted from a difference in the distribution of ancestral binding affinity across TFBSs between real and simulated data. If binding affinity was generally low in the common ancestor of D. melanogaster and D. simulans, selection for low binding affinity would manifest mainly as purifying selection against mutations that increase affinity instead of directional selection. Ancestral sequences for simulations, if I understood correctly, are observed peaks in D. melanogaster (line 715-719), which would include high fraction sequences that could be rarer in the real ancestral sequences.

The description of this particular result does not refer to a figure or table, nor is it revisited in the Discussion. Figure 5 treats peaks under directional selection as a single category. Taken together, it is hard to tell how this observation should be interpreted. If the authors consider this result as biologically meaningful, I would suggest adding more details (e.g., the number of each side).

(3) Selection in non-focal lineages

Regarding the detected signals of directional selection for stronger binding in certain tissues (Figure 6), I wonder if it is the focal species or those very tissues that are "special": did the human lineage undergo more adaptive regulatory evolution than the chimpanzee lineage, or do nervous and male reproductive systems have a high "propensity" for adaptive regulatory evolution? Assuming that the binding preference of the same TF did not undergo a significant change since human-chimpanzee split (which, I believe, is a built-in assumption in both RegEvo and the permutation test), it should be possible to perform the same test using chimpanzee sequences that are homologous to the human ChIP-seq peak regions. In the case of coding sequences, for example, Bakewell et al. (2007) found that it was the chimpanzee that had more genes under positive selection than humans; I wonder if TFBSs show the same or a different pattern.

(4) Comments on terminology

a) Meaning of "function"

The word "function" has had different meanings in the biology literature, with some authors using "functional" to refer to anything with a phenotypic effect and some using it only for targets of selection. A (putative) TFBS would be considered "functional" as long as it has TF binding affinity if we follow the effect-based definition, but only if its binding affinity is under selection if we follow the selection-based definition. In this manuscript, the term "function" appears to have been used to refer to TF binding but not selection, most notably in the first Results section. There are also places where it is less clear what "function" means exactly (e.g., "deeply conserved elements that are likely to be functionally important" of line 61). Since this paper is about evolution, it is likely that many readers prefer the selection-based definition or assume that the selection-based definition would be used. Thus, using "function" to refer to just TF binding could be confusing. To this end, I would suggest that the authors drop the word "function" or give an explicit definition early in this paper.

b) Directional selection in different directions

In this paper, selection for increased TF binding affinity is referred to as "positive directional selection", and selection in the opposite direction is called "negative directional selection" (as exemplified in Figure 2). I understand that using such shorthand names would make the text less clumsy, but these two terms could potentially be confusing, as "positive selection" and "negative (purifying) selection" are also terms referring to specific types of selection and have

some connection to directional and stabilizing selection. Therefore, I suggest that the authors use something like "selection for increased/decreased binding affinity" instead, or note explicitly in the text that "positive/negative directional selection" would be used as shorthand.

https://doi.org/10.7554/eLife.111237.1.sa2

Reviewer #2 (Public review):

Summary:

The manuscript by Laverre et al. provides an interesting new test of selection on TF binding. Rather than focusing on sequence changes, this test is specifically for changes in predicted TF binding affinity. The authors report directional selection on 5.1% of tested regions in Drosophila, as well as a signal of selection on CTCF binding in the human CNS and male reproductive system.

Strengths:

Overall, I think this represents an important direction for the field of molecular evolution: now that TF binding can be predicted fairly well from sequence, it can be a very useful focus for tests of selection.

Weaknesses:

As mentioned several times in the manuscript, Jiang and Zhang (2024) pointed out some issues with a previous permutation-based version of this test. Foremost among these was the issue of ascertainment bias: when testing only experimentally supported TF binding sites from a focal species, and then asking what type of selection (or lack of selection) led to those sites, one is guaranteed to find more substitutions that increase affinity, simply because the sites were selected in the first place as those with maximum (empirically measured) affinity.

