Leveraging Deep Learning Models to Study Enhancer Evolution

A Thesis

submitted to

Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS–MS Dual Degree Programme

by

Amruthamshu A Koundinya

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2026

Supervisor: Dr. Julia Zeitlinger
Investigator, Stowers Institute for Medical Research

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

© 2026 Amruthamshu A Koundinya. All rights reserved.

Certificate

This is to certify that this dissertation entitled **"Leveraging Deep Learning Models to Study Enhancer Evolution"** towards the partial fulfillment of the BS–MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by **Amruthamshu A Koundinya** at the Stowers Institute for Medical Research, Kansas City, USA under the supervision of **Dr. Julia Zeitlinger**, Investigator, Stowers Institute for Medical Research, during the academic year 2025-2026.

Dr. Julia Zeitlinger

Investigator, Stowers Institute for Medical Research

Dr. Leelavati Narlikar

Associate Professor and Deputy Chair, Data Science, IISER Pune

Declaration

I hereby declare that the matter embodied in the report entitled **"Leveraging Deep Learning Models to Study Enhancer Evolution"** are the results of the work carried out by me at at the Stowers Institute for Medical Research, Kansas City, USA under the supervision of **Dr. Julia Zeitlinger**, Investigator, Stowers Institute for Medical Research, and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Amruthamshu A Koundinya

20211224

This thesis is dedicated to my grandfather (Ajja)

Contributions

Based on the CRediT (Contributor Roles Taxonomy), the specific contributions of the authors to this thesis are outlined in the table below:

Role	Contributor(s)
Conceptualization	Dr. Julia Zeitlinger, Haining Jiang, Amruthamshu A Koundinya
Methodology	Amruthamshu A Koundinya, Haining Jiang
Software	Dr. Charles McAnany, Amruthamshu A Koundinya, Stowers Institute for Medical Research
Validation	Amruthamshu A Koundinya
Formal analysis	Amruthamshu A Koundinya
Investigation	Amruthamshu A Koundinya, Haining Jiang
Resources	Zeitlinger Lab
Data Curation	Zeitlinger Lab , Amruthamshu A Koundinya
Writing - Original Draft	Amruthamshu A Koundinya
Writing - Review & Editing	Amruthamshu A Koundinya, Haining Jiang, Dr. Julia Zeitlinger
Visualization	Amruthamshu A Koundinya
Supervision	Dr. Julia Zeitlinger
Project Administration	Dr. Julia Zeitlinger
Funding Acquisition	Stowers Institute for Medical Research, Kansas City, USA

AI Usage Statement

This work was primarily human-generated, with generative artificial intelligence assisting in the technical details, linguistic clarity, and structural formatting of the manuscript. Significant portions of the literature review (1) were facilitated by NotebookLM to summarize key points from foundational papers, which were then rephrased and edited to fit the narrative of this thesis. All of the above references are properly cited to ensure academic integrity. To refine the narrative flow, Gemini 3 Pro was used for paraphrasing and stylistic adjustments. ChatGPT 5.3 and Gemini 2.5 assisted the formatting of mathematical notation and data tables into LaTeX. Gemini 3 Pro and NotebookLM were utilized to cross-check the logical consistency of theorem proofs and to identify potential conceptual limitations, with the resulting refinements and finalized proofs detailed in the Appendix A. This iterative process involved using AI to both proofread human-derived arguments and assist in the initial drafting of proof structures; however, in all instances, the final logical steps were manually audited and verified to ensure mathematical and conceptual integrity. Finally, GitHub Copilot (Claude Sonnet 4.5, Gemini 3 Pro) was used to assist in the generation of Python code for data analysis, though all such code underwent manual line-by-line verification and testing prior to execution. Despite the use of these tools, the authors maintain full responsibility for the intellectual content, the accuracy of the verified proofs, and the final conclusions presented in this work.

AI Attribution (AIA): Primarily Human-generated, Multi-tool assisted, Human-verified, Gemini 3 Pro / NotebookLM / ChatGPT 5.3 / GitHub Copilot (Claude Sonnet 4.5)

Tools used: NotebookLM (summarization/verification), Gemini (paraphrasing/ verification), ChatGPT (LaTeX formatting), and GitHub Copilot (code generation).

Acknowledgements

This thesis represents the culmination of immense effort, and it would not have been possible without the support, guidance, and encouragement of many individuals and institutions.

My deepest appreciation goes to my supervisor, Dr. Julia Zeitlinger. Your vision and intellectual rigor were the driving forces behind this research. More than just guidance, I want to thank you for giving me the space to truly explore my own ideas and lead the project. Your consistent encouragement shaped this work from a mere concept into a complete thesis.

I am indebted to my mentor, Haining Jiang. Thank you for introducing me to this world of research in deep learning and genomics; and for patiently helping me grasp everything from the smallest technical details to the bigger concepts. Your hands-on help with the computational skills and your dedication to shaping how I think critically were absolutely instrumental to my progress.

To my colleagues and fellow Zeitlinger Lab members: thank you for the intellectual companionship, the countless discussions, and the collective energy that made the lab an enriching place to work every day. My thanks also go to the lab manager, Sara Jackson, for her efficient management and reliable technical assistance.

I am sincerely grateful to the Stowers Institute for Medical Research for providing a welcoming and highly stimulating environment, and crucially, the funding without which this work would not have been possible. The institutional resources of the Stowers Institute for Medical Research and the Zeitlinger Lab created an exceptional research atmosphere.

As an international student, I want to extend my distinct appreciation to the administrative staff; your logistical support and efficiency made my transition and time here remarkably smooth.

This opportunity was also fostered by the progressive outlook of IISER Pune. Thank you for championing your students to seek global research experiences and for providing the institutional freedom that made this international collaboration a reality. I also thank Dr. Leelavati Narlikar, who not only introduced me to the world of genomics but also a part of my thesis committee, which enabled this journey. I must also express my profound gratitude to the DST-INSPIRE scholarship; being selected for a scholarship awarded to the top 1% of students is an honor that deeply enabled this journey.

Finally, I want to express my deepest gratitude to the friends and family who provided the essential emotional backbone for this endeavor. To my friends: thank you for the

crazy times, the endless emotional support, and the deliberate effort you made to keep in touch so frequently, even with us being on opposite sides of the world, and even if it meant staying up all night to chat. Your friendship was a lifeline that kept me grounded and motivated throughout this journey. Specifically, I want to extend a very special thank you to my batchmates with whom I shared both this academic journey at Stowers and the experience of living together in a new country - your companionship, the shared laughs, and the funny memories made the stay in a foreign country easy, bearable and most importantly, fun. I am so grateful for the bond we share and the support we provided each other during this time.

Most importantly, my deepest and most profound gratitude belongs to my parents and my entire family. Your belief in me has never wavered, even from the very beginning. Thank you for your never-ending support and the continuous motivation you provided toward my goals, especially during the times when the finish line felt out of reach. More than anything, I want to thank you for standing by me when I chose to pursue scientific research. I know it is a path that is not always the most common, predictable, or financially rewarding, but you supported my vision and decision unconditionally. You never questioned my choices; instead, you celebrated my curiosity and gave me the courage to take a less-traveled road. Your enduring love and encouragement from afar have been a constant source of strength. This milestone is not just a reflection of my hard work, but a testament to the sacrifices you have made and the foundation you built for me. This achievement is as much yours as it is mine.

List of Figures

List of Tables

List of Abbreviations

Abbreviation	**Meaning**
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
CWM	Contribution Weight Matrix
DNA	Deoxyribonucleic Acid
LAM	Low Affinity Motif
PWM	Position Weight Matrix
RNA	Ribonucleic Acid
SHAP	Shapley values
TF	Transcription Factor

Abstract

Gene regulation is fundamental to shaping morphological diversity, driven by cis-regulatory regions containing transcription factors motifs to orchestrate precise gene expression. While protein-coding genes are very well conserved in sequence, cis-regulatory regions are subject to strong sequence divergence. How enhancer sequences change while maintaining their function is not clear. Recent research has shown that chromatin accessibility depends on motif cooperativity that follows a flexible motif syntax and includes low-affinity motifs, providing a possible avenue by which motifs arise de novo and diverge over time. To test if evolutionary selection occurs at the level of chromatin accessibility, we used *Drosophila* trichome development as a model system. We comprehensively mapped the chromatin accessibility landscape across several *Drosophila* species by performing ATAC-seq on *D. melanogaster*, *D. erecta*, *D. ananassae*, and *D. mojavensis* embryos at the appropriate stage. This revealed that, despite considerable sequence divergence, the amount of chromatin accessibility in regulatory regions is highly conserved across species, consistent with evolutionary selection at this level.

To precisely identify in an unbiased way which motifs and motif cooperativity rules drive the levels of chromatin accessibility in each species, we trained BPReveal deep learning models to predict bias-free accessibility profiles from DNA sequence. Interpreting these models revealed that strong sequence divergence between species is associated with a high turnover of individual motif instances across orthologous regions, as well as changes in motif affinity. Nevertheless, the type of motifs and their syntax rules are largely conserved across species, suggesting that the trans-environment of transcription factors is conserved. Consistent with this, models trained on one species perform well in predicting the ATAC-seq data from another species, with only small losses in performance with larger evolutionary distances.

This suggests that cis-regulatory regions are not only subject to strong sequence divergence, but also change in the way they encode chromatin accessibility over evolutionary time. Since the chromatin accessibility levels are under strong evolutionary selection, these results suggest that cis-regulatory regions diverge rapidly because sequence changes have a relatively high probability of producing similar amounts of chromatin accessibility through an alternative sequence encoding. Taken together, our data support the hypotheses that the highly flexible sequence rules of chromatin accessibility are a facilitator of cis-regulatory sequence evolution.

Contents

Chapter 1

Introduction

1.1 Principles of gene regulation and the cis regulatory code

1.1.1 Fundamentals of gene regulation

During the development of any complex multicellular organism, a single precursor cell - the fertilized egg - must divide, multiply, and differentiate to give rise to a staggering variety of highly specialized cell types. The complete set of instructions required to orchestrate this remarkable biological transformation is encoded within the organism's genome.

To understand this, one can think of the organism's genome as an immense, inherited instruction manual necessary for building and maintaining life, with its text written entirely in the four-letter string of alphabets - DNA. Within this vast manual, a gene represents a specific, discrete segment of DNA that serves as a blueprint for producing a functional biological molecule, which is typically a protein. Following the central dogma of molecular biology, the cell "reads" a gene through a multi-step process: the linear DNA sequence is first transcribed into an mRNA molecule, which is subsequently translated into a chain of amino acids to build a protein. This translation step is governed by the universal genetic code where specific triplet sequences of RNA (codons) translate directly into specific amino acids. These proteins have very diverse functions: some, like actin and tubulin, form the rigid structural scaffolding of the cell; others act as enzymatic catalysts, driving the metabolic reactions necessary for survival; while others function as signaling molecules, transmitting information from the cell surface to the nucleus. In this sense, the gene repertoire functions as the cell's vocabulary.

However, viewing the genome simply as a collection of genes is akin to viewing a complex machine as a disorganized box of gears. A list of components, no matter how comprehensive, does not constitute a functioning system. Every single cell in a multi-

cellular organism contains an essentially identical genome, and therefore possesses the exact same library of genes. Yet, a neuron, a heart cell, and a liver cell exhibit completely different morphologies and perform entirely different physiological tasks. If the instruction manual is identical in every cell, how does this incredible cellular diversity arise? This is where gene regulation comes into the picture. Cells achieve their unique identities, structures, and functions not by possessing different genes, but through the precise, differential expression of the genes they share. In any given cell type, at any given moment, only a specific, highly regulated subset of genes is actively being "read" and converted into proteins.

Gene regulation is not a simple, binary system of "on" and "off" switches. It is highly dynamic, quantitative, and complex. It dictates exactly *when*, *where*, and *how much* of a specific gene should be activated or silenced in response to developmental cues or environmental stimuli. Achieving this precision is an absolute necessity for survival and proper development. Even subtle changes in the expression level of a single gene—can dramatically alter an organism's phenotype, disrupt development, or drive the progression of disease. For example, a slight downregulation of the *sox9* gene results in the severe craniofacial disorder, whereas the abnormal upregulation of genes like *myc* accelerates cell proliferation and drives cancer (Kim and Wysocka 2023).

The sheer complexity of this spatiotemporal control dictates that the genome must encode far more than just the structural blueprints for proteins; it must also encode a vast and highly sophisticated regulatory program. However, the protein-coding sequences of the genes themselves do not contain the instructions for their own deployment which historically led researchers to question where this regulatory information was stored and how complex traits evolve. For example, in 1971, Britten and Davidson proposed a model - where they hypothesized that batteries of "producer genes" (protein-coding genes) were controlled by coordinated "receptor" and "integrator" sequences (Britten and Davidson 1971). They speculated that the origin of evolutionary novelty occurs when repetitive sequences are translocated throughout the genome, effectively rewiring existing regulatory pathways to construct entirely new regulative systems. Though we now know that this is not true, to understand how cells orchestrate this incredibly precise control of gene expression, early molecular biologists realized they had to look beyond the genes themselves and investigate the non-coding genome.

1.1.2 Cis regulatory elements

The non-coding genome harbors the instructions for cis-regulation—the control of a target gene's expression by DNA elements located on the same chromosomal allele. These functional segments of non-coding DNA are collectively known as cis-regulatory elements (CREs). CREs function as the primary integration hubs for regulatory information,

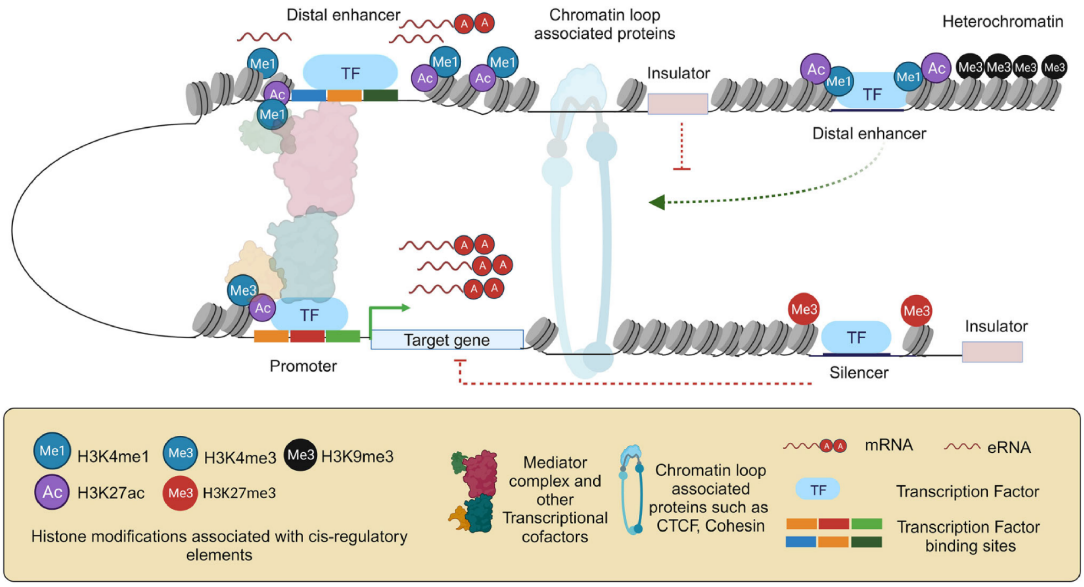

Figure 1.1: Cis regulatory elements
Source: https://doi.org/10.1007/s12038-024-00431-0 Adapted from Dsilva and Galande (2024)

processing combinatorial signals from the cellular environment to dictate the precise spatiotemporal levels of gene transcription. There are different types of CREs like promoters, enhancers, insulators etc - each playing distinct but coordinated roles in shaping the transcriptional landscape.

The most proximal of these elements is the core **promoter**, typically defined as the DNA region immediately surrounding the transcription start site (TSS). The core promoter serves as the foundational docking station for the assembly of the pre-initiation complex, which consists of general transcription factors and RNA Polymerase II. However, while the core promoter is required to initiate transcription, transcription is often weak in the absence of regulatory regions that are more distant from the TSS; most notably enhancers. **Enhancers** were first discovered over forty years ago as a short sequence in the SV40 viral genome that could stimulate the transcription of a linked reporter gene (Yáñez-Cuna et al. 2013). They can activate transcription largely independently of their distance relative to the target promoter. Unlike promoters, animal enhancers can be located hundreds, or even over a million base pairs away from the genes they regulate, residing within introns, downstream sequences, or intergenic regions. For example, the enhancer regulating the Sonic hedgehog (*Shh*) gene in the developing vertebrate limb is located inside the intron of an entirely different gene, over a million base pairs away from the *Shh* promoter (Cho 2012; Panigrahi and O'Malley 2021).

How such distal segments of non-coding DNA exert their regulatory influence relies fundamentally on the action of transcription factors (TFs). TFs are specialized proteins that serve as the primary trans-acting readers of the genome. They possess structured DNA-binding domains that allow them to scan the genome and recognize specific, short (typically 6 to 12 base pairs), degenerate DNA sequences called motifs or transcription

factor binding sites (TFBS). Enhancers typically contain multiple such TFBS for precise regulation.

In their inactive state, enhancers are occupied by nucleosomes, rendering them sterically inaccessible. Overcoming this chromatin barrier is not achieved through a single, universal pathway; rather, cells employ multiple mechanisms of enhancer activation. A prominent mechanism involves specialized *pioneer* transcription factors (TFs) that are capable of binding directly to nucleosomal DNA. These pioneers act as anchors, recruiting ATP-dependent chromatin remodeling complexes to evict or reposition nucleosomes, thereby establishing an accessible chromatin region. Alternatively, enhancer accessibility can be driven by collaborative competition, wherein multiple TFs bind transiently and collectively outcompete nucleosome occupancy to maintain an open chromatin state.

Once bound to an accessible enhancer, TFs do not generally catalyze transcription directly. Rather, their activation domains serve as physical adapters to recruit multi-protein arrays of coregulators. Together, these multi-protein assemblies alter the surrounding chromatin landscape, depositing characteristic epigenetic marks—such as histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac), which serve as robust biochemical signatures of active enhancers.

A persistent question in eukaryotic gene regulation is how these distal enhancers transmit their activating signals across vast genomic distances to target promoters. Several theoretical models have been proposed to explain this long-range communication.

The most widely accepted and experimentally validated framework is the looping model. In this paradigm, the intervening DNA between the enhancer and the promoter loops out, allowing the enhancer-bound TFs and recruited coactivators to come into direct physical proximity with the RNA Pol II machinery at the core promoter. Recent biophysical studies have further refined the static looping concept into a highly dynamic model involving liquid-liquid phase separation. In this framework, the intrinsically disordered regions of transcription factors and the mediator complex engage in multivalent, weak protein-protein interactions. These interactions drive the formation of highly concentrated, phase-separated droplets or "transcription bubbles". Rather than a rigid one-to-one loop, these condensates compartmentalize the transcriptional machinery, allowing multiple enhancers to simultaneously interact with a single or multiple promoters—facilitating rapid bursts of transcriptional initiation.

A key feature of the cis-regulatory architecture is its high degree of modularity. Genes governing complex developmental processes are rarely controlled by a single "master" enhancer; instead, they are regulated by an archipelago of multiple enhancer modules. The classic even-skipped (*eve*) gene in *Drosophila*, for instance, is regulated by five distinct enhancer modules, each responsible for integrating specific TF inputs to drive expression in distinct anterior-posterior stripes within the early embryo (Yáñez-Cuna et al. 2013; Meireles-Filho and Stark 2009). Similarly, the *Shh* gene mentioned above utilizes a mod-

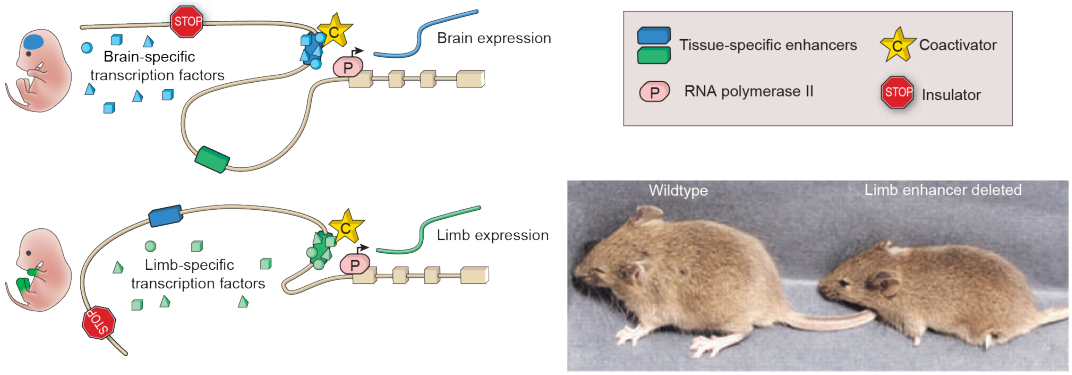

Figure 1.2: Brain and limb specific enhancers of *Shh* gene
Source: https://doi.org/10.1038/nature08451 Adapted from Visel et al. (2009)

ular architecture to direct its expression independently in the brain, spinal cord, and limbs. The distal limb enhancer of *Shh* acts as a highly specific "spatial switch" that restricts expression exclusively to the posterior margin of the developing limb bud (Cho 2012). By precisely activating *Shh* in only these cells, it establishes the morphogen gradient necessary for proper digit patterning and limb development. If this specific enhancer is deleted or mutated, the limb-specific program fails—resulting in severe truncation or malformations—while *Shh* expression in other tissues, such as the central nervous system, remains entirely unaffected. Ultimately, this modular architecture provides immense evolutionary flexibility. It allows genetic mutations to alter a single enhancer—thereby modifying a gene's expression in just one specific tissue or developmental stage—without causing detrimental, pleiotropic disruptions to the gene's essential functions elsewhere in the organism.