To address this issue, the authors simulated Drosophila CTCF peaks evolving neutrally and then tested different ascertainment cutoffs in Figure 4D. It was not entirely clear to me what is shown in Figure 4D: the text says the bins were stratified by derived delta-SVM, whereas the figure says SVM, and the legend says derived SVM (both without the delta). I was unable to find any clarification of this in the Methods section. In any case, I am not really convinced by this, for two main reasons. First, when analyzing empirical ChIP-seq data, I would guess that only a tiny fraction of the genome is bound (far less than 1%, especially in mammalian genomes). However, the most extreme bin in Figure 4D is taking the top 10% of (delta?) SVM values. What would Figure 4D look like at bins of the highest 0.1%, 0.001%, etc? My guess is there would be a strong uptick in the FPR. The second reason is actually more important and fundamental than the first. As long as this method is working as described, I cannot see any way that it would ‘not’ be impacted by ascertainment bias. As an extreme case, imagine that all TF binding sites tested had the maximum possible SVM scores; then none of them would have any chance of showing directional selection against binding, while even those that evolved neutrally would appear to have directional selection in favor of binding. Of course, real empirical data are not as extreme as this, but the same concept applies in less extreme scenarios.

This bias could explain patterns observed in the real data. For example: "We observe much more positive than negative directional selection, a pattern likely biological rather than methodological, since it is absent from simulations." This is exactly the pattern predicted under ascertainment bias (in the extreme-scenario thought experiment above). I suspect it is absent from simulations simply because the authors did not properly account for this bias in their simulations.

If the main result reported by the authors had been a lack of any directional selection in favor of binding, and instead only neutrality or directional selection against binding, then this ascertainment bias would not be an issue- it would only have made their results conservative. Unfortunately, this is not the case, and the directional selection in favor of binding, which is the main result emphasized from the empirical analysis, could be inflated by this bias.

Minor point:

The following statement: "In contrast, phastCons and phyloP scores lack such enrichment and have a lower dynamic range, suggesting that the conservation scores are less sensitive to fine-scale variation of TF occupancy and thus regulatory region function" is only true if one assumes that TF binding is the only function of this region. One could even turn this around and say the fact that the sites affecting TF binding are not the most conserved is actually evidence that TF binding is not a good indicator of these regions' entire function. I suggest the authors soften this claim that conservation scores are less sensitive to regulatory region function.

https://doi.org/10.7554/eLife.111237.1.sa1

Author response:

> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *In this manuscript, the authors present a method to detect natural selection on transcription factor binding sites (TFBSs), which is an upgraded version of a previously published method (Liu and Robinson-Rechavi, 2020). This upgraded version of the test implements more explicit models of evolution and is shown to outperform its predecessor in terms of both power and false positive rate. I think this method can be a valuable resource for the community and can be helpful not only to studies of TFBSs but also broader evolutionary questions related to genotype-phenotype maps or fitness landscapes.*
>
> *Major comments:*
>
> *(1) Questions related to Figure 1*
>
> *Figure 1, along with the first section of the Results, shows that the SVM score and its sensitivity to mutations are generally correlated with the strength of ChIP-seq signals. It is not very clear to me, however, what the motivation is behind this part of the paper. It seems that the model used to predict binding strength is a pre-existing one, and it is unclear what is new in this section. Was the prediction model retrained using different data? Was its validity confirmed using new data? I would appreciate some more elaboration on how these results differ from what was presented in the previous study of Liu and Robinson-Rechavi (2020).*

We agree that the current manuscript does not clearly distinguish which parts of Figure 1 are novel and which are foundational. The SVM itself is not new and is the same as in Lee et al. (2016), as used in Liu & Robinson-Rechavi (2020). In the revision, we will explicitly state that the SVM used in Figure 1 is the standard gapped-kmer SVM (ls-gkm) approach. We retrained all gkm-SVM models de novo for each species-TF dataset, ensuring consistency across all analysed ChIP-seq peaks. For this, we recalled all ChIP-seq peaks in a homogeneous and robust manner using the nf-core ChIP-seq pipeline v2.0 (Ewels et al. 2022). Figure 1A confirms that the predicted binding affinity from the SVM correlates with experimental ChIP peak height. In addition, examining scores per site rather than per peak is new compared with Liu

and Robinson-Rechavi (2020). The correlations between the SVM-derived scores and other features had not been shown before to the best of our knowledge, thus Figure 1B-C is entirely novel. In other words, this analysis is meant to show that our phenotypic metric (SVM score per site) indeed tracks binding intensity, i.e. molecular phenotype.

> *The existence of weak or negative correlations between SVM and coverage, which reportedly reflects low-quality peaks, seems applicable not only to this paper, but also to previous ones, so I would like to have it confirmed whether the question and the authors' answers apply to previous studies as well.*

Yes, this is a well-known issue in ChIP-seq studies. Low coverage often matches weak predicted binding affinity scores because noisy or unreliable peaks naturally have weaker signals. This is not specific to our work, and it has been observed in many other studies (e.g., Bailey et al. 2013 doi:10.1371/journal.pcbi.1003326; Nakato and Shirahige 2017 doi:10.1093/bib/bbw023). It is simply an expected property of the data.

> *It is reported that SVM scores capture TF binding signals better than conservation-based statistics do. My intuitive interpretation is that both ChIP-seq peaks and SVM scores are supposed to reflect binding strength, whereas conservation is supposed to reflect selection (i.e., different definitions of "function" as mentioned above). It is not explicitly explained in the Results, however, what the difference indicates, leaving only an impression that the SVM score is "better" than the conservation statistics.*

While the reviewer is correct that there are different definitions of function, both conservation-based statistics and RegEvol seek to capture selected function. The difference is that RegEvol aims to measure functional change, whereas conservation-based statistics aim to detect sequences that retain the same function across species. In both cases, we expect a correlation with causal function (i.e., binding). We will clarify these concepts and how they apply to our results in the revised manuscript.

> *(2) Lack of directional selection for low binding affinity*
>
> *In the analysis of Drosophila melanogaster ChIP-seq peaks, there were more cases of directional selection for higher binding affinity than directional selection for lower binding affinity. The authors suggested that this observation is "likely biological" because the same pattern was not seen in simulations (line 412-413). I wonder if this could have resulted from a difference in the distribution of ancestral binding affinity across TFBSs between real and simulated data. If binding affinity was generally low in the common ancestor of D. melanogaster and D. simulans, selection for low binding affinity would manifest mainly as purifying selection against mutations that increase affinity instead of directional selection. Ancestral sequences for simulations, if I understood correctly, are observed peaks in D. melanogaster (line 715-719), which would include high fraction sequences that could be rarer in the real ancestral sequences.*
>
> *The description of this particular result does not refer to a figure or table, nor is it revisited in the Discussion. Figure 5 treats peaks under directional selection as a single category. Taken together, it is hard to tell how this observation should be interpreted. If the authors consider this result as biologically meaningful, I would suggest adding more details (e.g., the number of each side).*

We appreciate this insight. We agree that the text was not clear, but in fact, the simulations were performed using the reconstructed ancestral sequences of ChIP-seq peaks themselves. Thus, simulated and empirical results should be directly comparable, and different results should be due to biology. We will revise the Manuscript to explicitly state that simulations are performed from reconstructed ancestral sequences and why. We will also add more descriptive statistics of the simulated and real data.

> *(3) Selection in non-focal lineages*
>
> *Regarding the detected signals of directional selection for stronger binding in certain tissues (Figure 6), I wonder if it is the focal species or those very tissues that are "special": did the human lineage undergo more adaptive regulatory evolution than the chimpanzee lineage, or do nervous and male reproductive systems have a high "propensity" for adaptive regulatory evolution? Assuming that the binding preference of the same TF did not undergo a significant change since human-chimpanzee split (which, I believe, is a built-in assumption in both RegEvo and the permutation test), it should be possible to perform the same test using chimpanzee sequences that are homologous to the human ChIP-seq peak regions. In the case of coding sequences, for example, Bakewell et al. (2007) found that it was the chimpanzee that had more genes under positive selection than humans; I wonder if TFBSs show the same or a different pattern.*

This is an excellent suggestion. To compare in an unbiased manner, we would need transcription factor ChIP-seq from the same organs in chimpanzees and humans. We are not aware of such a dataset. If one is identified, we would be very interested in analysing it, and thus answer this question. As suggested by the reviewer, we will analyse the human homologous sequences. Although it should be clear that this will provide a biased estimate for comparing adaptation between the two species, as we will lack newly acquired binding sites in the chimpanzee.