1.1.3 Cis regulatory code

While the physical elements of the non-coding genome constitute the "hardware" of transcriptional regulation, the actual instructions dictating their spatiotemporal deployment are embedded directly within the DNA sequence. The comprehensive set of rules governing how these short sequence motifs are arranged and interpreted by the cellular machinery is referred to as the cis-regulatory code. Whereas the genetic code relies on a nearly universal, discrete, and modular mapping of triplet codons to amino acids, the cis-regulatory code is highly context-dependent, quantitative, and degenerate (Boer and Taipale 2023). A specific regulatory sequence may be interpreted entirely differently depending on the precise combinations, concentrations, and signaling states of the transcription factors (TFs) present in a given cell type.

Short TF motifs appear millions of times across the genome by random chance, but the vast majority of these motif matches are non-functional and are never bound by TFs in vivo. Hence, the simple presence of a single motif is generally not sufficient to drive

enhancer activity. Instead, functional elements rely on the combinatorial requirement of multiple TFs, spacing, orientation etc. which we will see further. The architectural constraints of this syntax exist on a spectrum. At one extreme lies the rigid enhanceosome model, where the precise positioning, orientation, and phasing of motifs are strictly required for the cooperative assembly of a stable multi-protein complex. At the other extreme is the highly flexible billboard model, which integrates regulatory inputs from independently binding TFs with little constraint on their exact arrangement (Panigrahi and O'Malley 2021; Meireles-Filho and Stark 2009). Recent high-resolution genomic studies indicate that the majority of enhancers operate using *soft* syntax rules that lie between these two extremes (Section 1.3.2). In this study, we will explore the cis-regulatory rules at a certain developmental stage and its evolutionary dynamics across species.

Furthermore, a counterintuitive yet critical feature of the cis-regulatory code is its reliance on non-consensus, low-affinity binding motifs. While high-affinity sites strongly attract TFs, low-affinity motifs result in much shorter TF dwell times and thus require high local concentrations of TFs to become functionally occupied. This requirement establishes a strict concentration threshold for enhancer activation, providing a crucial mechanism for generating precise, tissue-specific expression patterns. Low-affinity motifs are hypothesized to fine-tune the cis-regulatory code, ensuring robust developmental specificity (Weilert et al. 2025). We shall explore what role these motifs play across evolution and their conservation in this study.

Because the cis-regulatory code relies on such nuanced features, traditional computational approaches have historically struggled to decode it. Predictive models relying on strict position weight matrices (PWMs) often fail to capture the degenerate and cooperative nature of TF binding in vivo. Accurately deciphering this hidden grammar requires advanced computational architectures capable of learning arbitrary, higher-order sequence features directly from raw genomic data without relying on rigid biological assumptions. As we shall see in Section 1.3, the advent of deep learning has revolutionized regulatory genomics, providing the ideal framework to finally decipher the rules of the cis-regulatory code.

1.2 Evolution of regulatory elements

1.2.1 Genetic basis of morphological evolution

In his 1977 essay "Evolution and Tinkering," François Jacob famously posited that nature operates not as an engineer designing from scratch, but as a tinkerer, continuously modifying and repurposing existing materials (Jacob 1977). For decades, the precise molecular substrate of this evolutionary tinkering remained unclear. It was once widely assumed that the emergence of distinct morphological traits and novel body plans required the continuous invention of entirely new protein-coding genes. The advent of comparative genomics provided a more nuanced reality. Mutations within protein-coding sequences certainly occur and play a vital role in evolution. Changes in coding regions are frequently responsible for physiological and biochemical adaptations—such as the evolution of specialized metabolic enzymes or distinct immune system components that differentiate species like mice and humans. However, coding sequence mutations alone cannot account for the vast diversity of animal forms.

Animals as morphologically disparate as fruit flies, mice, and humans share an ancient, deeply conserved genetic "toolkit" of developmental regulators. Because the functional integrity of these core structural and regulatory proteins is highly constrained across vast evolutionary distances, and these coding regions are very well conserved in terms of sequence. But the non-coding regulatory regions are not. Despite this sequence divergence, they seem to maintain their function in regulation, given that development is very robust even across species. This leads us to the question of what these regulatory changes are, how this regulatory activity is encoded and how it evolves.

These sequence changes in the regulatory regions that are responsible for rewiring the expression patterns are broadly categorized into two classes: trans-regulatory and cis-regulatory changes. Trans-regulatory mutations alter the sequence, activity, or availability of diffusible molecules, such as transcription factors (TFs) or signaling proteins. Because these factors operate diffusibly within the nucleus, a trans-mutation will typically affect the expression of both alleles of a target gene in a diploid cell. More importantly, because most TFs are deployed across multiple tissues and developmental stages to regulate hundreds of downstream targets, trans-acting mutations are highly pleiotropic. Modifying a TF to change a specific trait often disrupts its essential functions elsewhere in the organism, resulting in deleterious fitness consequences that are frequently purged by purifying selection. By contrast, cis-regulatory mutations occur in the local, non-coding DNA sequences (Wittkopp and Kalay 2011; Signor and Nuzhdin 2018). As mentioned in previous sections, a mutation within a single cis-regulatory element can alter gene expression only within the discrete spatial or temporal domain controlled by that specific module. This modular architecture allows evolution to bypass the constraints of pleiotropy: an

organism can tinker with a single tissue-specific enhancer to produce a morphological novelty without disrupting the gene's vital functions in other cell types (Wray 2007).

The distinct evolutionary dynamics of cis- and trans-regulation have been quantified using allele-specific expression analyses in F1 hybrids, a methodology pioneered in *Drosophila* by Wittkopp et al. (2004). Viewing some genes with this approach has revealed that while trans-acting mutations arise frequently due to the vast mutational target size of upstream regulatory networks, they primarily contribute to segregating intraspecific variation. In contrast, cis-regulatory mutations account for a significantly greater proportion of the fixed expression divergence between species. This indicates a model of regulatory evolution where trans-mutations are generally selected against in the wild, allowing the more precise, modular cis-regulatory mutations to preferentially accumulate and drive long-term phenotypic adaptation.

Concrete examples of this cis-regulatory modularity in nature illustrate how morphological traits evolve. Research by Sean Carroll and others has demonstrated that the independent gain and loss of male-specific wing pigment spots in different *Drosophila* lineages resulted directly from regulatory changes at the highly pleiotropic *yellow* gene. Rather than evolving entirely new pigmentation genes, these species evolved new wing patterns by co-opting and modifying distinct ancestral cis-regulatory elements. By accumulating simple mutations, these modular enhancers gained novel binding sites for highly conserved TFs, effectively rewiring the cis-regulatory code to paint a novel morphological trait onto the wing while leaving the essential, body-wide functions of the *yellow* gene intact (Gompel et al. 2005). Another textbook example involves the recurrent loss of pelvic fins in freshwater populations of threespine sticklebacks. Genetic mapping and transgenic analyses revealed that this dramatic skeletal reduction is driven by recurrent, independent deletions of a specific tissue-specific enhancer controlling the *Pitx1* gene. While *Pitx1* expression is abolished in the developing pelvic region of these fish, its expression in the thymus and olfactory pits remains completely intact, perfectly demonstrating how cis-regulatory mutations permit extreme morphological adaptation without lethal pleiotropy (Cho 2012; Wray 2007).

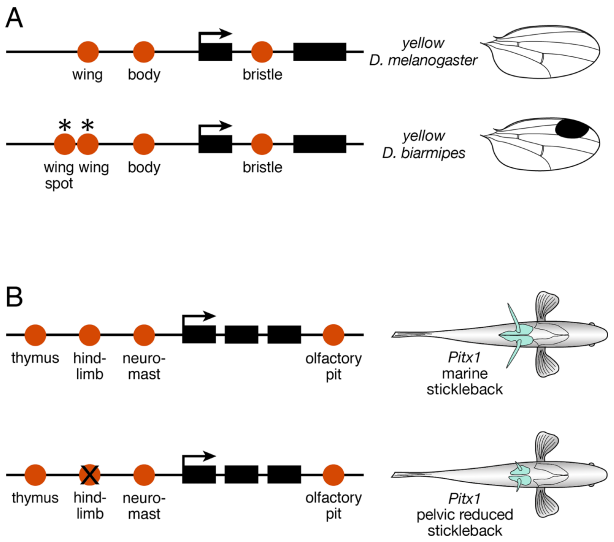

Figure 1.3: Examples of morphological changes due to cis-regulatory evolution
Source: https://doi.org/10.1371/journal.pbio.0030245
Adapted from Carroll (2005)

While these locus-specific studies provide elegant evidence that cis-regulatory changes

are a primary driver of morphological evolution, they inherently represent isolated cases. They strongly hint at a broader evolutionary principle. Historically, it has been incredibly difficult to track these evolutionary dynamics on a true genome-wide scale to see if these locus-specific rules apply universally across thousands of genes and regulatory elements. Here, in this research study, we harness the power of deep learning to fully understand the global dynamics of regulatory evolution.

To conduct such comprehensive, genome-wide evolutionary studies, selecting an appropriate model organism is critical. This study utilizes the fruit fly, *Drosophila*, as the primary model system. The *Drosophila* genus is exceptionally well-suited for studying the evolution of the cis-regulatory code because it encompasses a vast diversity of species that have diverged over tens of millions of years (with the family *Drosophilidae* diverging roughly 60-80 mya), offering rich morphological and genetic diversity. Furthermore, the genus benefits from a wealth of fully sequenced genomes, allowing for high-resolution comparative genomics across closely and distantly related species. Compared to mammalian genomes, *Drosophila* genomes are highly compact and exhibit a relatively high intrinsic rate of sequence evolution. This rapid evolutionary turnover combined with deeply conserved core developmental trajectory, provides a highly dynamic yet experimentally tractable landscape. Ultimately, it makes *Drosophila* an ideal model to study evolutionary changes.

1.2.2 Introduction to DNA models of evolution

In molecular phylogenetics, the primary goal is to infer the evolutionary history of organisms by comparing their genetic sequences. However, raw sequence data alone is often an unreliable metric for evolutionary time. To bridge the gap between observed sequence differences and actual evolutionary history, we use DNA substitution models. The fundamental challenge these models address is that the history of a DNA sequence is not fully preserved in its modern state. Over long evolutionary timescales, a single nucleotide site may undergo multiple changes that are invisible in the final comparison. DNA substitution models provide a statistical framework to correct for these unobserved changes.

It is important to distinguish that these are phenomenological models. They do not attempt to explicitly simulate the biochemical mechanisms of mutation. Instead, they describe the pattern and relative rates of the substitutions that accumulate over time, following the neutral theory of molecular evolution (Wikipedia contributors 2003; Wikipedia contributors 2005; Wikipedia contributors 2006). They ultimately enable us to build accurate phylogenetic trees and test complex biological hypotheses. Here, in this study, we use these models to test the effect of neutral evolution of the regulatory elements.

Math behind Substitution models

I have proved the theorems we need rigorously in Appendix A. But those not mathematically inclined can skip the whole math section and just read this section. All the math required is included here.

In the context of DNA, a single site/position in a sequence can exist in one of four distinct states: Adenine (A), Guanine (G), Cytosine (C), or Thymine (T). Over evolutionary time, a mutation may cause this site to switch from one state to another (e.g., A $\rightarrow$ G). It may cause more complex things like deletion or insertion of nucleotides, etc. Molecular evolution is modeled as a stochastic process describing the substitution of these residues/nucleotides over evolutionary time. The standard framework for DNA evolution relies on a Continuous-Time Markov Chain (CTMC) acting on a finite state space.

A Markov chain is a process describing a sequence of events where the probability of moving to the next state depends only on the current state, not on the history of how the system arrived there. This "memorylessness" is called the Markov property. For DNA, this means that if a nucleotide is currently an 'A', the probability that it will mutate to a 'G' in the next instant depends only on the fact that it is currently an 'A', not on whether it was a 'T' or 'C' millions of years ago. While some Markov chains move in discrete steps, DNA evolution occurs in continuous time. Mutations can happen at any random moment. Therefore, we model DNA evolution as a Continuous-Time Markov Chain (CTMC). Let us put this down mathematically.
The state space (i.e, the possible values of different states) of the DNA model S consists of the four nucleotides: $S = \{A, C, G, T\}$. Memorylessness - the probability of a transition from state i to state j depends exclusively on the current state i, independent of the prior evolutionary history (the path taken to reach state i). For a sequence of states X_t:

$$P(X_{t+h} = j \mid X_t = i, X_{t-1} = x_{t-1}, \dots) = P(X_{t+h} = j \mid X_t = i)$$

While describing this process, we deal with 2 primary matrices - P and Q.

- P matrix: $P(t)$ is the transition probability matrix.
  It answers the question: If I start at state i (e.g., A), what is the probability I will be at state j (e.g., G) after a specific time t? ($p_{ij}(t)$ = probability that a site currently in state i will be in state j after time t)

- Q matrix - Q is the instantaneous rate matrix.
  The Q-matrix defines the relative rates of change between nucleotides at an infinitesimal time step. It acts as the generator of the Markov chain. The elements q_{ij} $(i \neq j)$ represent the instantaneous rate of substitution from base i to base j.

Some properties of these matrices are:

- Because probabilities sum to 1, the rows of P always sum to 1 ($\forall i \sum_j P_{ij} = 1$)
- $P(t)$ depends on time. The probability of a change changes depending on whether we look at a branch of 1 million years or 100 million years. But Q is not dependent on time.
- The diagonal elements of Q, q_{ii} are defined such that the sum of each row is zero (representing the total rate of leaving state i) as we will see further

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

Assumptions of the model

1. Independent and identical distribution (i.i.d.)
   The model assumes that evolution at one site (nucleotide position) occurs independently of evolution at all other sites. The mutation probability at site k is not influenced by the state of site $k-1$ or $k+1$. Unless rate heterogeneity is explicitly added, the model assumes all sites are drawn from the same underlying distribution of rates and frequencies.

2. Time homogeneity
   The model assumes that the substitution process is constant over time. The Q-matrix is fixed across all branches of the phylogenetic tree. This implies that the "rules" of evolution (e.g., the transition-transversion bias κ) are the same for an ancient lineage as they are for a typically evolving modern lineage.

3. Stationarity
   The model assumes that the process is at equilibrium. The nucleotide frequencies $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ remain constant over time. Mathematically, this requires that the frequency vector π is the stationary distribution of the rate matrix Q, such that $\pi Q = 0$. This implies that if a sequence starts with frequencies π, it will maintain those frequencies indefinitely, despite individual mutations occurring.
   (*Why does such a π even exist? Go through the appendix for the mathematical details if interested*)

4. Time reversibility
   Most standard models (including HKY85) assume the process is chemically reversible. This means the amount of change from $i \to j$ is equal to the amount of change from $j \to i$ when the system is at equilibrium. This is defined by the balance equation:

$$\pi_i q_{ij} = \pi_j q_{ji}$$

Significance: Reversibility allows for the calculation of the likelihood of a tree without knowing the location of the root. It implies that the evolutionary process looks statistically identical whether time flows forward or backward, enabling the use of unrooted phylogenetic trees.

Let us first consider the probability matrix or the transition matrices which gives us the probability of transition from one state to another (state space $\mathcal{S} = \{A, G, C, T\}$).

$$P(t) = \begin{pmatrix} p_{\mathrm{AA}}(t) & p_{\mathrm{AG}}(t) & p_{\mathrm{AC}}(t) & p_{\mathrm{AT}}(t) \\ p_{\mathrm{GA}}(t) & p_{\mathrm{GG}}(t) & p_{\mathrm{GC}}(t) & p_{\mathrm{GT}}(t) \\ p_{\mathrm{CA}}(t) & p_{\mathrm{CG}}(t) & p_{\mathrm{CC}}(t) & p_{\mathrm{CT}}(t) \\ p_{\mathrm{TA}}(t) & p_{\mathrm{TG}}(t) & p_{\mathrm{TC}}(t) & p_{\mathrm{TT}}(t) \end{pmatrix}$$

Consider a DNA sequence of fixed length m evolving in time by base replacement.
Let $\mathbf{p}(t) = (p_A(t), p_G(t), p_C(t), p_T(t))$ be their respective probabilities at time t.
For two distinct $x, y \in S$, let μ_{xy} be the transition rate from state x to state y. Similarly, for any x, the total rate of change from x becomes

$$\mu_x = \sum_{y \neq x} \mu_{xy}.$$

Now, if at t_0, the Markov chain is in state E_i, then the probability that at time $t_0 + t$ it will be in state E_j depends only upon i, j and t.

$$p_A(t + \Delta t) = p_A(t) - p_A(t)\mu_A \Delta t + \sum_{x \neq A} p_x(t)\mu_{xA}\Delta t.$$

Basically, the probability of A at time $t + \Delta t$ is equal to the probability of A at time t minus the probability of the lost A plus the probability of the newly created A's. Same thing follows for other bases also - $p_G(t)$, $p_C(t)$ and $p_T(t)$. These equations can be written compactly as

$$\mathbf{p}(t + \Delta t) = \mathbf{p}(t) + \mathbf{p}(t)Q\Delta t$$

where Q is known as the rate matrix, $Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$

Result 1.2.1. $\mathbf{p}'(t) = \mathbf{p}(t)Q$ *where* Q = *rate matrix of a Continuous Time Markov chain.*

Proof. We have already seen that

$$\mathbf{p}(t+\Delta t) = \mathbf{p}(t) + \mathbf{p}(t)Q\Delta t$$
$$\frac{\mathbf{p}(t+\Delta t) - \mathbf{p}(t)}{\Delta t} = \mathbf{p}(t)Q$$

Now let $\Delta t \to 0$. We get that

$$\mathbf{p}'(t) = \mathbf{p}(t)Q.$$

Now we need to verify if Q is a proper matrix.
$\mathbf{p}(t)$ is a probability vector. So, it has to follow $p_i(t) \geq 0 \; \forall i$ and $\sum_i p_i(t) = 1 \; \forall t$.
That is, probability has to be non-negative and it should sum to 1.
Let

$$S(t) = \sum_i p_i(t)$$

Differentiating on both sides, we get

$$S'(t) = \sum_i p_i'(t) = \sum_i (p(t).Q)_i$$

From the matrix, we get that

$$S'(t) = \sum_x p_x(t)\left(\sum_i Q_{xi}\right)$$

We know that $\sum_i Q_{xi} = 0$ because each row of Q sums to 0, from our assumption and definition.
$\therefore S'(t) = 0 \; \forall \, \mathbf{p}(t)$ - the total probability is conserved. Q is a valid transition matrix and $\mathbf{p}'(t) = \mathbf{p}(t)Q$ □

Result 1.2.2. Stationarity : *The equilibrium row vector π must be annihilated by the rate matrix Q. ($\pi Q = 0$)*

Proof. We have proved above that $\mathbf{p}'(t) = \mathbf{p}(t)Q$. If the system is at stationarity, the base frequencies are not changing. So, the rate of change is 0. $\implies \mathbf{p}'(t) = 0$. Since, at equilibrium, $\mathbf{p}(t) = \pi$ (the equilibrium state frequencies) $\therefore \pi Q = 0$.
At equilibrium, the total "flux" leaving any state equals the total flux entering it. □

Getting from Q to P

We have derived that $\mathbf{p}'(t) = \mathbf{p}(t)Q$ (for those seeing the mathematical details, you have seen the more rigorous proof of this). Since we assumed time homogeneity (Q does not

depend on time), this differential equation can be solved with the initial condition that at time 0, no change has occurred ($P(0) = I$, the identity matrix). We get the following form from the exponent of a matrix.

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} Q^n \frac{t^n}{n!},$$

Different Models of Evolution

JC69 model

The simplest model was proposed by Jukes and Cantor (1969). It assumes equal base frequencies ($\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$) and equal mutation/substitution rates μ. This is very simplistic and does not allow us to do a lot of meaningful biological analysis.

HKY85 model

This model was proposed by Hasegawa, Kishino and Yano in 1985 (Hasegawa et al. 1985). I am introducing this because I have chosen this for further analysis, since though being not very simple and primitive, it is not too complicated and it takes into account several constraints.

It takes into account the difference in rates of transitions (purine $\leftrightarrow$ purine & pyrimidine $\leftrightarrow$ pyrimidine) and transversions (purine $\leftrightarrow$ pyrimidine). A,G are purines while C,T are pyrimidines. Transitions are more common than transversions because they cause smaller changes to DNA structure, making them harder for DNA repair systems to detect. Many frequent chemical mutations, like cytosine deamination, naturally produce transitions, and these mutations are more likely to slip through replication. Transitions are also less disruptive to proteins, so natural selection removes them less often than transversions, which tend to distort the DNA helix more and cause more harmful amino-acid changes. It also allows unequal base frequencies ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$), which is really useful since we know that *Drosophila* genomes are GC-rich.

These are the parameters used by the model: κ - ratio of transversions to transitions; π - the equilibrium state frequencies; μ - substitution rate; t - time duration of evolution (ν - we use branch length instead of μ, t - as we will see further)

$$\text{Rate matrix } Q = \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix}$$

1.2.3 Advantages of DNA models of evolution

Neutrality

At their core, these Continuous-Time Markov chain (CTMC) models align with the Neutral Theory of Molecular Evolution proposed by Motoo Kimura (Kimura 1983). (they rely on the mechanics of genetic drift rather than the mechanisms of mutation, we are distinguishing between how a change physically happens and how that change becomes a permanent part of the species' genome).

- Random genetic drift: The Neutral Theory asserts that most mutations are neither beneficial nor harmful; they are "neutral." That is, the majority of molecular changes are not due to organisms adapting to their environment, but rather due to random sampling errors acting on selectively neutral mutant alleles. Whether a neutral mutation spreads to the whole population or disappears is purely a game of chance—a "random walk" of allele frequencies known as genetic drift. The Markov chain treats evolution as a stochastic (random) process where a nucleotide flips from one state to another based on probability, not based on whether that change improves the organism's fitness.

- Molecular clock: A key prediction of Neutral Theory is that if mutations are neutral, they accumulate at a constant rate over time. DNA substitution models rely on this assumption to estimate branch lengths; they calculate the expected number of substitutions per site, assuming a steady accumulation of random changes. Motoo Kimura showed that if evolution is driven by drift, the rate at which substitutions occur (k) is exactly equal to the rate at which mutations arise (v).
  Because of this "mechanic" of drift, we don't need to model the complex population dynamics of every generation. We can assume that substitutions accumulate at a constant, steady rate over time. This is what allows us to use a constant rate matrix Q to calculate probabilities over millions of years.

- Independence: The models generally assume that every site evolves independently. This ignores complex selective interactions (epistasis) where a mutation at one site is only beneficial if another site changes, a hallmark of complex adaptive evolution.

They incorporate parameters that act as proxies for biological constraints—factors as mentioned before. They do so by allowing unequal base frequencies, allowing different transition rates for residues, etc. For example, mutational biases and purifying selection favoring conservative changes are probably both responsible for the relatively high rate of transitions compared to transversions in evolving sequences. However, model described above only attempts to capture the effect of both forces in a parameter that reflects the relative rate of transitions to transversions (κ).

In summary, these models are "neutral" because they simulate a random process without explicit fitness parameters, but they are biologically constrained and hence, not completely neutral because they parameterize the biases that result from the physical and chemical limitations of life.

Phylogenetic tree

Evolutionary relationships between biological entities are typically modeled using phylogenetic trees, which are branching structures inferred from morphological or sequence data. In these models, the terminal ends represent extant taxa, and the interior nodes signify hypothetical ancestors at historical divergence events. Furthermore, when represented as a phylogram, the tree's branch lengths are explicitly scaled to represent the precise amount of genetic substitution, illustrating both the hierarchy of the relationships and the accumulated evolutionary divergence (Wikipedia contributors 2002).

Building a tree using DNA substitution model

We have established the mathematical foundation for the DNA models of evolution/substitution. Now, given a DNA model of substitution, which tree topology makes our observed data most probable? In the Maximum Likelihood (ML) framework, the objective is to identify the specific tree topology (τ) and the associated set of branch lengths (ν) that maximize the probability of observing the sequence alignment given our model. For a candidate tree topology with specified branch lengths, the algorithm computes the probability of observing the specific configuration of nucleotides at the "tips" of the tree, for every single column in the DNA alignment. Since the ancestral states at the internal nodes are unknown, the likelihood for a single site is calculated using Felsenstein's Pruning Algorithm. This method sums over all possible nucleotide assignments (A, C, G, T) for every internal node, weighted by their probabilities. This marginal probability is derived using the transition probability matrix $P(t) = e^{Qt}$ defined earlier, which provides the probability of a substitution occurring over a branch of length t. The total likelihood of the tree is the product of the likelihoods of all individual sites. Though finding a global maximum is an NP-hard problem, softwares such as IQ-TREE distinguishes itself from standard heuristic approaches by employing a stochastic tree search algorithm (Methods 4.8.1).

There are several reasons why model based methods are better than building trees using simple methods like UPGMA, parsimony etc.

- If a site mutates from $A \to T$, and then later $T \to A$, simpler methods see zero changes. They assume the two species are identical at that site. A substitution model knows that over long time scales, "back-mutations" are statistically probable.

It will calculate a branch length that is longer than the raw number of differences that simple methods suggest.

- Imagine two species (A and B) evolve very fast (long branches), and two others (C and D) evolve slowly. By pure chance, A and B will accumulate some identical mutations (homoplasy). Parsimony sees these shared mutations and groups A and B together as sisters, which maybe wrong. ML and Bayesian methods separate rate from topology. The model expects long branches to have more random collisions. A model with rate heterogeneity can distinguish if the species are truly related or if they both have high mutation rates.

- Not all mutations are weighted equally. Based on the model and its parameters, mutations are weighted appropriately. For example, a shared *rare* mutation (like a transversion, $A \to T$) is stronger evidence of shared ancestry than a shared *common* mutation (like a transition, $A \to G$).

As a result, the observed differences often underestimate the true number of substitutions along the evolutionary path separating the sequences. Hence, we use these DNA substitution models to accurately estimate the evolutionary history.

Distances

Normally, a branch length of a phylogenetic tree is expressed as the expected number of substitutions per site. Sometimes a branch length is measured in terms of geological years. But some species evolve at faster rates than others. So, these two measures of branch length are not always in direct proportion. The expected number of substitutions per site per year is the widely used parameter (μ) (Wikipedia contributors 2005; Wikipedia contributors 2002).

Standard rate matrices provide relative substitution probabilities but must be normalized so that a branch length of 1 corresponds to exactly one expected substitution per site. We achieve this by replacing the standard time parameter (μt) with the product $\beta\nu$. In this formulation, ν represents the actual branch length estimated from the sequence data, while β acts as a fixed scaling factor derived directly from the rate matrix. By forcing the total expected substitution flux out of all states to 1, β is calculated as:

$$\beta = \frac{1}{-\sum_i \pi_i \mu_{ii}}$$

For the HKY85 model specifically, this scaling factor becomes:

$$\beta = \frac{1}{2(\pi_A + \pi_G)(\pi_C + \pi_T) + 2\kappa[(\pi_A \pi_G) + (\pi_C \pi_T)]}$$

1.3 Deep learning as a lens for genomics

1.3.1 Introduction to sequence-to-function models

As discussed in previous sections, the evolutionary plasticity of enhancers and the degenerate, highly combinatorial nature of the cis-regulatory code present a formidable decoding challenge. Historically, deciphering this regulatory grammar relied on traditional computational methods, such as sequence alignment for identifying conserved regions or scanning genomes with Position Weight Matrices (PWMs) to find transcription factor binding sites. However, these approaches are severely limited. PWM scanning often produces massive amounts of false-positive predictions because it ignores the broader sequence context and the cooperative syntax required for in vivo binding.

The advent of deep learning, specifically the application of convolutional neural networks (CNNs), has revolutionized regulatory genomics by providing a solution to these limitations. Unlike traditional machine learning algorithms that rely on rigid assumptions or manually hand-crafted features, deep neural networks perform adaptive feature extraction. They are predictive models composed of hierarchical layers that learn complex pattern detectors directly from the data. By taking raw, one-hot encoded DNA sequences as input and mapping them directly to experimental readouts, such as TF binding, chromatin accessibility, or histone modifications—these networks established the paradigm of "sequence-to-function" modeling. Early pioneering models such as DeepSEA and Basset successfully demonstrated that deep CNNs could simultaneously predict large-scale chromatin-profiling data across different cell types (Zhou and Troyanskaya 2015; Kelley et al. 2016). Basset, for example, learned the complex regulatory code of DNA accessibility across many cell types. By learning sequence representations at multiple spatial scales, these early models naturally captured the contextual dependencies of regulatory elements without requiring researchers to define motif syntax a priori. As computational power and architectural designs have advanced, the scope of sequence-to-function models has expanded dramatically. While early CNNs were restricted to relatively short input sequences, restricting their ability to detect distal regulatory elements, newer architectures have bypassed this limitation. Models such as Enformer incorporated self-attention mechanisms (transformers) to integrate information from long-range genomic interactions up to 100 kilobases away, effectively capturing complex enhancer-promoter communication (Avsec et al. 2021b). More recently, models like Borzoi and AlphaGenome have scaled this approach to megabase-length inputs, enabling the simultaneous, highly accurate prediction of diverse, multi-modal outputs ranging from 3D contact maps and splicing junctions to RNA-seq expression coverage (Linder et al. 2025; Avsec et al. 2026).

Despite their unprecedented predictive power, deep neural networks are frequently criticized as "black boxes," because their sequence features and regulatory logic are

learned in a highly distributed and hidden manner. However, in genomics, high prediction accuracy is merely a means to an end; the ultimate goal is biological discovery. Interpretability is what makes deep learning biologically relevant. It is through post-hoc interpretation methods that we can open the black box to extract the learned cis-regulatory grammar, identifying novel motifs, cooperative syntax rules, and the precise impact of evolutionary mutations. As we will explore in detail in Section 1.3.3, advanced interpretation toolkits transform these predictive networks from mere computational classifiers into powerful in-silico oracles for studying gene regulation.

1.3.2 Decoding motif syntax and regulatory code

While early convolutional neural networks (such as Basset) and newer transformer-based architectures (such as Enformer) successfully predict genomic features across large contexts, they typically suffer from a fundamental resolution trade-off. These models predict binary binding events or coarse, smoothed signals across large genomic windows. While sufficient for identifying broad regions of activity, this low-resolution aggregation blurs fine-scale regulatory features, such as the precise footprints of bound TFs and the exact spacing between adjacent motifs. Also, they are not trained on high resolution data like ChIP-nexus. Consequently, these models don't capture the high-resolution rules of motif syntax and TF cooperativity.

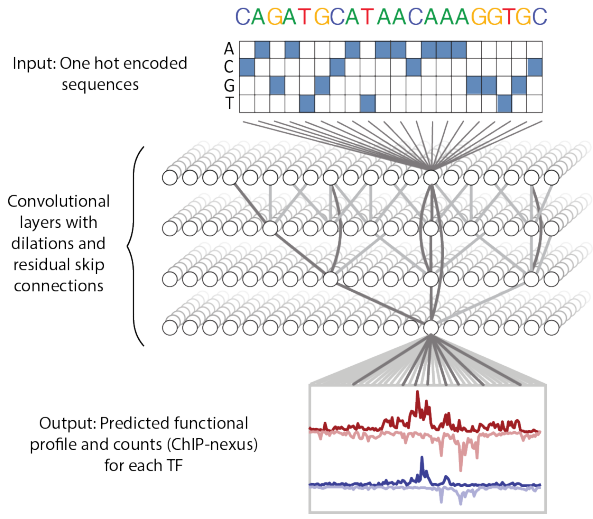

Figure 1.4: BPNet architecture

To overcome this limitation and extract the precise grammar of the cis-regulatory code, a deep learning model called BPNet was developed, designed explicitly to map DNA sequences to TF binding profiles at single-base resolution (Avsec et al. 2021a). Instead of predicting binary binding events or smoothed signals, BPNet is trained on high-resolution experimental data, such as ChIP-nexus, to directly predict the raw base-resolution read count profiles and the exact shape of TF footprints. To achieve this without sacrificing the necessary sequence context, BPNet employs a unique architectural innovation: it uses dilated convolutional layers with residual skip connections. By exponentially increasing the dilation rate—the number of skipped positions within the convolutional filter—at each successive layer, BPNet achieves a wide receptive field of over 1 kb. Crucially, it accomplishes this while strictly avoiding the pooling operations that reduce resolution in other

architectures, preserving base-pair precision throughout the network.

The resolution of BPNet allowed it to uncover sequence rules of the cis-regulatory code, specifically demonstrating that TF binding relies heavily on soft motif syntax. By utilizing the trained model as an in-silico oracle to test synthetic combinations of motifs, researchers discovered that TFs interact cooperatively in a flexible, distance-dependent manner rather than requiring rigid, exact spacing. For example, BPNet revealed a ∼10.5 base-pair helical periodicity for the pluripotency factor Nanog. This soft syntax rule facilitates cooperative protein-protein interactions by ensuring the interacting factors are positioned on the same face of the DNA molecule, without demanding a single, strict inter-motif distance. The principles established by BPNet have subsequently been extended to model chromatin accessibility data, such as ATAC-seq, through the development of the ChromBPNet architecture (Pampari et al. 2024). A critical challenge in modeling accessibility data at base resolution is the strong enzymatic cleavage bias introduced by the assay itself, such as the specific sequence insertion preference of the Tn5 transposase used in ATAC-seq. ChromBPNet addresses this by utilizing a bias factorization approach. It employs a separate neural network model to explicitly learn and regress out the assay's intrinsic cleavage bias during training. By separating these artifactual biases from the true biological signal, ChromBPNet can accurately predict functional TF footprints and high-resolution chromatin accessibility profiles directly from the sequence. We will be using a similar architecture in this study as we aim to understand the cis-regulatory rules underlying chromatin accessibility. Together, these base-resolution models provide an incredibly precise lens for deciphering the structural rules of gene regulation.

1.3.3 Model interpretation

Despite their unprecedented predictive power, deep neural networks are frequently criticized as opaque "black boxes" because their sequence features and regulatory logic are learned in a highly distributed, hidden manner. In regulatory genomics, however, high prediction accuracy is merely a means to an end; the ultimate goal is biological discovery. To extract the learned cis-regulatory grammar, researchers rely on feature attribution methods that quantify the exact importance of every single nucleotide within an input sequence for a given prediction.

A highly intuitive approach to feature attribution is in silico saturation mutagenesis (ISM), wherein we computationally introduce every possible single-nucleotide substitution along a sequence and directly measure the resulting change in the model's output. While ISM provides a straightforward measurement of mutational impact, it becomes computationally expensive when applied across large genomic windows.

To address the computational bottleneck of ISM, highly efficient backpropagation-based attribution algorithms were developed, most notably DeepLIFT (Shrikumar et

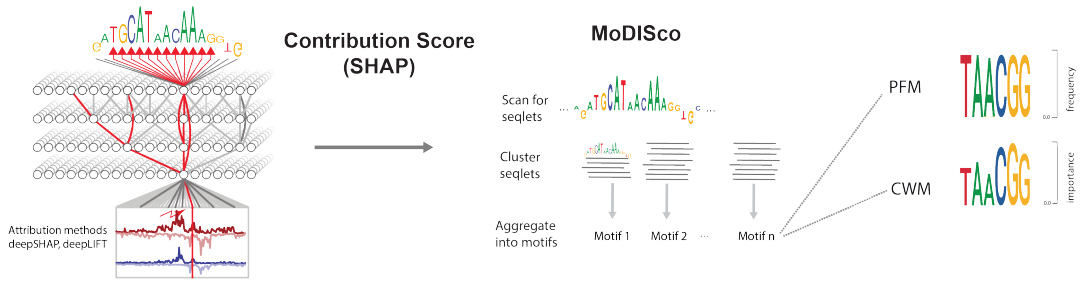

Figure 1.5: Interpretation tools - Shapley values, MoDISco etc.

al. 2017). DeepLIFT calculates base-resolution contribution scores by decomposing the difference between a model's prediction for a given sequence and the prediction for a neutral reference sequence. It assigns a specific weight to each base based on its predictive influence by systematically backtracking the signal from the output layer down to the input DNA sequence through the network.

However, both standard ISM and the original DeepLIFT algorithm evaluate feature importance in a relatively localized manner. In a highly combinatorial cis-regulatory code, calculating a nucleotide's true functional importance theoretically requires evaluating its effect across all possible combinations and subsets of other features—essentially performing a fully exhaustive, subset-wise ISM. This rigorous theoretical ideal is formally captured by SHAP (Shapley Additive exPlanations), a framework grounded in cooperative game theory. The mathematical concept was originally designed to fairly distribute a total "payout" among a coalition of "players." In the context of genomic sequence-to-function models, the "payout" is the model's overall prediction score, and the "players" are the individual nucleotides composing the input sequence. By computing the Shapley values, the framework objectively calculates the marginal contribution of each nucleotide across all possible combinations of sequence features.

Because calculating exact Shapley values is mathematically intractable for large deep learning networks, researchers utilize DeepSHAP, an adaptation that directly leverages the backpropagation algorithm to approximate Shapley values efficiently. While original DeepLIFT implementations often relied on a single reference sequence (such as an all-zero background or the expected value of inputs), utilizing a single baseline can introduce severe attribution biases, such as ignoring functional features in certain genomic contexts. DeepSHAP provides a theoretical justification for using a diverse background distribution instead of a single reference. By computing Shapley values against multiple references drawn from a background distribution and averaging the resultant attributions, DeepSHAP generates highly robust, theoretically sound saliency maps (McAnany et al. 2025).

Once these base-resolution contribution scores are calculated across thousands of regulatory regions, they must be synthesized into decipherable biological rules. This is

achieved through motif discovery algorithms such as TF-MoDISco, which systematically identifies, aligns, and clusters highly predictive subsequences (seqlets) into consolidated motifs (Shrikumar et al. 2018). A revolutionary output of this process is the creation of Contribution Weight Matrices (CWMs). Unlike traditional Position Frequency Matrices (PFMs) that merely capture the statistical over-representation of base frequencies, CWMs represent the average contribution scores of each base within a motif cluster. Consequently, CWMs highlight the exact sequence features that are functionally and predictively important for transcription factor binding in vivo, filtering out non-functional background noise. Beyond extracting static motifs, trained deep learning models can be utilized as in silico oracles to explicitly decode the complex, higher-order syntax of the cis-regulatory code. Because these networks capture the biophysical reality of transcription factor cooperativity without relying on predefined biological assumptions, we can use the models to design synthetic sequences or test the effect of genomic mutations. Ultimately, this suite of interpretation tools extracts the precise regulatory grammar of DNA.

In this study, we will use the above mentioned methods and tools to study the cis-regulatory rules in *Drosophila* embryo and dynamics of these rules across evolution (in different species). We shall be focusing on the sequence changes along with the functional changes, if any, in the regulatory regions genome-wide.

Chapter 2

Results

2.1 Accessibility is conserved across evolutionary time in *Drosophila* species

As discussed, there is a lot of sequence divergence in regulatory regions despite overall stable gene expression and robust developmental plan. To explore what property is conserved to maintain this function, we started with ATAC-seq since it gives us an overview of the regulatory landscape and grammar by mapping the accessible regions in the genome (Buenrostro et al. 2013). Additionally, ATAC-seq is a robust assay and highly adaptable across species unlike other methods like ChIP-seq. We have the ATAC-seq data of 4 species of *Drosophila*- *D. melanogaster*, *D. erecta*, *D. ananassae* and *D. mojavensis* at the same developmental stage. We are interested in the late embryo stage corresponding to 12-14hr in *D. melanogaster*; the other species were brought to the same developmental stage considering the their developmental time and rate.

Though the species in consideration belong to the same genus, they are quite evolutionarily divergent (Figure 2.1). To put their evolutionary distances into perspective, humans and gorillas diverged approximately 10 million years ago, whereas *D. melanogaster* and *D. mojavensis* diverged roughly 40 million years ago (they are even further apart in terms of substitution). We obtained the new, highly contiguous, curated genome assemblies of the other species from Kim et al. (2021). In order to align the whole genome of different species, we used a species alignment tool called Progressive Cactus (Armstrong et al.

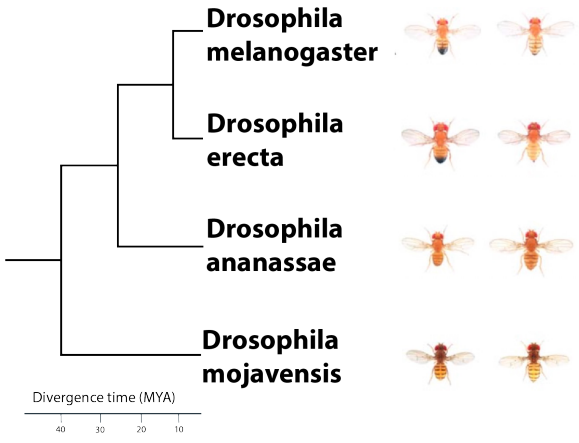

Figure 2.1: *Drosophila* species tree

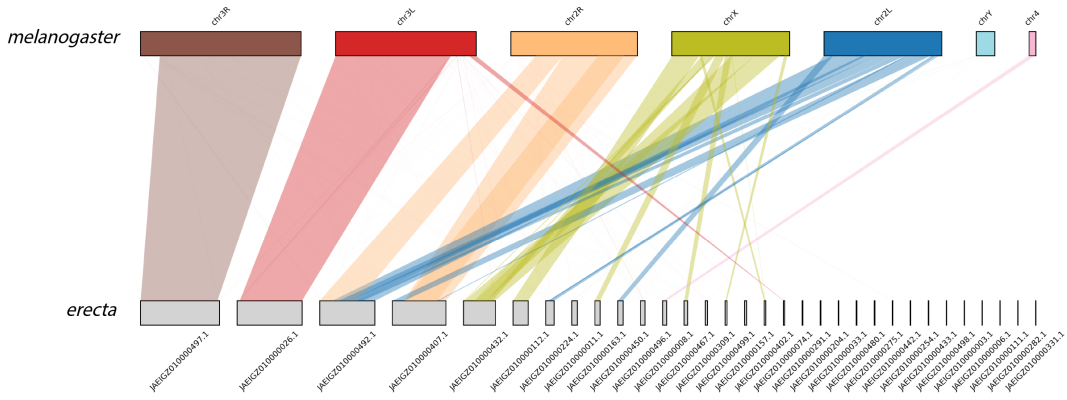

Figure 2.2: Syntenic regions between *D. melanogaster* and *D. erecta*

2020), a reference-free whole-genome alignment method. It constructs alignments using cactus graphs that explicitly model the evolutionary history of DNA sequences across multiple species (see Methods 4.5). Note that the genome assemblies of other species consist of multiple contigs and scaffolds, rather than fully resolved chromosomes. Figure 2.2 shows the syntenic regions throughout the genome for *D. melanogaster* and *D. erecta*. From the supplementary figure (Figure B.1), we can already see that the closest species have a lot of syntenic/orthologous regions and as we go across evolutionary time, in *D. melanogaster* vs *D. mojavensis*, the synteny is reduced and more fragmented.

Then we use a tool called LiftOver (Hinrichs 2006) on the Cactus-generated chain files to view the orthologous regions between any 2 species, based on our region of interest (see Methods 4.6). Figure 2.3 shows an example region (an ATAC peak) in *D. melanogaster* and its orthologous region in *D. erecta* and their respective accessibility profiles. The sequence is also shown below the figure as a heatmap. Even though there are more than 500 sequence mismatches in this 2kb region, their accessibility profiles are very similar.