> *(4) Comments on terminology*
>
> *(a) Meaning of "function"*
>
> *The word "function" has had different meanings in the biology literature, with some authors using "functional" to refer to anything with a phenotypic effect and some using it only for targets of selection. A (putative) TFBS would be considered "functional" as long as it has TF binding affinity if we follow the effect-based definition, but only if its binding affinity is under selection if we follow the selection-based definition. In this manuscript, the term "function" appears to have been used to refer to TF binding but not selection, most notably in the first Results section. There are also places where it is less clear what "function" means exactly (e.g., "deeply conserved elements that are likely to be functionally important" of line 61). Since this paper is about evolution, it is likely that many readers prefer the selection-based definition or assume that the selection-based definition would be used. Thus, using "function" to refer to just TF binding could be confusing. To this end, I would suggest that the authors drop the word "function" or give an explicit definition early in this paper.*

We thank the reviewer for this precision and fully agree, we will revise our terminology for clarity. We will clarify the distinction between selected function and causal function, and we will pay attention to their use throughout the manuscript.

> *(b) Directional selection in different directions*
>
> *In this paper, selection for increased TF binding affinity is referred to as "positive directional selection", and selection in the opposite direction is called "negative directional selection" (as exemplified in Figure 2). I understand that using such shorthand names would make the text less clumsy, but these two terms could potentially be confusing, as "positive selection" and "negative (purifying) selection" are also terms referring to specific types of selection and have some connection to directional and stabilizing selection. Therefore, I suggest that the authors use something like "selection for increased/decreased binding affinity" instead, or note explicitly in the text that "positive/negative directional selection" would be used as shorthand.*

We agree with this ambiguity in the current terminologies. We will replace the phrases "positive directional selection" and "negative directional selection" with, e.g., "selection for increased binding affinity" and "selection for decreased binding affinity" as suggested when presenting our biological result on ChIP-seq peaks. However, we will still use "positive/negative directional" for the general framework (genotype → phenotype →fitness map) and insert a note that we use "positive/negative directional" as shorthand to mean increasing/decreasing affinity in the case of CHIP-seq peaks.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *The manuscript by Laverre et al. provides an interesting new test of selection on TF binding. Rather than focusing on sequence changes, this test is specifically for changes in predicted TF binding affinity. The authors report directional selection on 5.1% of tested regions in Drosophila, as well as a signal of selection on CTCF binding in the human CNS and male reproductive system.*
>
> *Strengths:*
>
> *Overall, I think this represents an important direction for the field of molecular evolution: now that TF binding can be predicted fairly well from sequence, it can be a very useful focus for tests of selection.*
>
> *Weaknesses:*
>
> *As mentioned several times in the manuscript, Jiang and Zhang (2024) pointed out some issues with a previous permutation-based version of this test. Foremost among these was the issue of ascertainment bias: when testing only experimentally supported TF binding sites from a focal species, and then asking what type of selection (or lack of selection) led to those sites, one is guaranteed to find more substitutions that increase affinity, simply because the sites were selected in the first place as those with maximum (empirically measured) affinity.*
>
> *To address this issue, the authors simulated Drosophila CTCF peaks evolving neutrally and then tested different ascertainment cutoffs in Figure 4D. It was not entirely clear to me what is shown in Figure 4D: the text says the bins were stratified by derived delta-SVM, whereas the figure says SVM, and the legend says derived SVM (both without the delta). I was unable to find any clarification of this in the Methods section. In any case, I am not really convinced by his, for two main reasons. First, when analyzing empirical ChIP-seq data, I would guess that only a tiny fraction of the genome is bound (far less than 1%, especially in mammalian genomes). However, the most extreme bin in Figure 4D is taking the top 10% of (delta?) SVM values. What would Figure 4D look like at bins of the highest 0.1%, 0.001%, etc? My guess is there would be a strong uptick in the FPR.*

We apologise for the confusion in Figure 4D, we will clarify the caption and text and specify that bins are stratified by derived SVM (post-simulation binding affinity proxy), not genome % or ΔSVM.