Note that, if our region of interest is not in a syntenic block, or if there is no alignment possible as returned by Cactus, the region will not be lifted over to the other species. Hence, as we have seen in other papers (Phan et al. 2025), the percentage of liftover already gives us an estimate of conservation of the region across evolutionary timescale. Though, as we expect, the fraction of liftover regions ie., ATAC peaks of *D. melanogaster*, decreases with greater distance, the maintained fraction is quite high and does not drop drastically.

To test if these orthologous regions are also putative cis-regulatory regions, we took the ATAC peaks from *D. melanogaster* and analysed the orthologous liftover regions in the other species. Over 70% of these peaks are overlapping an ATAC peak that was called in the other species. When analysed quantitatively, the normalised counts correlation of the mapped ATAC peaks between two species was even higher, eg., it was 0.91

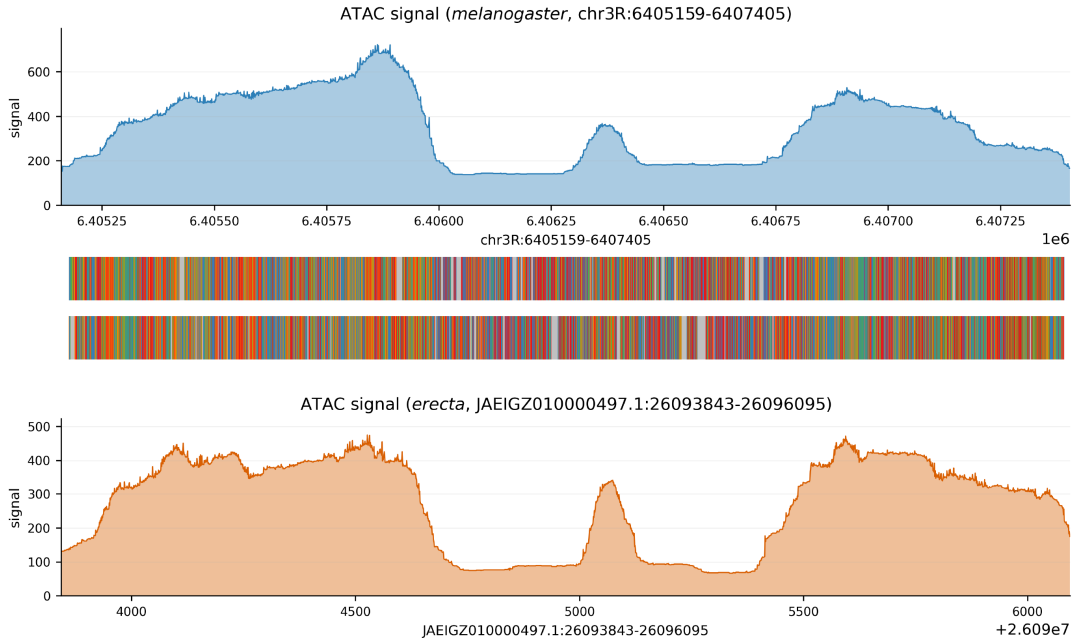

Figure 2.3: An example ATAC peak in *D. melanogaster* and its orthologous region in *D. erecta*

for *D. melanogaster* and *D. erecta*. This holds true for other species as well - the counts correlation remains high, despite the evolutionary distance and sequence divergence (Figure 2.4). This confirms that the regulatory landscape in terms of accessibility is largely conserved at this developmental stage in these *Drosophila* species.

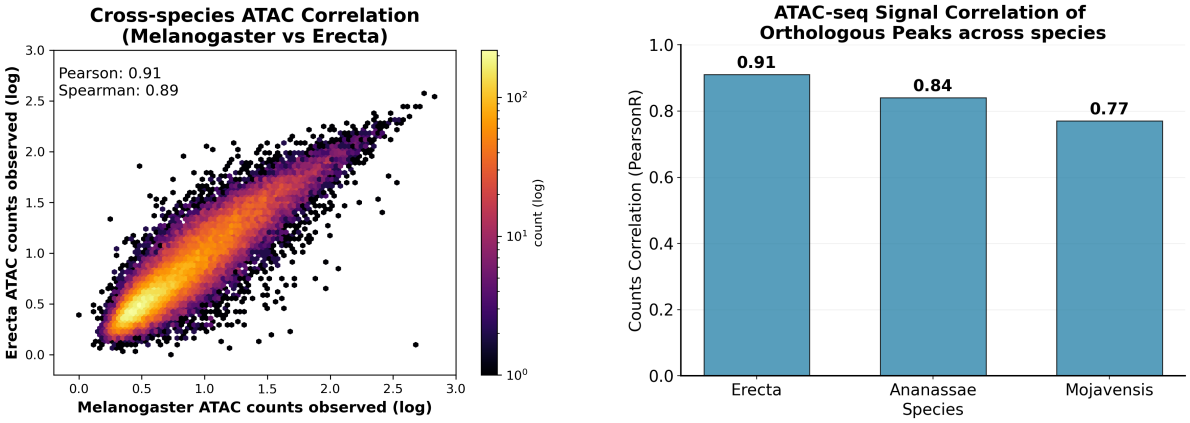

Figure 2.4: ATAC signal correlation across species. Left: Correlation between *D. melanogaster* and *D. erecta*. Right: Correlation across all species.

The slight reduction in correlation across evolutionary time hints at the changing regulatory landscape in these species. We shall now focus on the functional changes like TF motifs by analysing the regulatory rules using deep learning.

2.2 BPReveal accurately interprets the sequence rules of accessibility

Deep learning has emerged as a powerful paradigm in biology for decoding the regulatory language of the genome, effectively mapping raw DNA sequences to complex cellular functions. As described in the introduction (Section 1.3), previous works (Avsec et al. 2021a) and others (Pampari et al. 2024) have shown that deep learning models learn the sequence rules of many functional genomics data like binding (ChIP-nexus), accessibility, MNase-seq etc. And they not only accurately predict base resolution profiles but also identify the sequence's base contribution to function.

We have seen that the accessibility is conserved - but we don't know the mechanisms. We don't know if the trans regulatory environment is what's changing or the cis-regulatory architecture. Previous experimental studies on a select few loci have shown that the trans regulatory environment is largely conserved across species, and it is the cis-regulatory architecture that is changing. Deep learning models allow us to decipher the regulatory grammar and understand the sequence rules, genome-wide and not just limited to a few loci. Here, we use BPReveal (McAnany et al. 2025), a deep learning model to understand the sequence rules of accessibility.

ATAC-seq uses Tn5 transposase enzyme which cuts the DNA at open regions and inserts adapters which are then sequenced (Buenrostro et al. 2013). But this Tn5 transposase has a sequence preference, an enzymatic bias. To correct this bias, we first train a bias model which picks up this intrinsic sequence preference and not any underlying biology. We freeze this model and then we train a bias-reduced or a residual model, which learns the true biological signal that is not the enzymatic bias. Finally, we combine both the models to predict the final profile but we only use the bias reduced model for any meaningful biological interpretation (Figure 2.5, Methods 4.2).

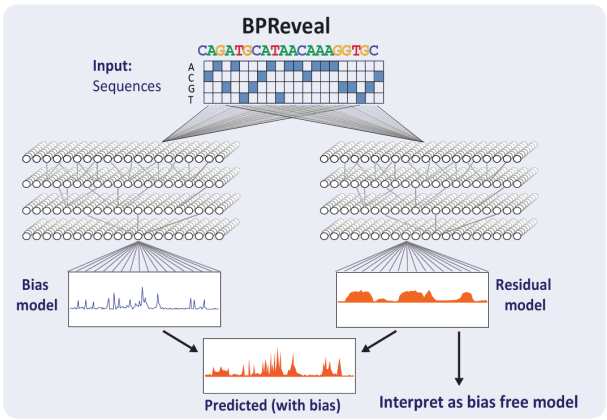

Figure 2.5: BPReveal model architecture

I trained a BPReveal model on the ATAC-seq data of *D. melanogaster* as detailed in Methods 4.2. Figure 2.6 (Top) shows a region of the genome showing observed and predicted ATAC signal. As we can see, the model does pretty well at predicting the ATAC signal. We measure the counts correlation across the observed and predicted tracks to quantify the performance of the model. The counts correlation for the test set, which the model has never seen, was 0.8 (PearsonR). Moreover, the genome-wide counts correlation (PearsonR) was 0.81 (2.6). This shows us that the model has learnt the sequence rules

and hence accurately predicts the accessibility data from sequence alone.

We don't just aim for high performing models, but we are interested in interpretability. To achieve this, we use Shapley values (referred to as SHAP scores or contribution scores from hereon). Originating from cooperative game theory, Shapley values provide a principled way to attribute a model's prediction to its input features. The method distributes the total prediction ("payout") among features by computing the average marginal contribution of each feature across all possible feature subsets (see Section 1.3.3). In our setting, the "features" correspond to the bases in DNA. SHAP scores quantify how much each base at each position contributes to the predicted binding signal. Thus, instead of treating the network as a black box, we obtain a base-level attribution map that highlights the important sequence features and interactions between them as learned by the model.

In this framework, contribution scores are defined relative to a reference input. A contribution is positive if the presence of a base at position i increases the predicted signal compared to the reference, and negative if it decreases it. Formally, the score quantifies the marginal effect of a specific base at position i on the predicted binding profile within a given genomic window. Thus, each attribution value reflects how strongly that base contributes to shaping the overall predicted profile.

We give the input sequences along with their SHAP scores to a motif-discovery tool called modisco-lite (Shrikumar et al. 2018) - which scans for high contribution seqlets and clusters similar ones together. These seqlets are for the majority motifs. These seqlets can be clustered into a traditonal PWM or PFM, and more importantly a CWM - contribution weight matrix, which tells us the contribution of each base in the motif, and overall contribution of the motif itself.

The motifs identified by the BPReveal model are shown in Figure 2.6. They include pioneer transcription factors - *GAF*(GAGA Factor), *GATA* and *Grh*(Grainyhead), which are very important for creating chromatin accessibility. The motifs also include promoter-associated motifs like *M1BP*, *Ohler 5* etc, insulators like *SuHW*, *CTCF* and repressors - *ttk* and *slbo*. Additionally, the model also found new motifs which may be putative binding sites of unknown factors.

After successfully training a model on the ATAC-seq data of *D. melanogaster*, I trained additional models on each of the ATAC-seq datasets of the other 3 species, keeping the architecture the same, as described in Methods 4.2. These models also perform well with a genome wide counts correlation of 0.72, 0.73 and 0.84 (PearsonR) on *D. erecta*, *D. ananassae* and *D. mojavensis* respectively. The motifs identified by these different models are largely the same as the ones found in *D. melanogaster* in Figure B.2. The CWM logos are also identical (Figure B.3). This suggests that the transcription factors active at this developmental stage are conserved at the genome-wide level.

Taken together, this lets us conclude that the trans-regulatory environment is con-

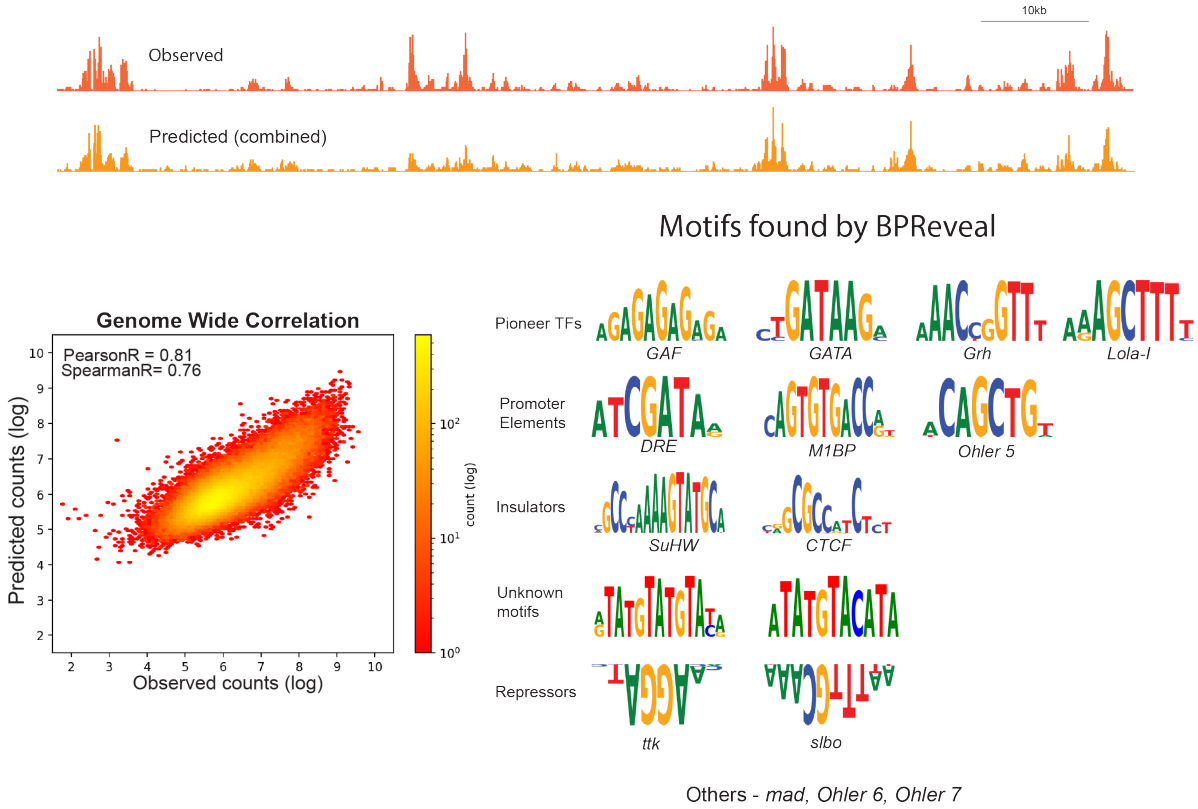

Figure 2.6: BPReveal model performance and motifs in *D. melanogaster*. Top: Example region showing observed and predicted signals. Bottom left: Genome-wide correlation. Bottom right: Motifs discovered by BPReveal.

served across these species, genome wide.

2.3 Cross species prediction reveals the conservation of cis regulatory code across evolution

We have different models trained on ATAC data of these different species and they largely discover the same underlying motifs, and hence we have established that the trans environment is the same. This means that the cis-regulatory architecture is what's changing across species. But how does it change, keeping accessibility the same?

This is where the advantage of deep learning models comes in - and we take a genome wide approach to answer this question. Now that we have the models and the sequences and their data, I wanted to see if there are any generalizable trends/patterns across species.

We have a model which has learnt sequence rules of accessibility in *D. melanogaster*. Using this model, we predict the accessibility profile on the other species, to see how well it performs. Figure 2.7 (top) depicts a region showing tracks of observed data in erecta, and the two predictions - one from a model trained on *D. erecta* data and one from the model trained on *D. melanogaster* data. Though the predicted profile of the

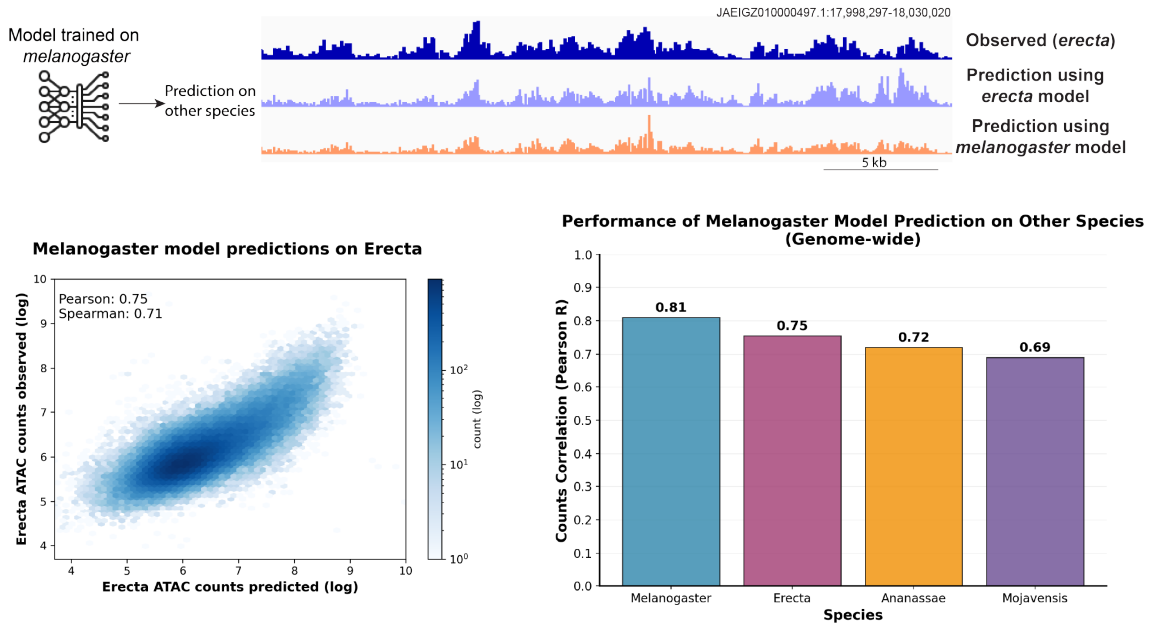

Figure 2.7: Cross Species Prediction. Top: Example region showing observed and predicted signals on *D. erecta* data. Bottom left: Genome wide correlation of observed vs predicted accessibility using the *D. melanogaster* model. Bottom right: Correlation across all species.

D. melanogaster model looks a bit worse than the *D. erecta* model, it still looks good compared to the observed profile. (This is a handpicked example showing the difference between the predicted profiles of these models; when we look genome-wide, one can hardly make out the difference between the two profiles.) Then I calculated the genome wide counts correlation of observed vs predicted counts (predictions using the *D. melanogaster* model). This yielded a genome-wide correlation of 0.75 (PearsonR) for *D. erecta* as well as fairly good correlations across the other species as shown in Figure 2.7.

So, we can conclude that the majority of the sequence rules seem to be conserved across these species. Also, there is a slight reduction in correlation that reflects the evolutionary distance between these species. This is more evidently seen when we see the performance of the different species models on the other species, as shown in Figure B.4.

Though this reflects the evolutionary distance, it might mean several things: changes in syntax, or species specific rules, or it may even be an artefact of the model which has not had enough instances to learn from. But overall, the high performance indicates that the sequence rules governing accessibility are largely conserved across species.

2.4 Leveraging multi-species data improves model performance

We have seen that these models learn the sequence rules of accessibility accurately, but we have also seen in the previous section that the model performance gets slightly worse over evolutionary time. Since the majority of the sequence rules are conserved but multiple species provide more sequence variation, we wondered whether we could leverage this increased variation and the increase in the amount of data to train a better model. For example, the model might learn additional sequence rules, for which there were not enough examples in each single species data. Furthermore, more data alone could improve the performance. If so, we would like to know whether the improved performance also helps the model interpretation.

The multi-species model was built using two different paradigms. For the first model, all species data were normalized and pooled together. The second model is made species aware, by altering the structure of the model such that 4 channels of information are added at the level of one-hot encoding, which each represent the species (Figure 2.8). This means that the input to the model contains information of which species the particular sequence and data is coming from. The same model training structure as described in Methods 4.7 was followed for both models. I will refer to the first model as the *species-pooled* and the second as the *species-aware* model.

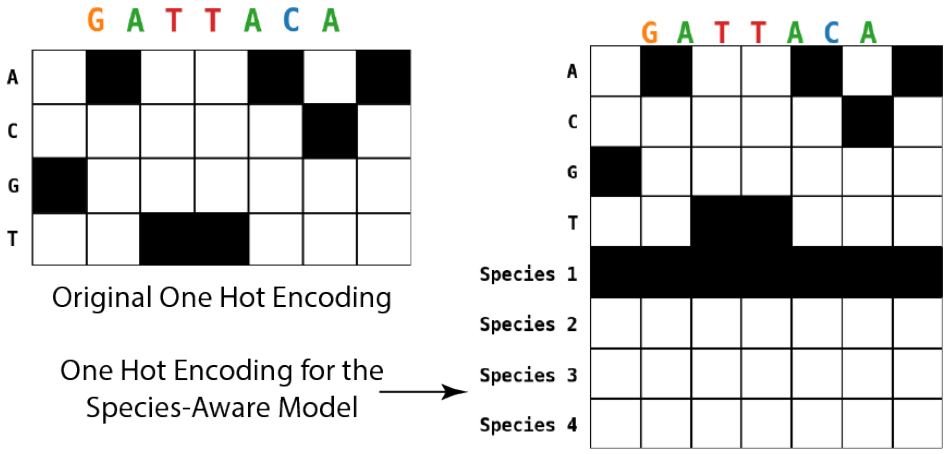

Figure 2.8: One Hot Encoding for the species aware model.

Each model - *species-pooled* and *species-aware*, were trained in 3 different ways (as described in Methods 4.7). The first method did not consider the orthologous regions and used random splits of input sequences in each species. This can be a cause of data leakage since the model can learn the signal from the region in a species and then predict accurately in the orthologous region of other species. Therefore, the second method was implemented as follows: The train, test and validation regions of *D. melanogaster* were

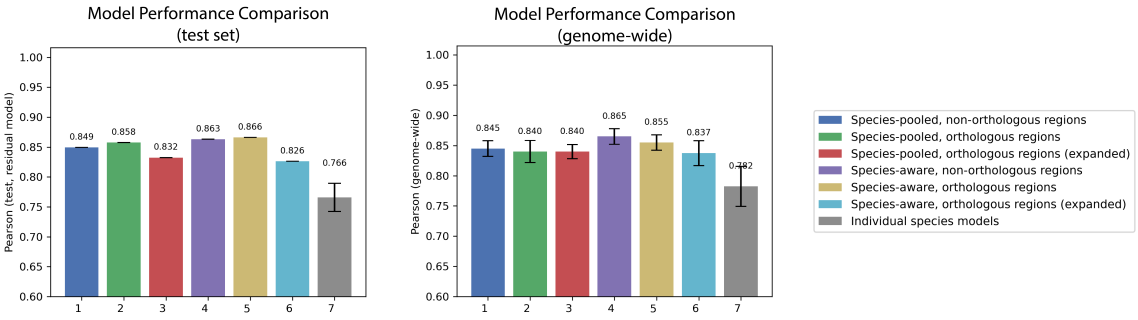

Figure 2.9: Comparison of multi species model performance - Pearson correlation.

first selected, and then pooled together with the orthologous regions of the other 3 species. This way, we can be sure that the model has never seen the test sequence or any instance of the test sequences' orthologous regions. This has a slight drawback that we are not considering the species specific accessible regions. So, I also use a 3rd method - where I add the species-specific peaks to only the test and validation sets to accurately test the performance on these regions.

I then analyzed the performance of the models using the counts correlation between observed and predicted data of the test sets or genome wide. This revealed that, regardless of the model or training paradigm, the multi-species models (both species-pooled and species-aware) perform better than the single species models (Figure 2.9, B.8). This could be attributed to having more data and the natural variation in sequences having the same function, which allows the model to learn better. The performance increase is not huge, which means that the cis-regulatory syntax in the species-specific regions is not that different from the genome-wide cis-regulatory code, indicating the conserved cis-regulatory code that we have already seen earlier.

We can see that the performance of the species-pooled model (genome-wide) is largely the same across the 3 different training methods. But the species-aware model performs better, if not equal to the species-pooled model. We do see that the first method has the highest performance, and it is especially higher in the species aware model. This can not be totally attributed to data leakage between species since the performance of the second method is fairly comparable to the first method. The decrease in performance in the third method, more apparent in the performance on the test set, does show that there might be some species-specific rules. For example, transposable elements are often species-specific.

At the interpretation level, the motifs found by the species-pooled model were the same as those found by the single species model. But the clusters and CWMs were cleaner and well defined. As we expected from the previous results, there are no new motifs found, but there is an increase in performance. (The interpretation of the species-aware model has not been done yet, but it will be interesting to see the species-specific contribution scores for the same regions). This shows that multi-species information can

be utilized not only to understand the evolution of regulatory regions, but also to improve the performance of the model so that the interpretability is better and more robust.

2.5 Modeling neutral DNA evolution demonstrates the effect of sequence divergence on accessibility

We have seen from previous sections that the accessibility is maintained despite the sequence divergence, but also that the sequence rules of accessibility are largely conserved across species. We know that the sequence changes - mutations are not randomly passed on to further generations, but are subject to natural selection. If we consider the ATAC peaks, the levels of accessibility are remarkably conserved. But is this because of selection? To answer this, we ask the question - if we were to have a similar sequence divergence between 2 species as seen in the observed data, but without any selection, could the mutated new sequence still have the same level of function as measured by the model predictions? This would argue for neutral evolution, rather than selection. To make this framework rigorous and robust, we use simulations to synthetically model DNA evolution. As introduced in Chapter 1, we don't want to mutate the sequences randomly to see its effect. We would like to incorporate some biological constraints that sequences undergo, into our simulation; and that is where we turn to DNA models of evolution. These are phenomenological models and do not hint at any mechanism of evolution, but what they give us is the relative rates of different mutations that we actually see in the observed data. These models were originally built to correct the non-linear relationship between observed sequence differences and the actual time elapsed since divergence. By modeling this mathematically, these frameworks were originally designed to recover the "true" evolutionary distance—the expected number of substitutions per site—thereby restoring the linearity required for molecular clock analyses and the construction of accurate phylogenetic trees.

We shall now use these models to introduce mutations in the DNA sequences to assess to what extent the sequence diverges under no selection, as we will see in a while. And for this, we use the HKY85 model discussed in Section 1.2.2, with some modifications of our own. But before that, we need to get the different parameters needed for the simulation such as - branch length or evolutionary distance, rate of transversions to transitions, equilibrium frequencies of the bases, etc.

The method of simulation and parameter estimation has been explained in detail in Methods 4.8. I will briefly state them here (Figure 2.10). I took the ATAC peaks in *D. melanogaster* which had a valid orthologous region in all the 4 species, extracted the orthologous sequences, and used MAFFT (Kuraku et al. 2013) to align these regions. Then I used IQ-TREE (Wong et al. 2025) to build a phylogenetic tree with the HKY

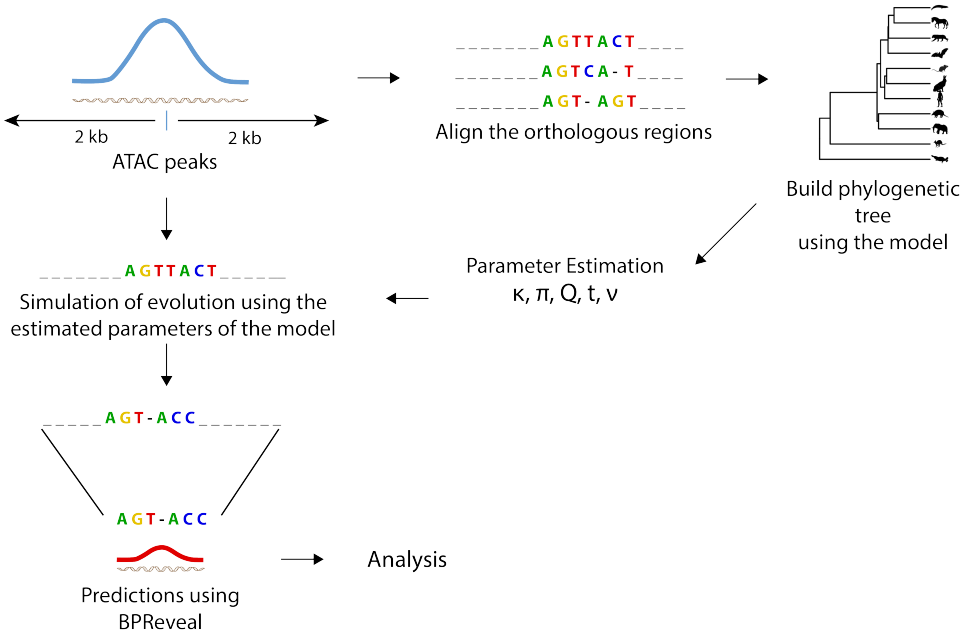

Figure 2.10: Workflow of synthetic evolution simulation

model on our orthologous regions to get the model parameters.

I am using a modified method to simulate evolution by including indels (insertion-deletions) onto the already existing rates based on the HKY substitution matrix. Hence, the model parameters required for simulation were obtained from the already executed IQ-TREE model and from the alignment data. We use the following considerations and constraints in terms of model parameters - evolutionary distance in terms of branch length, ratio of rates of transversion to transition, indel rate, insertion probability & distribution of insertion and deletion length. (Refer to Methods 4.8, Table 4.2)

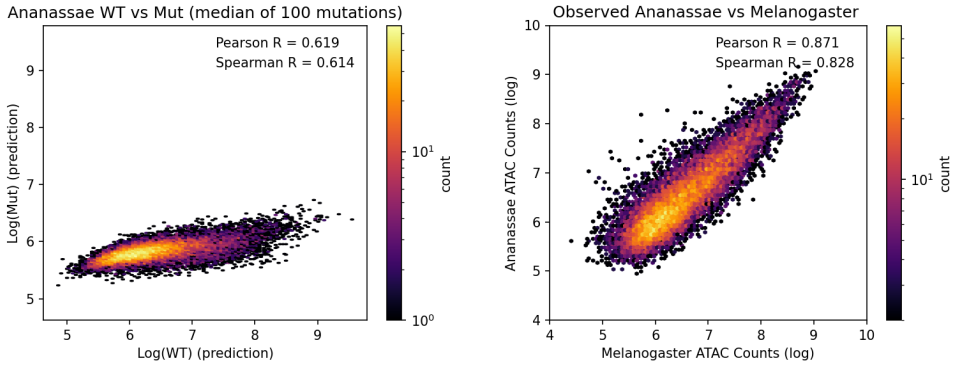

Figure 2.11: Correlation of predictions on synthetically evolved sequences and correlation of observed data for *D. ananassae* w.r.t *D. melanogaster*.

With the above set of parameters, one for each species with respect to *D. melanogaster*, I ran the synthetic evolution simulation 100 times per ATAC peak in consideration. In this manner, we have 100 variations of the original ATAC region for each of the species at the appropriate level of sequence divergence. These mutated sequences were for predic-

tions by our original BPReveal model to assess the divergence in the level of accessibility.

We then computed the counts correlation between the predictions on the wildtype sequences (the original *D. melanogaster* ATAC peaks we started with) and the predictions on the mutated sequences. (Here I considered the median over the 100 runs per peak. If we consider all 100, the correlation is very poor). This shows that the accessibility levels of the synthetically evolved sequences are overall much lower compared to those of the observed data, e.g. see those for *D. ananassae* in 2.11, suggesting strong loss-of-function effects. For *D. erecta*, the counts correlation between synthetic and observed sequences is still similar, which can be attributed to the fact that *D. erecta* is very close to *D. melanogaster*. But for species of larger evolutionary distances, the correlations are drastically reduced, with 0.62 for *D. ananassae* and 0.34 for *D. mojavensis*; when we compare with the correlation of the observed data as seen in Figure 2.12 & Figure B.5. (The same thing was also observed for fold change). This suggests that these regulatory regions are under high selective pressure to maintain the accessibility of these regions at appropriate levels.

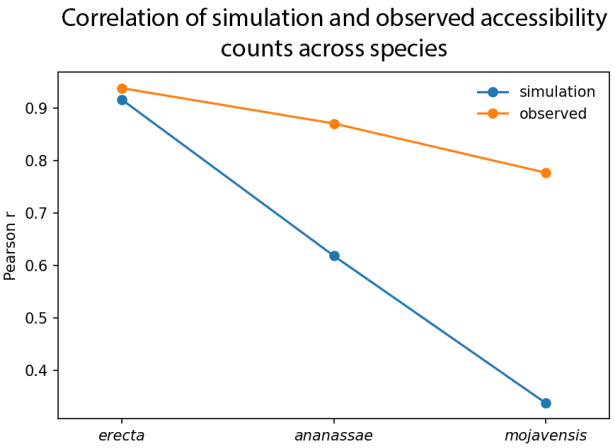

Figure 2.12: Correlation across species

Does every region take an equal amount of hits? Is every ATAC peak equally vulnerable to mutations? To answer this, I calculated the log fold change of the predictions– i.e., the log of predicted signals of mutated and wildtype sequences. Figure 2.13 (top) shows the variation of this predicted log fold change across the observed counts of the region. Figure 2.13 (bottom) shows the same for the observed log fold change. We can see that the predicted fold change of highly accessible regions is low, and this reduction is drastic when we go across evolutionary timescale. However, when we see the observed fold change, even though there is a lot more variation when we go to greater evolutionary distances, the accessibility is still maintained. Also, there is a considerable fraction of regions with a positive log fold change, which is not seen in the simulations.

The above data tells us that these regulatory regions are under a considerable level of selective pressure, and that highly accessible regions are probably under an even higher level of selection.

We know that the accessibility of such regulatory regions are mediated by transcription factor motifs. We shall therefore next analyze the conservation of motifs and their turnover to better explain how evolution maintains regulatory regions at similar levels of accessibility.

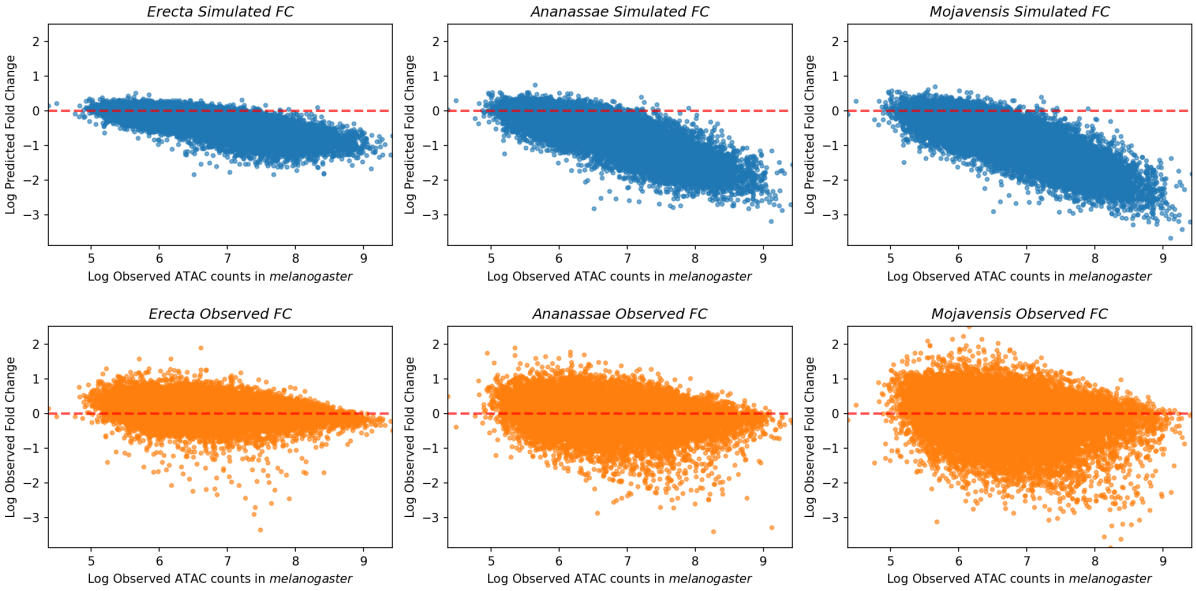

Figure 2.13: Predicted and observed log fold change across species

2.6 Motif composition is maintained at highly accessible regions despite a lot of motif turnover

We know that any function, here being accessibility, is mediated by motifs, from previous results and previous papers. But if the sequence is diverging a lot, do the motifs remain conserved? Do they also turnover?

First of all, we see that these motifs (motif instances retrieved from the model) are indeed more conserved than random genomic regions indicating some selective pressure (Figure B.6). Also, the motif instances identified by the model are relatively more conserved than the ones found by PWM scanning. This is because the model identifies the motifs contributing to function (accessibility) and doesn't just look at the sequence match. Hence, these motifs, which are contributing to function should be more conserved, which is what we see in Figure B.6. We use this list of model derived motif instances for further analysis.

The curated mapped motif instances from the BPReveal model (Methods 4.4) was lifted over to other species to compare the sequence changes in the list of motif instances. The sequence changes were categorised in terms of number of mismatches in the alignment of the 2 motif regions - which includes substitution or gaps. Note that these comparisons can only be made for the motifs which are lifted over to the other species (have a valid orthologous region). A few times, the motif region may be mapped onto a huge region in another species, in which case the alignment will give us a lot of mismatches. Figure 2.14 shows what percentage of motifs have sequence mismatches. We can see that over 50% of motifs are lost - either not lifted over or have more than 2-3 mismatches. These 2-3 changes can destroy the motif since a normal motif is around 7-10bp. This shows that

motifs also undergo a lot of mutations.

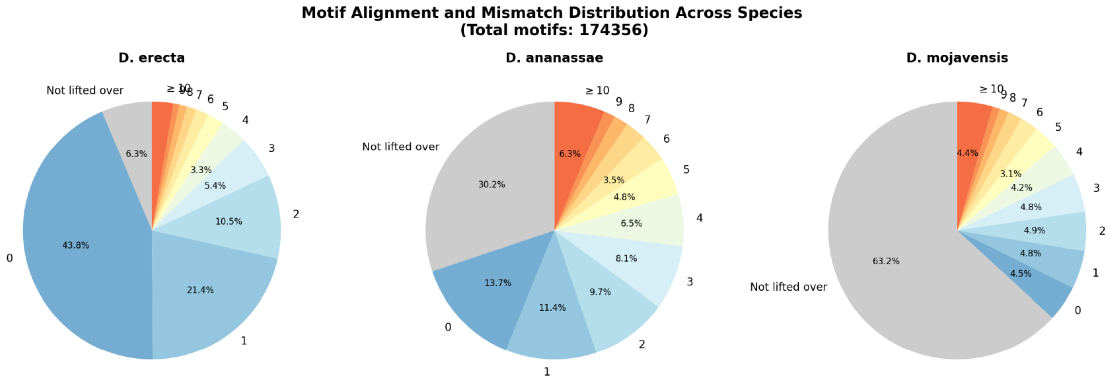

Figure 2.14: Motif turnover across species. Percentage of motifs with sequence mismatches across species.

But we don't know if the orthologous regions of motifs are functional or contributing to accessibility. It may retain the same sequence, but not be functional because of many reasons. The opposite can also happen - the motif can have a few sequence changes, but still function as motifs. This is possible due to the flexibility of the motif. This has been shown before - even if the PWM score of the sequence is high, there need not be any specific binding of the TFs, and even if the PWM score is low, there can be binding of the TFs (Weilert et al. 2025). So, I checked if these orthologous regions were called as motifs in the respective species by the model. If they are, this would say that the particular motif region is functional and is contributing to accessibility in both species. We see that only 20% of the motif instances are actually called as motifs (for *D. erecta*). So, these motifs are undergoing a lot of turnover. But then, how is the accessibility maintained?

We will now zoom out and focus on the motif composition at the orthologous regions. From the previous section, we know that the highly accessible regions are under a higher level of selection. In those regions, are the motifs more conserved? Both in terms of sequence and in terms of function? To answer this, we look at a concept called Jaccard similarity which is a measure of similarity between 2 sets. Here, we look at the list of TF motifs in a region in *D. melanogaster* and the list of TF motifs in the orthologous region in *D. erecta* or other species. Then we calculate the Jaccard similarity between these 2 sets - which is nothing but the number of motifs found in both species divided by the total number of motifs found in either of the species at the respective orthologous regions (Figure 2.15 & Methods 4.9).

We can see the distribution of Jaccard similarity across the orthologous ATAC peaks of *D. melanogaster* in the other 3 species (Figure 2.16). We see that the distribution is moving towards 0 as we go across evolutionary time. The density of peaks with a Jaccard similarity of 1 is reducing across species and the density at 0 is increasing. This shows that across the same regions, the motif composition is changing across species, and this

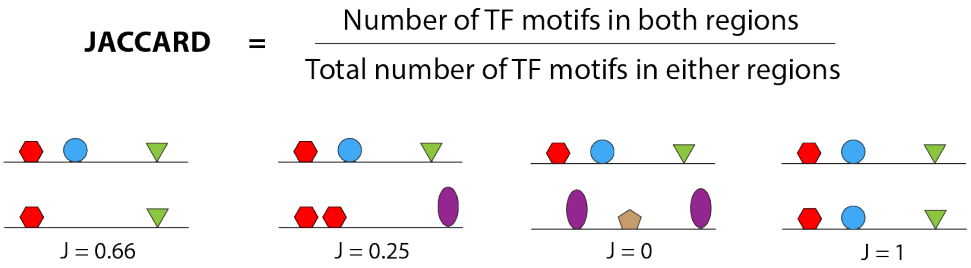

Figure 2.15: Jaccard similarity

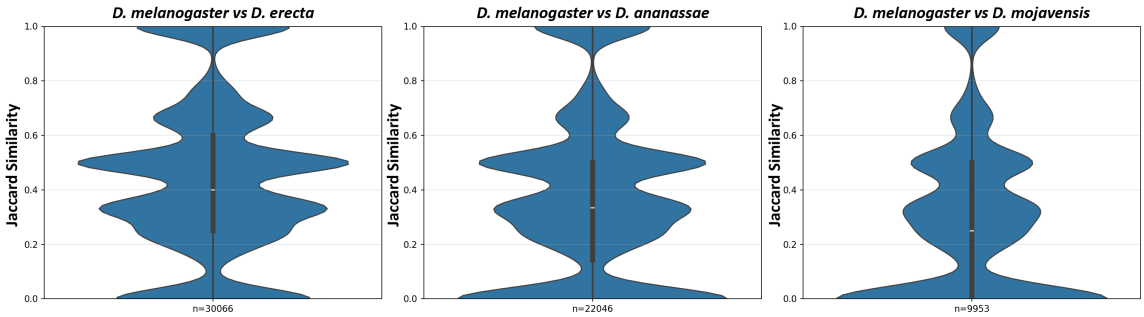

Figure 2.16: Distribution of Jaccard similarity across species.

change is more drastic as we go across evolutionary time.

One thing to note here is that - Jaccard similarity sees only the different motifs, but it does not take into account the number of instances of the same motif. It may so happen that one motif is lost but there are 2 instances of another motif, which can compensate for the loss of the first motif. This analysis does not take these cases into account. Regardless, this is still a good metric to analyse the motif conservation at orthologous regions across species, and has been used in previous studies too (Khoueiry et al. 2017).

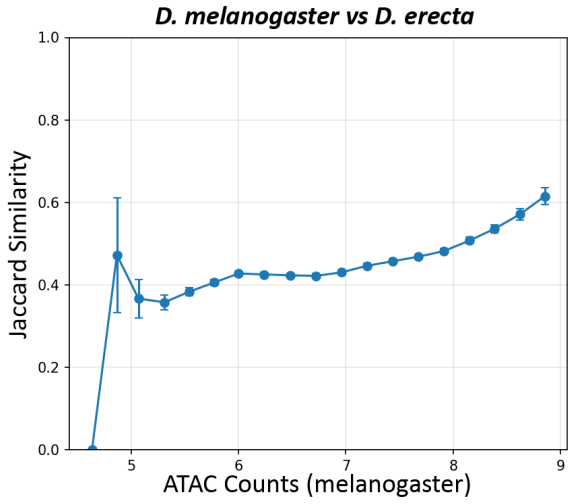

Figure 2.17: Jaccard similarity vs ATAC counts/signal

Figure 2.17 shows the Jaccard similarity of the peaks (w.r.t *D. erecta*) versus the observed ATAC counts of the region in *D. melanogaster*. We can see that the jaccard similarity increases with the accessibility of the region. This holds true for other species also as shown in Figure B.7 (it holds even if we include the other species peaks, not shown). This also explains the previous result as to why the highly accessible regions lose accessibility across evolutionary time when we synthetically mutate the sequence. These regions have a lot of motifs, and their motif composition is more maintained across species

than other regions. So, there is a higher level of selection acting on these regions - not only to maintain accessibility but also to maintain the motif composition at the sequence level.