We want to note that we used the same subsampling approach as Jiang and Zhang (2024) to evaluate ascertainment bias, and that Figure 4 both confirms the issue that they identified with Liu and Robinson-Rechavi (2020), and shows very clearly that RegEvol does not have the same issue (flat red lines). Following the reviewer's suggestion, we can extend the figure to 1% or 0.1% bins. We note that the % of the total genome is different from the % of peaks: while actual peaks cover a very small proportion of the genome, the subsampling in Figure 4

(and in Jiang and Zhang 2024) aims to estimate the impact of detecting only the strongest peaks.

One difference between Jiang and Zhang (2024) and our study is that we simulated using whole empirical peaks, whereas they simulated 10-nucleotide transcription-binding sites, meaning that each substitution represented a 10% change. We will clarify these differences in the revised text.

> *The second reason is actually more important and fundamental than the first. As long as this method is working as described, I cannot see any way that it would 'not' be impacted by ascertainment bias. As an extreme case, imagine that all TF binding sites tested had the maximum possible SVM scores; then none of them would have any chance of showing directional selection against binding, while even those that evolved neutrally would appear to have directional selection in favor of binding. Of course, real empirical data are not as extreme as this, but the same concept applies in less extreme scenarios.*
>
> *This bias could explain patterns observed in the real data. For example: "We observe much more positive than negative directional selection, a pattern likely biological rather than methodological, since it is absent from simulations." This is exactly the pattern predicted under ascertainment bias (in the extreme-scenario thought experiment above). I suspect it is absent from simulations simply because the authors did not properly account for this bias in their simulations.*
>
> *If the main result reported by the authors had been a lack of any directional selection in favor of binding, and instead only neutrality or directional selection against binding, then this ascertainment bias would not be an issue- it would only have made their results conservative. Unfortunately, this is not the case, and the directional selection in favor of binding, which is the main result emphasized from the empirical analysis, could be inflated by this bias.*

There is indeed a possible ascertainment bias, although we believe it concerns only the detection of negative directional selection, as long as we have only empirical peaks in the focal species and not the sister species. This is not so much a limitation of our method as an intrinsic limitation of asymmetrical sampling of species: to study both gain and loss of function, function must be studied experimentally in several species. We will revise the manuscript to highlight this limitation.

Concerning positive directional selection, the mathematical foundation of RegEvol makes it inherently robust to ascertainment bias for positive directional selection. RegEvol calculates the likelihood of the entire sequence of observed substitutions accounting for the starting ancestral state and the mutational landscape. In other words, the model does not assume a uniform probability of phenotypic change; instead, it models the probability of each nucleotide mutation to result in a substitution (i.e., go to fixation) depending on its phenotype.

In an extreme case where all tested TF binding sites had the maximum SVM score, detecting negative directional selection would indeed be impossible, as ancestral states would have had equivalent or lower scores. However, positive directional selection would be inferred only if the likelihood of observing the substitution pattern's deltaSVM distribution significantly exceeded that expected under the mutational landscape. If a sequence evolved neutrally but reached a maximum SVM score, the likelihood of detecting directional selection would depend on: either the ancestral state being close to maximum with few substitutions increasing SVM (resulting in low statistical power), or the ancestral state being distant with many neutral substitutions and rare chance shifts to maximum (where the substitution distribution would be indistinguishable from neutrality). Then, even in such an extreme

dataset, neutral evolution remains detectable, demonstrating RegEvol's strength beyond deltaSVM comparisons between two states.

> *Minor point:*
>
> *The following statement: "In contrast, phastCons and phyloP scores lack such enrichment and have a lower dynamic range, suggesting that the conservation scores are less sensitive to fine-scale variation of TF occupancy and thus regulatory region function" is only true if one assumes that TF binding is the only function of this region. One could even turn this around and say the fact that the sites affecting TF binding are not the most conserved is actually evidence that TF binding is not a good indicator of these regions' entire function. I suggest the authors soften this claim that conservation scores are less sensitive to regulatory region function.*

We thank the reviewer for this comment, the text will be revised to soften this claim. We will explicitly state that sequence conservation reflects general functional constraints, whereas sequence-to-phenotype predictions capture highly specific and lineage-specific TF-DNA interactions.

https://doi.org/10.7554/eLife.111237.1.sa0