2.7 Low affinity motifs show syntax-dependent conservation

As we saw in the previous section, there is a lot of motif turnover. Because a motif's PWM score directly correlates with empirical in vitro binding measurements derived from protein-binding microarrays, it serves as a proxy to quantify a sequence's intrinsic binding affinity. Because these sequences exhibit weak TF binding strength in vitro and poor PWM matches, they have historically been difficult to map and study without deep learning methods. Functionally, low-affinity motifs demand higher local TF concentrations to become fully occupied, allowing enhancers to precisely sense and respond to shifting TF dosage thresholds. However, the regulatory impact of a low affinity motif is not dictated by its binding strength alone; rather, deep learning models have revealed that their function relies heavily on their surrounding genomic sequence context. Specifically, these weak motifs achieve outsized functional effects on chromatin accessibility by cooperating with nearby, stronger pioneer TF motifs within short distances, utilizing a flexible soft motif syntax to cooperatively fine-tune enhancer activity. Here, I want to look closely at the conservation of these low affinity motifs (LAMs).

I would like to focus on a specific transcription factor motif for this analysis - *Grh*, a pioneer transcription factor. It is discovered in all the species - as shown before (Figure B.2), and though the information content of the motif is high, it allows for some flexibility in the sequence changes of the motif as we can see in the motif PWM logo. Since I want to compare the motif across species, I am considering *D. melanogaster* and *D. erecta*, and all further analysis will be with respect to these two species.

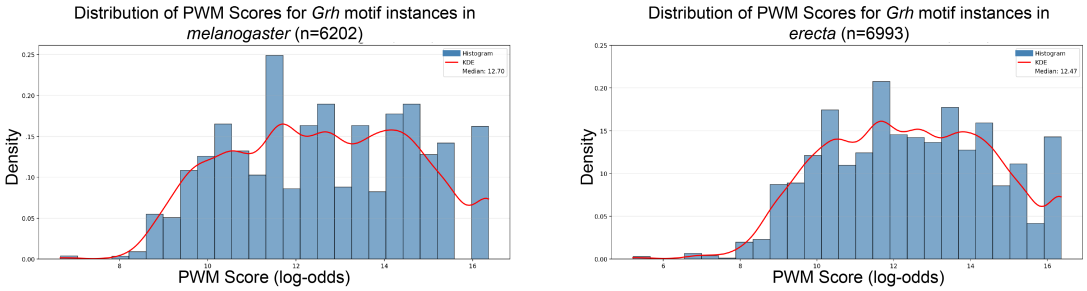

Figure 2.18: *Grh* PWM distribution across species.

I took all the mapped *Grh* motif instances in *D. melanogaster* and lifted them over to *D. erecta*, and vice versa as well. We can see the PWM distribution of all the mapped *Grh* motif instances, in the 2 species, which are similar (Figure 2.18).

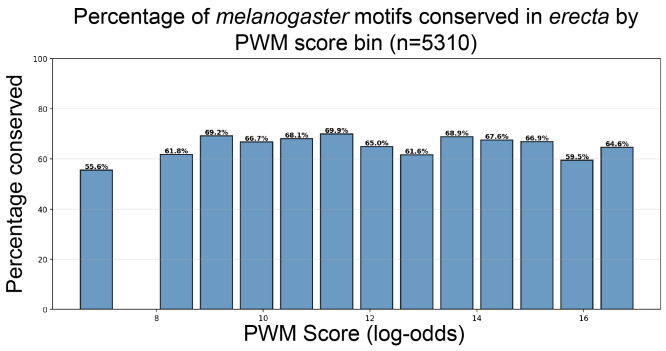

Figure 2.19: Percentage of conserved *Grh* motif instances across PWM score bins.

To see the extent of motif turnover, I categorised the *D. melanogaster* motif instances (which had a successful liftover) into 2 categories: "Conserved", if it has the same sequence in *D. erecta* as in *D. melanogaster*, and "Turnover" if it doesn't. Around 63% of the motif instances fall into the Conserved category. For the Turnover category, the distribution goes down and has a greater spread while Conserved category has the same PWM distribution since the sequence is maintained (Figure B.9). We see that the PWM distribution of the 2 categories in *D. melanogaster* is very similar. To make this precise - to see if conservation is dependent on affinity, I calculated the percentage of motif instances falling in the Conserved category for each PWM score bin (Figure 2.19). This shows that the percentage of Conserved motif instances in each PWM score bin is largely the same - which tells us that the sequence conservation of a motif is largely independent of its affinity.

But we still cannot see how each motif is turning over. To see the instance-specific turnover, I categorised the motif instances based on the PWM score into 5 bins based on the quartiles of the PWM distribution of mapped *Grh* instances (Table 4.3 in Methods 4.10). Now it comes to the question of - what is the motif turning into, in the other species? Since we have already defined the categories of the motifs, we can calculate the frequency of transitions between these categories. Here is the transition probability matrix for the motif instances in the Turnover category (Figure 2.20). This is normalised by the row and it answers the question - Given that a motif is a certain affinity 1 in *D. melanogaster*, what is the probability that is affinity 2 in *D. erecta*? (Affinity 1 and 2 being variable here)

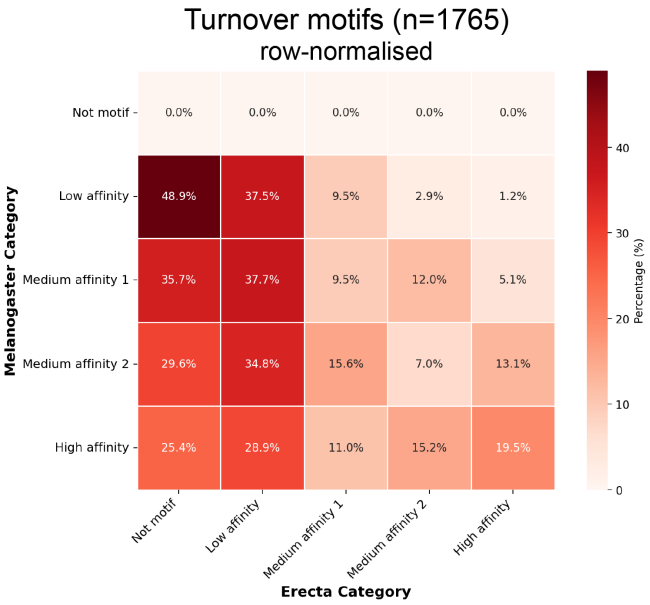

Figure 2.20: PWM score transitions for *Grh* instances in the Turnover category.

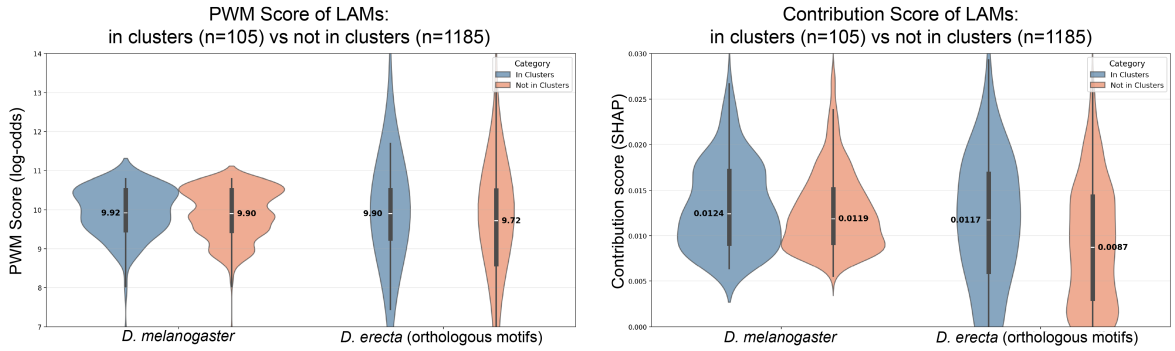

Figure 2.21: PWM and contribution score distribution of LAMs in *D. melanogaster* and its orthologous regions of *D. erecta* in clusters and non clusters.

We can see that there is a lot of motif turnover involving the categories of low affinity motifs or non-motifs in *D. erecta*. But towards the right side, we see that there is a high probability of being high affinity in *D. erecta*, given that it is high affinity in *D. melanogaster*. This shows that high affinity motifs

It has been shown that LAMs cooperate with pioneer motifs to increase accessibility of that region. We ask if the conservation of LAMs is dependent on its neighbouring motifs. I consider all LAMs in *D. melanogaster*, and filter them based on whether there is a high affinity pioneer motif in its neighbouring region or not ($\pm$ 300bp). If they do have a high affinity pioneer motif nearby, I put them into the "LAM cluster" category (Methods 4.10). When we see the PWM score distribution of the LAMs in these 2 categories and their orthologous motifs in *D. erecta*, we don't see a huge difference (Figure 2.21 left). And we see that 78% of LAMs in clusters fall in the Conserved category whereas 66% of LAMs in non clusters fall in the Conserved category. This already hints that LAMs in clusters are relatively more conserved.

When we look at the contribution scores of these motif instances in the respective categories, the syntax dependent conservation becomes apparent. LAMs in clusters largely retain their contribution score in the other species, but for the LAMs in non-clusters, the contribution score in *D. erecta* goes down (Figure 2.21 right). Furthermore, we observe that this is the case if there is any pioneer motif in the neighbouring region of the LAM, regardless of its affinity. Therefore, there is a significant change in the contribution score rather than in the PWM score, which tells us that functional conservation of motifs is dependent on its surrounding motifs (probably due to cooperativity).

Chapter 3

Discussion

Here we explored how chromatin accessibility evolves across species to maintain an important developmental embryonic stage. Using ATAC-seq data, we found that chromatin accessibility is very well conserved across *Drosophila* species. Since we know that the regulatory rules of chromatin accessibility is driven by transcription factors and their combinatorial syntax like spacing, orientation etc, we used deep learning models to elucidate the regulatory rules. Firstly, this gave us the specific TF motifs contributing to accessibility (like pioneers, insulators, repressors etc). Different models further revealed that the transcription factors at play during this developmental stage are largely conserved across species and hence, the trans environment is conserved. It would be interesting to compare the expression levels of these TFs across species, to see if at all they have varying levels of expression, and to connect it to the contribution scores in each respective model.

Our results also show that the majority of the cis-regulatory rules are conserved across species, but there are also hints at some changes in sequence rules at species-specific accessible regions. On top of the general cis-regulatory rules, there might be slight changes in syntax or species-specific transposable elements that are more frequent in a species and hence, contribute to a small reduction in performance (not only the *D. melanogaster* model, but also the multi-species model)

Our multi-species model showed that leveraging data from multiple species can help improve model performance due to the larger dataset and the diversity of sequences, wherein the model has more instances and variants to learn from. This can also be used to design some elegant in-silico experiments such as contribution of a certain sequence motif to accessibility in a certain species, cooperativity of TFs across different species etc. Furthermore, the fully sequenced genomes of *Drosophila* species provides us a massive dataset - we can use these models to predict the regulatory regions in other species too where we don't have the data. This has been shown to work above, across diverse evolutionary timescale, and it will be really helpful to narrow down the sequence specific changes that are linked to any regulatory changes in different species.

Combining DNA models of evolution along with deep learning models showed that the accessible regions which are conserved, are indeed under a strong selective pressure. They also show a variable level of selection across different regions, the highly accessible regions being under the highest pressure. The above analysis not only quantifies the level of selection, but also provides some hints as to how much of a change in the accessibility can be tolerated by the organism until a mutation is weeded out of the population by natural selection. This lead us to investigate the motif level changes across species - which showed that there is a lot of motif turnover. So, we zoomed out and analysed the motif composition in these accessible regions. The motif composition across the same orthologous regions showed a greater difference in more distant species. Also, the highly accessible regions have a higher similarity of motif composition across orthologous regions, explaining the variation in the level of selection. This suggests that at such highly accessible regions, selection is also acting at the motif level in the sequence, and not only at the accessibility level. This is probably due to the fact that there are not too many ways (in terms of motifs) and there is not too much flexibility to create such a high level of accessibility in the region. This is interesting since it might shed light on how evolution works at the sequence level.

And focusing on a specific TF - *Grh*, we saw the syntax dependent conservation of low affinity motifs. The conservation of motifs is largely independent of its affinity, though there was a decent percent of high affinity motifs staying at high affinity despite sequence changes. The low affinity motifs which are near pioneer motifs had a higher conservation of their function (in terms of contribution score) when compared to the ones which were away from pioneer motifs. Further analysis needs to be done to understand the flexibility of the motifs in the region - to gain insights into compensatory changes and the role of cooperatvity of such TFs. The level of sequence changes at motifs was really surprising. It would be really interesting to find some differentially expressed genes and focus on the sequence changes in their regulatory regions.

The above study has been done using bulk ATAC-seq data in different species which does not reflect the cell-type level heterogeneity of the *Drosophila* embryo. Though different TFs and different enhancers maybe active in different cell types, this allows us to infer the general regulatory regions genome-wide in the whole embryo at that stage. Also, bulk ATAC has the obvious advantage of cleaner data, having a much higher signal-to-noise ratio and a deeper coverage of the genome - which makes it very suitable to study evolution.

Overall, this study shows that using deep learning models gives us a strong advantage to study evolution of regulatory regions - by analysing motif turnover, composition, synthetic evolution etc. which was not possible by earlier primitive methods of comparative genomics.

Chapter 4

Methods

4.1 ATAC-seq data processing

ATAC-seq data was generated by Gerardo Mendoza. The developmental rate of each species was taken into account to bring them to the same developmental stage as the late embryo stage in *D. melanogaster* (12-14h). The genome assemblies for other species were obtained from Kim et al. (2021).

Table 4.1: Genome assemblies of the 4 *Drosophila* species

Species	**Genome Assembly**
D. melanogaster	dm6
D. erecta	GCA_018904525.1
D. ananassae	GCA_018148915.1
D. mojavensis	GCF_018153725.1

Reads were pre trimmed for adapters using Cutadapt (v.4.2) (Martin 2011) and aligned to the following genome assemblies using bowtie2 (v.2.3.5.1) (Langmead and Salzberg 2012). Duplicated alignments were marked using Picard (v.2.23.8) (Broad Institute 2019) and deduplicated, filtered for fragment lengths corrected for dovetailed alignment pairs and end adjusted to accommodate the Tn5 enzymatic cut correction. Normalised RPM ATAC-Seq tracks were generated by scaling the aligned coverage to the weighted total reads. Cut site ATAC seq coverage used for training BPReveal models were generated by isolating the bases on each end of an aligned fragment, consolidating those cut sites into coverage tracks. ATAC peaks were mapped using MACS2 (v.2.2.7.1) (Zhang et al. 2008) with default settings. To assess reproducibility between biological replicates, we calculated the genome-wide Pearson correlation coefficient between the two replicate bigWig signal tracks. The resulting correlation coefficients were 0.88, 0.96, 0.83, 0.98 for the 4 species respectively, indicating high concordance between replicates. Given this strong agreement, replicate 1 was selected as the representative dataset for all subsequent downstream analyses and model training.

4.2 BPReveal model training

BPNet (Avsec et al. 2021a) is a convolutional network designed to learn genomic data from input DNA sequences alone. BPReveal (McAnany et al. 2025) is a suite that reimplements and extends the original BPNet and ChromBPNet (Pampari et al. 2024) architectures to be more versatile and interpretable. While retaining the core convolutional neural network structure for base-resolution prediction, BPReveal generalizes the framework to support a wider array of genomic data types beyond the original ChIP-nexus focus, including ATAC-seq, ChIP-seq, and MNase-seq. It also incorporates advanced bias correction mechanisms similar to ChromBPNet but with enhanced capabilities for regressing out complex experimental biases, and includes technical improvements like adaptive counts loss to stabilize training across diverse datasets. We use BPReveal (version 5.2.0) in all of the model training, predictions, interpretation etc.

Model Training for ATAC data in D. melanogaster
Since Tn5 enzyme has an enzymatic bias on where it cuts the DNA, we need to correct for it. For this bias correction, we first train a small "bias" model on non-peak or low-count regions, to pick up this intrinsic sequence preference of Tn5. Then, using this bias model, we train a transformation model on the peak regions - so that this transformation network learns to correct the ATAC data into bias corrected or bias reduced signal. Then we train the residual model, a larger CNN, on the peak regions to explain the bias removed ATAC-seq accessibility. We finally combine the transformation and residual models to get the final combined model. The bias model was trained on 30882 inaccessible, low count regions. The residual model was trained on 78102 peak regions split between train, validation and test sets. During this training step, we considered only the positional information provided by the bias model when assessing output predictions. A bunch of different model architectures and parameters were tested and trained on, for optimisation and selection of the model based on performance and interpretability. Model performance was determined based on Pearson & Spearman correlations of predicted versus observed total experimental counts, Jensen-Shannon distances of predicted and observed profiles and multinomial negative log-likelihoods of predicted and observed profiles. The final residual model BPReveal ATAC model was trained with 9 convolutional layers, a filter depth of 128, counts loss weight of 100, input filter length of 7 and output filter width of 25.

With the associated architecture, the input DNA sequences were 3074 bp long and the desired output prediction window of 1000bp. BPReveal's reimplementation of BPNet does not pad through convolutions, but rather requires a larger input sequence to provide the receptive field with sufficient sequence information as the model deepens.

Model Training for the other 3 species

The model architecture was kept the same as the melanogaster model mentioned above, while training models on other 3 species. However, after parameter and architecture optimisation, an overall performance increase was seen - an increase of 0.05 for *D. erecta*, an increase of 0.08 for *D. ananassae* in terms of Spearman correlation and performance remained the same for *D. mojavensis*. But since we are not using these models for any further motif-level analysis, the original same architecture model was retained for any analysis.

Apart from that, to ensure a fair and controlled comparison across datasets, the model architecture was kept constant throughout all experiments. This design choice isolates the effect of dataset characteristics on model performance, preventing architectural differences from confounding the results. The performance variations from these models can be attributed primarily to dataset-specific factors such as size, complexity, and distribution.

4.3 tfmodisco-lite

TF-MoDISco (Transcription Factor Motif Discovery from Importance Scores) (Shrikumar et al. 2018) is a post-hoc interpretability algorithm designed to extract consolidated transcription factor binding motifs from the attribution scores of deep learning models. Unlike traditional motif discovery tools that rely on the statistical overrepresentation of nucleotide sequences (PWMs), TF-MoDISco operates on contribution scores (DeepLIFT or SHAP values) that quantify the predictive importance of each base.

The method operates in three main stages. First, it extracts high-importance sequence segments, known as seqlets, by scanning per-base contribution scores to identify regions where the model detects strong regulatory signals. Next, it computes pairwise similarity between seqlets by optimally aligning them based on their contribution profiles rather than just nucleotide identity, enabling recognition of motifs with slight sequence variations but conserved functional importance. Finally, similar seqlets are clustered and the resulting clusters are consolidated into Contribution Weight Matrices (CWMs), which summarize both base frequency and average contribution, capturing complex regulatory patterns beyond simple sequence alignment. tfmodisco-lite is the standard, optimized, and faster implementation of the algorithm in TF-MoDISco and BPReveal's implementation of this `tfmodisco-lite` (v2.2.0) was used in all the analysis. Total number of seqlets was set to 100000 and a window size of 1000 was used. The report feature of modisco was executed using the JASPAR 2022/2024 insect motif database.

4.4 Motif mapping

For the trained BPReveal accessibility model, I generated the sequence contribution scores using BPReveal's implementation of deepSHAP, modified to generate hypothetical con-

tribution scores similar to those from DeepLIFT. Using the counts contribution scores as an input to tf-modiscolite, I generated contribution motif representations called Contribution Weight Matrices (CWM) for each model. Then I manually curated motif identities based on existing literature.

To map the modisco-lite CWM motif representations back to the genome, we performed CWM scanning implemented by BPReveal. Across the whole genome, I mapped a motif if the genomic region matched the criteria designated by the corresponding modisco-motif distributions : (1) the Jaccardian similarity between the motif CWM and the genomic site's sequence contribution exceeded the 20th percentile of the TF-MoDISco motif representations and (2) the contribution L1 magnitude is higher than the TF-MoDISco motif seqlets lowest contribution score and (3) the sequence match (PWM score) is higher than the 10th percentile of TF-MoDISco motif PWM.

For the other species, genome-wide SHAP scores were calculated using the predictions of the *D. melanogaster* ATAC model on those species. Modisco was then run on these scores to get the motif instances via CWM scanning. This was done so as to retain the same model for further analysis and also because the *D. melanogaster* model performed fairly well across all species.

4.5 Progressive cactus

Progressive Cactus (Armstrong et al. 2020; Paten et al. 2011) is a reference free whole genome aligner which uses a progressive alignment strategy. It uses cactus graphs to model the evolutionary history of a DNA sequence.

Unlike many other aligners, Progressive Cactus avoids reference bias and instead uses cactus graphs to model the universal substructure of genome alignments, allowing accurate representation of complex evolutionary events such as rearrangements and copy-number variations. Because of these features, it can effectively model complex evolutionary histories - something traditional matrix-based aligners struggle to achieve. Unlike linear alignments, these graphs model homology as a network where any edge belongs to at most one simple cycle, a property that prevents combinatorial entanglement. This allows the global alignment problem to be naturally decomposed into a hierarchy of independent subproblems, enabling the software to scale linearly to thousands of genomes while accurately preserving the non-linear history of genome evolution.

Algorithmically, the method follows a guide tree to recursively align small sets of genomes at each internal node. For each subproblem, ingroup genomes are aligned together with selected outgroups to infer an ancestral genome. These ancestral reconstructions are then propagated upward and reused as inputs for higher-level comparisons. At each stage, alignments are represented and refined within the cactus graph framework to remove inconsistencies, consolidate homologous regions, and resolve duplication histories,

ultimately producing a hierarchical alignment that captures both sequence homology and deep evolutionary structure across many genomes.

This tool requires 2 primary inputs - A guide tree - a phylogenetic tree that defines the evolutionary relationships, and the genome FASTA file containing the sequences. The genome FASTA files obtained from Kim et al. (2021) were filtered using `repeatMasker` (Smit et al. 2015) and a phylogenetic tree was created using a tool called `mashtree` (Katz et al. 2019) before giving these inputs to Cactus (version 2.9.7). Supporting tools from HALTools (Hickey et al. 2013) like `hal2chain, hal2maf` were used to obtain the chain files and the multi alignment files.

4.6 LiftOver

UCSC liftOver (Hinrichs 2006) is a tool that converts genomic coordinates between assemblies of same or different species by referencing a chain file, which contains pairwise alignment data describing how segments of the source genome correspond to the target genome. For each input interval, the tool locates the relevant alignment block within a chain and mathematically projects the start and end positions onto the target coordinate system, accounting for strand orientation and offsets. This tool has an attribute/parameter called `-minMatch`, which checks if a specified percentage of bases map successfully. If an interval fails this threshold or falls into deleted regions, it is excluded and written to an unmapped file.

All ATAC peaks were lifted over with a `minMatch` threshold of 0.1 and all motif instances were lifted over with a `minMatch` threshold of 0.01. The threshold is kept really low here, since we are not focused on the degree of sequence divergence.

4.7 Multi-species model training

The multi species model was built in 2 different approaches as described in Results 2.4. The data used was not the same as the one used before. Here, the 2 replicate bigwig files were added together to get the combined ATAC bigwig file for each species. The peaks were also called on this combined bigwig file. These were considered for the model training as discussed below.

The first method was by just pooling all the species data together - all the 4 ATAC seq bigwig files from the 4 species were combined together into a mega bigwig file after normalisation by the read depth. The second method is the species aware method. The individual ATAC-seq files from different species were first normalised by the read depth as before. The structure of BPReveal was modified to take an input of 8 channels instead of 4 as in the canonical one hot encoding. The additional 4 channels were to encode

species information. If the sequence was from species i, the whole row/channel of i+4 was filled with 1s and the other 3 rows to be zeroes, as shown in Figure 2.8. So, the model is seeing the species information along with the sequence input. For the species aware model, bias correction was also based on the above model structure and hence the bias model was also species aware. For all of the above models, the model architecture was kept the same as the initial *D. melanogaster* model - consisting of 9 layers, counts weight loss of 100 and 128 filters. These multi species models were trained and tested on 3 different criteria:

- Non orthologous regions: The train, test & validation sets were constructed based on individual species ATAC data. The respective train, test and val sets were pooled together and given as input to the model. (total regions - 262740)
- Orthologous regions: The train, test & validation sets were constructed for *D. melanogaster* based on *D. melanogaster* ATAC data. The orthologous regions in other species of these *D. melanogaster* regions were added to the respective category of train, test, val and given to the model. (total regions - 236216)
- Orthologous regions expanded: In addition to what was done in the second procedure, species-specific peaks were considered—defined as accessible regions in each species that had a successful liftover onto other species but were not identified as peaks in those other species. These regions were added onto the test and validation sets only. (Total regions - 313669; train - 139534; test - 87324; val - 86811)

4.8 Modelling DNA evolution

4.8.1 IQ-TREE

IQ-TREE (Wong et al. 2025) is a software for inferring phylogenetic trees by searching for the tree topology that maximizes the likelihood of the observed sequence data. It uses a stochastic algorithm that overcomes the limitations of traditional hill-climbing methods, which often get trapped in suboptimal local peaks by maintaining a dynamic set of candidate trees. This efficient exploration allows it to locate the global maximum likelihood tree more reliably than old methods.

The search begins by generating a set of initial candidate trees, typically using fast distance-based methods like Neighbor-Joining or Parsimony. Then, the algorithm takes a candidate tree and applies topological rearrangements on these candidates and generates new topologies. In this step, the algorithm slightly alters the topology or optimizes the parameters and re-calculates the likelihood using the pruning algorithm. If the modification results in a statistically significant improvement in the log-likelihood score, the new tree is retained; otherwise, it may be discarded (or kept with a low probability to

escape local optima). This cycle repeats until the algorithm converges on a topology. To avoid getting trapped in local optima, it applies random rearrangements to the current best tree (to search a different part of the tree space), while also maintaining the pool of top candidate trees found so far. It also performs Ultrafast Bootstrap on this optimized topology to estimate the reliability of each split without re-doing the entire search.

4.8.2 MAFFT (Multiple Alignment using Fast Fourier Transform)

MAFFT (Kuraku et al. 2013) is a high-speed algorithm designed to rapidly align large nucleotide datasets by identifying homologous regions through spectral analysis. It converts DNA sequences into numerical vectors and applies the Fast Fourier Transform (FFT) to quickly calculate cross-correlations, identifying conserved segments without the exhaustive computational cost of standard dynamic programming. This allows for the efficient handling of thousands of sequences by rapidly establishing anchor points for the alignment. It consists of a progressive alignment stage that builds an initial draft based on a guide tree, followed by an iterative refinement stage. In the refinement phase, the algorithm repeatedly partitions the alignment into subgroups and re-aligns them to maximize a weighted sum-of-pairs objective score. This iterative process corrects misalignment errors introduced by the initial greedy tree-building approach, resulting in a statistically robust final alignment topology.

4.8.3 Problem with HAL files

To simulate DNA evolution using a particular substitution matrix, we first need to obtain the model parameters based on the observed data. Using the already existing HAL/MAF files from CACTUS ran into a lot of issues while extracting this data. We need the alignment in a fasta file, and tools like `maf-convert, maf_parse` (PHAST) and `hal2fasta` (ComparativeGenomicsToolkit) either discarded the alignment information or did not have the option of fasta output. `msa_view` (PHAST) tool extracted the whole alignment block that overlaps my region of interest and does not allow me to specify the coordinates while extracting the fasta sequences. This was especially problematic if the region of interest spans multiple blocks. We wouldn't know where they come from. This will no longer be ideal since we want the parameters for our regions of interest. Hence, I had to turn to the approach of extracting the orthologous coordinates using LiftOver of cactus generated chain files, and then using an external software to align these orthologous regions, which is where MAFFT was used, as described below.

4.8.4 Workflow

Since our regions of interest are accessible regions or ATAC peaks, I took all the ATAC peaks of melanogaster which had a successful liftover in all the 4 species (23,220). These peaks were resized to 4000bp before liftover to account for the input window of the BPReveal model. I extracted the sequences from species' genomes of these orthologous regions and ran the following command on MAFFT (v7.480) to align them. (`mafft --localpair --maxiterate 1000 --adjustdirectionaccurately`)

I gave these aligned sequences to IQTREE (v3.0.1) to extract the model parameters, using the HKY substitution model (`iqtree3 -m HKY -T AUTO`). The bootstrap parameter was added and tested which yielded the same results confirming the topology of the final tree. I got the model parameters from the above tree on our regions of interest (Table 4.2). Note that - since we are using the HKY model for simulation, the same model must be used to construct the tree. Otherwise the model parameters would not make sense under a different model framework. ($\kappa = 2.095, \pi = \{A : 0.286, C : 0.214, G : 0.214, T : 0.286\}$)

Table 4.2: Parameters for simulation of evolution across *Drosophila* species

Species	**Branch length**	**Indel rate**	**Insertion probability**	**Insertion length** (mean)	**Deletion length** (mean)
D. erecta	0.1299	0.00397	0.4055	6	3
D. ananassae	0.4148	0.00887	0.5242	6	6
D. mojavensis	0.7489	0.01127	0.5811	7	6

Then, I use these parameters to simulate evolution keeping the base/starting sequences as the *D. melanogaster* 4kb sequences. I did not use any software since I wanted to incorporate indels into my evolution simulation, and not only the substitution rate, which most of the already existing softwares don't do (eg. SeqGen). So, my simulation of evolution is a modified algorithm based on the HKY85 substitution model. I calculated the gaps in each alignment of a peak to get the distribution of probability and lengths of insertions, deletions. While simulating the evolution of a sequence, before a substitution occurs at a site, a random number is drawn. If it is less than (indel rate $\times$ insertion probability), then an insertion occurs by sampling the length from a geometric distribution with the predefined mean. If it is less than the indel rate, then a deletion occurs similarly. If neither, then, it proceeds to the substitution of the current site using the substitution matrix Q. So the total indel probability per base is the indel rate, with insertions taking a fraction of {insertion probability} and deletions, the remainder.

I simulate this evolution of sequence for every ATAC peak in consideration (23,220), for each of the 3 different species in hand, using the different species-specific parameters obtained from IQTREE. I do this simulation 100 times for each peak. So finally, per ATAC seq peak in *D. melanogaster*, we have 100 mutated peaks at 3 different branch

lengths, each corresponding to one of the other species.

After mutation of these sequences, I resize them to 3074 bp, the exact input window of the BPReveal model, and I give these sequences to be predicted on, by the same model. The predictions across the complete window size of 1000bp were considered for further analysis.

4.9 Jaccard similarity

The Jaccard similarity coefficient serves as a fundamental metric for gauging the diversity and similarity of sample sets. Mathematically, it is defined as the size of the intersection of two sets divided by the size of their union: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. This metric is constrained to a range between 0 and 1, where a value of 1 signifies identical sets and 0 indicates completely disjoint sets. A primary property of Jaccard similarity is its robustness against "negative matches"; unlike simple matching coefficients, it does not account for elements absent in both sets, making it particularly effective for sparse data environments.

Here we use it to measure the similarity of motif composition between 2 orthologous regions. This metric is blind to how many instances of the motif there are in that region and where they are located in that region, and just gives us a measure of the diversity of motifs in that region. We calculate it as the number of TF motifs mapped in both regions divided by the number of TF motifs mapped in either of the 2 regions, as seen in Figure 2.15.

All *D. melanogaster* peaks were resized to 1kb centered at the summit and the motif instances overlapping the region were allocated to that peak. These summits were lifted over to the other species, resized to 1kb and the motif instances in those species (Methods 4.4) overlapping the orthologous peaks were allocated to those peaks. Then the Jaccard similarity was calculated between the motif sets of the orthologous peaks. This was repeated by adding the species specific peaks also (defined as peaks that had a successful liftover but not overlapping a peak in the other species), which revealed the same trend (not shown).

4.10 Low affinity motifs

All the mapped *Grh* instances were considered in both *D. melanogaster* and *D. erecta* (n=6202 and n=6993). They were lifted over to the other species in consideration (*D. melanogaster* → *D. erecta* and *D. erecta* → *D. melanogaster*). These motif instances were not combined but kept separate; all the analysis has been shown for *D. melanogaster* → *D. erecta*, but it was repeated for the other direction as well, which revealed the same trend (not shown).

The motif instances were divided into 2 categories based on their sequence conservation - Conserved category, where the sequence in *D. erecta* is exactly equal to the sequence in *D. melanogaster*, and Turnover category, where the sequence in *D. erecta* is different from the sequence in *D. melanogaster*. This gave 3545 and 1710 motifs in the Conserved and Turnover category respectively. The motif instances were also categorized into 5 different categories based on their PWM score (Table 4.3). This was done using the distribution of PWM scores of the mapped *Grh* motif instances in the two species. PWM score refers to the log-odds score wherever mentioned in the text.

Table 4.3: PWM score categories for *Grh* motif instances

Category of Affinity	**Percentile**	**PWM Score (log-odds)**
Not motif	$\leq 0\%$	< 5
Low	0–25%	5–10.8
Medium 1	25–50%	10.8–12.6
Medium 2	50–75%	12.6–14.1
High	75–100%	> 14.1

To categorise the LAMs based on its surrounding, a neighbourhood of 300bp on either side of the motif instance was considered. Any motif overlap was checked, and the LAM was categorised as "Cluster" if there was a high affinity pioneer motif in the neighbourhood. A high affinity pioneer motif was defined as one of *Grh, GAGA, GATA* with a PWM score higher than the 75th percentile of the respective motif instances in that species. If there was no such motif in the neighbourhood, then the LAM was categorised as "Non-cluster".

References

1. Armstrong, Joel et al. (Nov. 2020). "Progressive Cactus is a multiple-genome aligner for the thousand-genome era". In: *Nature* 587.7833, 246–251. DOI: 10.1038/s41586-020-2871-y.
2. Avsec, Žiga et al. (Feb. 2021a). "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature Genetics* 53.3, 354–366. DOI: 10.1038/s41588-021-00782-6.
3. Avsec, Žiga et al. (Oct. 2021b). "Effective gene expression prediction from sequence by integrating long-range interactions". en. In: *Nature Methods* 18.10, 1196–1203. DOI: 10.1038/s41592-021-01252-x.
4. Avsec, Žiga et al. (Jan. 2026). "Advancing regulatory variant effect prediction with AlphaGenome". en. In: *Nature* 649.8099, 1206–1218. DOI: 10.1038/s41586-025-10014-0.
5. Boer, Carl G. de and Jussi Taipale (Dec. 2023). "Hold out the genome: a roadmap to solving the cis-regulatory code". In: *Nature* 625.7993, 41–50. DOI: 10.1038/s41586-023-06661-w.
6. Britten, Roy J. and Eric H. Davidson (1971). "Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty". In: *The Quarterly Review of Biology* 46.2, pp. 111–138. (Visited on 03/09/2026).
7. Broad Institute (2019). *Picard Tools*. Version 2.20.3. Accessed: 2026-02-12. URL: http://broadinstitute.github.io/picard/.
8. Buenrostro, Jason D et al. (Oct. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10.12, 1213–1218. DOI: 10.1038/nmeth.2688.
9. Carroll, Sean B (July 2005). "Evolution at Two Levels: On Genes and Form". In: *PLoS Biology* 3.7, e245. DOI: 10.1371/journal.pbio.0030245.
10. Cho, Ken W.Y. (Apr. 2012). "Enhancers". In: *WIREs Developmental Biology* 1.4, 469–478. DOI: 10.1002/wdev.53.
11. Dsilva, Greg Jude and Sanjeev Galande (Mar. 2024). "From sequence to consequence: Deciphering the complex cis-regulatory landscape". In: *Journal of Biosciences* 49.2. DOI: 10.1007/s12038-024-00431-0.

12. Gompel, Nicolas et al. (Feb. 2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila". In: *Nature* 433.7025, 481–487. DOI: 10.1038/nature03235.
13. Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of Molecular Evolution* 22.2, pp. 160–174. DOI: 10.1007/BF02101694.
14. Hickey, Glenn et al. (Mar. 2013). "HAL: a hierarchical format for storing and analyzing multiple genome alignments". In: *Bioinformatics* 29.10, 1341–1342. DOI: 10.1093/bioinformatics/btt128.
15. Hinrichs, A. S. (Jan. 2006). "The UCSC Genome Browser Database: update 2006". In: *Nucleic Acids Research* 34.90001, D590–D598. DOI: 10.1093/nar/gkj144.
16. Jacob, François (1977). "Evolution and Tinkering". In: *Science* 196.4295, pp. 1161–1166. DOI: 10.1126/science.860134. eprint: https://www.science.org/doi/pdf/10.1126/science.860134.
17. Jukes, Thomas H. and Charles R. Cantor (1969). "Evolution of Protein Molecules". en. In: *Mammalian Protein Metabolism*. Elsevier, 21–132. ISBN: 9781483232119. DOI: 10.1016/B978-1-4832-3211-9.50009-7. URL: https://linkinghub.elsevier.com/retrieve/pii/B9781483232119500097.
18. Katz, Lee et al. (Dec. 2019). "Mashtree: a rapid comparison of whole genome sequence files". In: *Journal of Open Source Software* 4.44, p. 1762. DOI: 10.21105/joss.01762.
19. Kelley, David R., Jasper Snoek, and John L. Rinn (July 2016). "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks". en. In: *Genome Research* 26.7, 990–999. DOI: 10.1101/gr.200535.115.
20. Khoueiry, Pierre et al. (Aug. 2017). "Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity". In: *eLife* 6. DOI: 10.7554/elife.28440.
21. Kim, Bernard Y et al. (July 2021). "Highly contiguous assemblies of 101 drosophilid genomes". In: *eLife* 10. DOI: 10.7554/elife.66405.
22. Kim, Seungsoo and Joanna Wysocka (Feb. 2023). "Deciphering the multi-scale, quantitative cis-regulatory code". In: *Molecular Cell* 83.3, 373–392. DOI: 10.1016/j.molcel.2022.12.032.
23. Kimura, Motoo (Oct. 1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. ISBN: 9780511623486. DOI: 10.1017/cbo9780511623486. URL: http://dx.doi.org/10.1017/CBO9780511623486.
24. Kuraku, Shigehiro et al. (May 2013). "aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity". In: *Nucleic Acids Research* 41.W1, W22–W28. DOI: 10.1093/nar/gkt389.
25. Langmead, Ben and Steven L Salzberg (Mar. 2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4, 357–359. DOI: 10.1038/nmeth.1923.

26. Linder, Johannes et al. (Apr. 2025). "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation". en. In: *Nature Genetics* 57.4, 949–961. DOI: 10.1038/s41588-024-02053-6.
27. Martin, Marcel (May 2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1, p. 10. DOI: 10.14806/ej.17.1.200.
28. McAnany, Charles E. et al. (Apr. 2025). "Positional Interpretation of Cis-Regulatory Code and Nucleosome Organization with Deep Learning Models". In: DOI: 10.1101/2025.04.07.647613.
29. Meireles-Filho, Antonio CA and Alexander Stark (Dec. 2009). "Comparative genomics of gene regulation—conservation and divergence of cis-regulatory information". In: *Current Opinion in Genetics & Development* 19.6, 565–570. DOI: 10.1016/j.gde.2009.10.006.
30. Pampari, Anusri et al. (Dec. 2024). "ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants". In: DOI: 10.1101/2024.12.25.630221.
31. Panigrahi, Anil and Bert W. O'Malley (Apr. 2021). "Mechanisms of enhancer action: the known and the unknown". In: *Genome Biology* 22.1. DOI: 10.1186/s13059-021-02322-1.
32. Paten, Benedict et al. (June 2011). "Cactus: Algorithms for genome multiple sequence alignment". In: *Genome Research* 21.9, 1512–1528. DOI: 10.1101/gr.123356.111.
33. Phan, Mai H. Q. et al. (May 2025). "Conservation of regulatory elements with highly diverged sequences across large evolutionary distances". In: *Nature Genetics* 57.6, 1524–1534. DOI: 10.1038/s41588-025-02202-5.
34. Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning Important Features Through Propagating Activation Differences". In: DOI: 10.48550/ARXIV.1704.02685.
35. Shrikumar, Avanti et al. (2018). *Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5*. DOI: 10.48550/ARXIV.1811.00416. URL: https://arxiv.org/abs/1811.00416.
36. Signor, Sarah A. and Sergey V. Nuzhdin (July 2018). "The Evolution of Gene Expression in cis and trans". In: *Trends in Genetics* 34.7, 532–544. DOI: 10.1016/j.tig.2018.03.007.
37. Smit, Arian F. A., Robert Hubley, and Phil Green (2015). *RepeatMasker Open-4.0*. http://www.repeatmasker.org. Accessed 2013–2015.
38. Visel, Axel, Edward M. Rubin, and Len A. Pennacchio (Sept. 2009). "Genomic views of distant-acting enhancers". In: *Nature* 461.7261, 199–205. DOI: 10.1038/nature08451.

39. Weilert, Melanie et al. (Nov. 2025). "Widespread low-affinity motifs enhance chromatin accessibility and regulatory potential in mESCs". In: DOI: 10.1101/2025.11.18.685822.
40. Wikipedia contributors (2002). *Phylogenetic tree*. https://en.wikipedia.org/wiki/Phylogenetic_tree. [Online; accessed 9-March-2026].
41. — (2003). *Neutral theory of molecular evolution*. https://en.wikipedia.org/wiki/Neutral_theory_of_molecular_evolution. [Online; accessed 9-March-2026].
42. — (2005). *Substitution model*. https://en.wikipedia.org/wiki/Substitution_model. [Online; accessed 9-March-2026].
43. — (2006). *Models of DNA evolution*. https://en.wikipedia.org/wiki/Models_of_DNA_evolution. [Online; accessed 9-March-2026].
44. Wittkopp, Patricia J., Belinda K. Haerum, and Andrew G. Clark (July 2004). "Evolutionary changes in cis and trans gene regulation". In: *Nature* 430.6995, 85–88. DOI: 10.1038/nature02698.
45. Wittkopp, Patricia J. and Gizem Kalay (Dec. 2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". In: *Nature Reviews Genetics* 13.1, 59–69. DOI: 10.1038/nrg3095.
46. Wong, Thomas et al. (Apr. 2025). "IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models". In: DOI: 10.32942/x2p62n.
47. Wray, Gregory A. (Mar. 2007). "The evolutionary significance of cis-regulatory mutations". In: *Nature Reviews Genetics* 8.3, 206–216. DOI: 10.1038/nrg2063.
48. Yáñez-Cuna, J. Omar, Evgeny Z. Kvon, and Alexander Stark (Jan. 2013). "Deciphering the transcriptional cis-regulatory code". In: *Trends in Genetics* 29.1, 11–22. DOI: 10.1016/j.tig.2012.09.007.
49. Zhang, Yong et al. (Sept. 2008). "Model-based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9. DOI: 10.1186/gb-2008-9-9-r137.
50. Zhou, Jian and Olga G Troyanskaya (Aug. 2015). "Predicting effects of noncoding variants with deep learning–based sequence model". In: *Nature Methods* 12.10, 931–934. DOI: 10.1038/nmeth.3547.

Appendix A

Markov Chains

I detail the mathematical foundations of Continuous-Time Markov Chains (CTMCs) that underpin the DNA substitution models discussed in Chapter 1.2.2. This appendix serves as a reference for the theoretical concepts and properties of Markov chains, providing the necessary background for understanding their application in modeling molecular evolution. The content is structured to be self-contained, allowing readers to grasp the essential mathematical details without needing to consult external sources.

A.1 Mathematical Details

Definition 1. *Let $(X_t)_{t\in T}$ be a sequence (or family) of random variables taking values in a state space E. The process is called a Markov chain if, for all times $t_0 < t_1 < \cdots < t_n < t$ and all states $i_0, i_1, \ldots, i_n, j \in E$,*

$$\mathbf{P}\left(X_t = j \mid X_{t_n} = i_n, X_{t_{n-1}} = i_{n-1}, \ldots, X_{t_0} = i_0\right) = \mathbf{P}\left(X_t = j \mid X_{t_n} = i_n\right).$$

Remark 1. *Continuous-time Markov chains are defined by transition probabilities which can be represented in the form of transition matrices which are, in addition, parameterized by time, t. The transition matrix $P(t) = \big(P_{ij}(t)\big)$ where each individual entry, $P_{ij}(t)$ refers to the probability that state E_i will change to state E_j in time t.*

Definition 2 (Continuous-Time Markov Chain)**.** *Let S be a countable state space. A family of random variables $\{X(t) : t \geq 0\}$ taking values in S is called a Continuous-time Markov chain if, for all $0 \leq t_1 < t_2 < \cdots < t_n < t_{n+1}$ and all states $i_1, \ldots, i_n, j \in S$,*

$$\mathbb{P}(X(t_{n+1}) = j \mid X(t_n) = i_n, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1) = \mathbb{P}(X(t_{n+1}) = j \mid X(t_n) = i_n).$$

Holding times and jump chain. If $X(t) = i$, the time spent in state i before the next transition is called the *holding time*. This is exponentially distributed with parameter $\lambda_i \geq 0$. The exponential distribution is the unique continuous distribution with the

memoryless property, which ensures the Markov property in continuous time.
When the holding time expires, the process must leave state i. The Jump Chain is the discrete-time mechanism that decides the next state. The process jumps to state j with probability P_{ij}^{jump}, where $\sum_{j \in S} P_{ij}^{jump} = 1$.

Note that for non-absorbing states, $P_{ii}^{jump} = 0$. 'Transitioning' from i to i in continuous time is indistinguishable from simply staying in i longer (which is already controlled by the holding time parameter λ_i). Therefore, a transition event always implies a change of state.

Therefore, A continuous-time Markov chain is completely characterized by the transition probabilities $\{P_{ij}^{jump}\}$ of the embedded discrete-time Markov chain and the holding rates $\{\lambda_i\}_{i \in S}$.

Transition Probabilities The process is said to be *time-homogeneous* if the conditional transition probabilities do not depend on the current time, that is,

$$\mathbb{P}(X_{s+t} = j \mid X_s = i) = \mathbb{P}(X_t = j \mid X_0 = i)\,, \qquad s, t \geq 0.$$

In this case, we define the transition probabilities by

$$P_{ij}(t) := \mathbb{P}(X_{s+t} = j \mid X_s = i) = \mathbb{P}(X_t = j \mid X_0 = i)\,, \qquad t \geq 0.$$

Remark 2. *For a discrete-time Markov chain, the one-step transition probability is*

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i).$$

The n-step transition probability is defined by

$$P_{ij}^{(n)} = \mathbb{P}(X_n = j \mid X_0 = i),$$

and is given by $(P^{(n)})_{ij}$, where $P = (p_{ij})$ is the one-step transition matrix.

The Chapman–Kolmogorov equations*: For all $m, n \geq 0$,*

$$P_{ij}^{(m+n)} = \sum_{k \in S} P_{ik}^{(m)}\, P_{kj}^{(n)}.$$

To go from state i to state j in $m+n$ steps, the chain must pass through some intermediate state k after m steps. Therefore, in terms of transition matrices,

$$P^{(m+n)} = P^{(m)} P^{(n)} \quad \text{and hence,} \quad P^{(n)} = P^n$$

For a continuous-time Markov chain (CTMC), the transition probabilities are defined by $P_{ij}(t) = \mathbb{P}(X(t) = j \mid X(0) = i)$, and the Chapman–Kolmogorov equations take the form

$$P(s + t) = P(s)P(t), \quad s, t \geq 0.$$

Definition 3. State probability: *Consider a Markov Chain having a state space S. For a DTMC with transition probabilities P_{ij}, the state probability is $\pi_j^{(n)} = P\{X_n = j\}$, the probability of finding the system in state j at time n.*
For a CTMC $\{X(t)\}_{t\geq 0}$, the state probability at time t is defined as $\pi_j(t) := \mathbb{P}(X(t) = j), j \in S$

Remark 3. *For a DTMC,*

$$\pi_j^{(m)} = \sum_{i\in S} \pi_i^{(m-1)} P_{ij} \quad \text{for all } m \geq 1 \text{ and all } i, j \in S.$$

$$\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(m-1)} P.$$

$$\boldsymbol{\pi}^{(m)} = \boldsymbol{\pi}^{(m-1)} P = \boldsymbol{\pi}^{(m-2)} P^2 = \cdots = \boldsymbol{\pi}^{(0)} P^m$$

where $\boldsymbol{\pi}^{(0)}$ is the initial state distribution.

For a CTMC, the state probabilities satisfy

$$\pi_j(t) = \sum_{i\in S} \pi_i(s)\, P_{ij}(t-s), \quad 0 \leq s \leq t.$$

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(s)\, P(t-s), \quad 0 \leq s \leq t.$$

In particular, taking $s = 0$,

$$\boldsymbol{\pi}(t) = \boldsymbol{\pi}(0)\, P(t),$$

where $\boldsymbol{\pi}(0)$ is the initial state distribution.

Definition 4. *Let $\{X_n\}$ be a Markov chain on a finite state space S with transition matrix P.*

1. *Irreducibility: A chain is irreducible if for any two states $i, j \in S$, there exists an $n > 0$ such that $(P^n)_{ij} > 0$. This means every state is reachable from every other state.*

2. *Periodicity: The period $d(i)$ of a state i is the greatest common divisor (GCD) of all n such that $(P^n)_{ii} > 0$. If $d(i) = 1$ for all i, the chain is aperiodic.*

3. *Recurrence: Let T_i be the first return time to state i. $T_i = \min\{n \geq 1 : X_n = i \mid X_0 = i\}$.*

   - *A state is transient if $P(T_i < \infty) < 1 \implies E[T_i] = \infty$.*
   - *A state is recurrent if $P(T_i < \infty) = 1$*
     - *A state is positive recurrent if $P(T_i < \infty) = 1$ and $E[T_i] < \infty$.*
     - *A state is null recurrent if $P(T_i < \infty) = 1$ but $E[T_i] = \infty$.*

   A chain is recurrent/transient if all of its states are recurrent/transient.

4. *Ergodicity: A state i is said to be ergodic if it is aperiodic and positive recurrent. A chain is ergodic if it is irreducible, aperiodic, and positive recurrent (which is guaranteed in finite state spaces).*

Remark 4. *The above definitions hold for a DTMC. For a CTMC, they are slightly altered, though the meaning still remains the same. Irreducibility* $\implies \exists t > 0 \quad s.t \quad P_{ij}(t) > 0$. *Periodicity is not defined for CTMCs. The first return time* T_i *is defined as* $T_i = \inf\{t > 0 : X(t) = i \mid X(0) = i\}$ *and recurrence definition still holds. A chain is ergodic if it is irreducible and positive recurrent.*

Lemma 1. Coupling Inequality : *Let* X *and* Y *be two random variables taking values in the same discrete state space* S. *The coupling inequality states that for any subset of states* $A \subseteq S$*:*

$$|P(X \in A) - P(Y \in A)| \leq P(X \neq Y)$$

Proof. Let X and Y be random variables taking values in the same discrete state space S, and let $A \subseteq S$.

We write the difference of probabilities as an expectation of indicator functions:

$$P(X \in A) - P(Y \in A) = \mathbb{E}\left[\mathbf{1}_{\{X \in A\}} - \mathbf{1}_{\{Y \in A\}}\right].$$

Taking absolute values and applying the triangle inequality yields

$$|P(X \in A) - P(Y \in A)| \leq \mathbb{E}\left[\left|\mathbf{1}_{\{X \in A\}} - \mathbf{1}_{\{Y \in A\}}\right|\right].$$

$$\left|\mathbf{1}_{\{X \in A\}} - \mathbf{1}_{\{Y \in A\}}\right| = \begin{cases} 1, & \text{if exactly one of } X, Y \text{ belongs to } A, \\ 0, & \text{otherwise.} \end{cases}$$

In particular, $\left|\mathbf{1}_{\{X \in A\}} - \mathbf{1}_{\{Y \in A\}}\right| \leq \mathbf{1}_{\{X \neq Y\}}$

$$\mathbb{E}\left[\left|\mathbf{1}_{\{X \in A\}} - \mathbf{1}_{\{Y \in A\}}\right|\right] \leq \mathbb{E}\left[\mathbf{1}_{\{X \neq Y\}}\right] = P(X \neq Y).$$

$$|P(X \in A) - P(Y \in A)| \leq P(X \neq Y),$$

□

Lemma 2. *In an irreducible chain, recurrence/transience is a class property. If one state is transient, every state* $j \in S$ *must be transient.*

Proof. First we prove a result. We know that state i is transient if $P(T_i < \infty) < 1$ where T is the first return time $T_i = min\{n \geq 1 : X_n = i \mid X_0 = i\}$.
Claim : State i is transient $\Leftrightarrow \sum_{n=0}^{\infty}(P^n)_{ii} < \infty$ where $(P^n)_{ij} = P(X_n = j \mid X_0 = i)$
Proof: Let N_i be the total number of visits to state i.

$$N_i = \sum_{n=0}^{\infty} \mathbf{1}_{\{X_n = i\}}$$

$$\mathbb{E}[N_i \mid X_0 = i] = \sum_{n=0}^{\infty} P(X_n = i | X_0 = i) = \sum_{n=0}^{\infty} (P^n)_{ii}$$

Let $f := P(T_i < \infty \mid X_0 = i)$ - probability of returning back to i. Suppose the chain returns to i at T_i. Then the future evolution does not depend on the past, and this return resets the whole process. So, the chain will return with a probability of f.

$$P(N_i = k) = f.f.f...(k-1 \text{ times}).(1-f) = f^{k-1}(1-f)$$

This is the probability that chain **returns** to i exactly $k-1$ times. (It is already there at $t = 0$)

$$\mathbb{E}[N_i \mid X_0 = i] = \frac{1}{1-f}$$

for a geometric random variable.

From $\sum_{n=0}^{\infty}(P^n)_{ii} = \mathbb{E}[N_i \mid X_0 = i]$, we can conclude the following: If $f < 1$, finite expected number of visits $\implies$ transient
If $f = 1$, infinite expected number of visits $\implies$ recurrent
Now, back to the main proof:
Fix any 2 states $i, j \in S$. Since the chain is irreducible, $\exists m, n \in \mathbb{Z}$ s.t $(P^m)_{ij} > 0$ and $(P^n)_{ji} > 0$.

For any $k \geq 0$, by the Chapman–Kolmogorov equations,

$$(P^{(m+k+n)})_{ii} = \sum_{l \in S} \sum_{r \in S} P^m_{il}.P^k_{lr}.P^n_{ri}$$

Since all the terms here are non-negative, take $l = r = j$ for some j which gives

$$(P^{(m+k+n)})_{ii} \geq P^m_{ij}.P^k_{jj}.P^n_{ji}$$

Summing over k,

$$\sum_{r=m+n}^{\infty} (P^r)_{ii} \geq P^m_{ij}.P^n_{ji} \sum_{k=0}^{\infty} P^k_{jj}$$

With a simple change of index, we get

$$\sum_{r=0}^{\infty} (P^r)_{ii} \geq P^m_{ij}.P^n_{ji} \sum_{k=0}^{\infty} P^k_{jj}$$

Now, if a state i is transient , then $\sum_{r=0}^{\infty}(P^r)_{ii} < \infty$. From the inequality above, $\sum_{k=0}^{\infty}(P^k)_{jj} < \infty$. Hence, state j is also transient. Since j was arbitrary, every state is transient.
Similarly, the proof holds for recurrence also. □

Theorem 3. *If a Markov chain on a finite space S is irreducible, it is positive recurrent.*

Proof. If the chain is irreducible and finite, at least one state must be recurrent.

If all states were transient, then for any state j,

$$\sum_{n=1}^{\infty} P(X_n = j \mid X_0 = i) < \infty$$

The probability of being in any state j would decay to zero as $n \to \infty$ ($\lim_{n\to\infty} P_{ij}^n = 0$). However, for any n, the sum of probabilities across the finite state space must be 1. $\sum_{j\in S} P_{ij}^n = 1$.

$$\lim_{n\to\infty} \sum_{j\in S} P_{ij}^n = \sum_{j\in S} \lim_{n\to\infty} P_{ij}^n = \sum_{j\in S} 0 \neq 1 \quad \Rightarrow\Leftarrow$$

(We could take the limit inside since S was finite). Hence, the chain is recurrent.
In finite spaces, Recurrence $\implies$ Positive recurrence
Fix some state $i \in S$ $(finite)$. Since the chain is irreducible, for any state j, $\exists\, n_j \in \mathbb{Z}\, s.t\ \ (P^{(n_j)})_{ji} > 0$. Let

$$N = \max_{j\in S} n_j \quad \text{and} \quad \varepsilon = \min_{j\in S} (P^N)_{ji} > 0$$

$\varepsilon > 0$ since the minimum is over a finite space S and irreducibility guarantees positivity. For any $k \geq 1$

$$P(T_i > kN \mid X_0 = i) \leq (1-\varepsilon)^k$$

At time N, no matter where the chain is, the probability of being at i is atleast ε from our definitions. $P(T_i > 2N \mid X_0 = i) = P(T_i > N \mid X_N)$ since we know that $T_i > N$ and at N, the state is $X_N (\neq i)$ and the markov chain does not depend on the past. So, it is a new block of length N, given X_N. Now, from any state j,

$$P(T_i \leq N \mid X_0 = j) \geq P(X_N = i \mid X_0 = j) = (P^N)_{ji} \geq \varepsilon$$

$$P(T_i > N \mid X_0 = j) \leq 1 - \varepsilon \quad \forall j$$

Putting it together,

$$P(T_i > 2N \mid X_0 = i) = E_i[\mathbf{1}_{\{T_i > N\}} P(T_i > N \mid X_N)] \leq E_i[\mathbf{1}_{\{T_i > N\}} (1-\varepsilon)]$$

$$= (1-\varepsilon) P(T_i > N \mid X_0 = i)$$

We can see that the RHS is $(1-\varepsilon)^2$. Similarly for $P(T_i > kN \mid X_0 = i) \leq (1-\varepsilon)^k$
Using,

$$E_i[T_i] = \sum_{n=0}^{\infty} P(T_i > n \mid X_0 = i)$$

$$E_i[T_i] \leq \sum_{k=0}^{\infty} N.P(T_i > kN \mid X_0 = i) \leq N \sum_{k=0}^{\infty} (1-\varepsilon)^k = \frac{N}{\varepsilon} < \infty$$

This shows that the expectation of retuning time is finite which proves positive recurrence.

□

Definition 5. *Consider a Markov chain X_t with transition probability matrix $P = (P_{ij}(t))$*

1. *Limiting Distribution: A vector $\{\pi_i\}_{i\in S}$ is called* limiting distribution *if* $\pi_i = \lim_{n\to\infty} P_{ji}^{(n)}, \qquad i, j \in S$ *(provided the limits exist) and* $\sum_{i\in S} \pi_i = 1$
2. *Stationary Distribution: A vector $\{\pi_i\}_{i\in S}$ is called a* stationary distribution *if* $\pi_i \geq 0$ *for all* $i \in S$, $\sum_{i\in S} \pi_i = 1$, *and* $\sum_{j\in S} \pi_j P_{ji} = \pi_i$ *for all* $i \in S$ *(i.e.,* $\boldsymbol{\pi} P = \boldsymbol{\pi}$*)*

For a CTMC, P_{ij} becomes $P_{ij}(t)$ and definitions still hold.

Theorem 4. *An irreducible Markov chain X_t on a finite state space S has a unique stationary distribution π.*

Proof. **Existence**

Consider the transition matrix P. A row vector π satisfying $\pi P = \pi$ is a left eigenvector of P with eigenvalue $\lambda = 1$. For any transition matrix P, the sum of each row is 1. Hence, $P.\mathbf{1} = \mathbf{1}$. $\lambda = 1$ is a right eigenvalue of P. It should also have a left eigenvalue of $\lambda = 1$ since a matrix and its transpose have same eigenvalues. Hence, $\exists v$ s.t $vP = v$.

The *Perron-Frobenius theorem* states that for any irreducible non-negative matrix, there is a largest real eigenvalue and it has a strictly positive eigenvector. Moreover, this eigenvalue is unique and dominant.

Now, we prove that $\lambda = 1$ is the largest eigenvalue for P. We have shown that $\lambda = 1$ is an eigenvalue with eigenvector $\mathbf{1}$. Let β be any eigenvalue of P with eigenvector $x \neq 0$. W.K.T $Px = \beta x$. Take the component i where $|x_i|$ is maximal.

$$|\beta|\,|x_i| = |(Px)_i| = \left|\sum_j p_{ij} x_j\right| \leq \sum_j p_{ij}|x_j| \leq \sum_j p_{ij}|x_i| = |x_i|.$$

$$|\beta| \leq 1$$

Hence, the largest eigenvalue is 1 and by the *Frobenius theorem*, it has a positive eigenvector ($v > 0$). We can now normalise this vector by setting $\pi_i = \frac{v_i}{\sum v_j}$ to ensure that $\sum \pi_i = 1$. This π is our stationary distribution. Therefore, $\exists \pi$ s.t $\pi P = \pi$

Uniqueness

Suppose there are two distinct stationary distributions, α and β, such that $\alpha P = \alpha$ and $\beta P = \beta$. Since the chain is irreducible, we know from *Perron-Frobenius* that α_i, β_i

$> 0 \; \forall i$. Consider the ratio $\frac{\alpha_i}{\beta_i} \quad \forall i \in S$. Since S is finite, there must be a state k where the ratio attains its minimum. Let $c = min_i\left(\frac{\alpha_i}{\beta_i}\right) = \frac{\alpha_k}{\beta_k}$
From $\alpha P = \alpha$, we can write the entry for state k:

$$\alpha_k = \sum_{j \in S} \alpha_j P_{jk}$$

Since $\alpha_j \geq c\beta_j \; \forall j$ (by our definition of c), we substitute:

$$\alpha_k = \sum_{j \in S} \alpha_j P_{jk} \geq \sum_{j \in S} (c\beta_j) P_{jk}$$

$$\alpha_k \geq c \sum_{j \in S} \beta_j P_{jk}$$

Because $\beta P = \beta$, we know $\sum \beta_j P_{jk} = \beta_k$. Thus:

$$\alpha_k \geq c\beta_k$$

For the equality $\alpha_k = c\beta_k$ to hold, it must be that $\forall j$ where $P_{jk} > 0$, the ratio α_j/β_j is exactly c. Because the chain is irreducible, we can reach any state from state k. Similarly, the ratio must be c for all states in the chain. If $\alpha_i/\beta_i = c \forall i$, then $\alpha = c\beta$. Since both α and β must sum to 1, $c = 1$. Therefore, $\alpha = \beta$. □

Corollary 4.1. *If a finite Markov chain is irreducible, then all states are positive recurrent, and it has a unique stationary distribution given by* $\pi_i = 1/E[T_i]$

Proof. From 3, we know that if a finite Markov Chain is irreducible, then it is positive recurrent. From 4, we can get that the irreducible markov chain has a unique stationary distribution $\pi P = \pi$.

Fix a state $i \in S$ and let

$$T_i = \inf\{n \geq 1 : X_n = i\}$$

be the first return time to i. Assume $X_0 = i$ and define, for each $j \in S$,

$$\rho_j = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_n = j\}} \right].$$

Clearly, $\rho_i = 1$, and

$$\sum_{j \in S} \rho_j = \mathbb{E}_i[T_i].$$

We show that ρ is an invariant measure. For any $k \in S$,

$$\sum_{j \in S} \rho_j P_{jk} = \sum_{j \in S} P_{jk}\, \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_n = j\}} \right] = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \sum_{j \in S} \mathbf{1}_{\{X_n = j\}} P_{jk} \right]$$

Now,

$$\sum_{j \in S} \mathbf{1}_{\{X_n = j\}} P_{jk} = \mathbb{E}[\mathbf{1}_{\{X_{n+1} = k\}} \mid X_n]$$

By the tower property of expectation, we get,

$$\mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \sum_{j \in S} \mathbf{1}_{\{X_n = j\}} P_{jk} \right] = \mathbb{E}_i \left[\sum_{n=0}^{T_i - 1} \mathbb{E}\big[\mathbf{1}_{\{X_{n+1} = k\}} \mid X_n\big] \right] = E_i \left[\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_{n+1} = k\}} \right]$$

Since T_i is the first return time,

$$\sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_{n+1} = k\}} = \sum_{n=0}^{T_i - 1} \mathbf{1}_{\{X_n = k\}} \quad \text{almost surely (since } X_0 = X_{T_i} = i)$$

$$\sum_{j \in S} \rho_j P_{jk} = \rho_k,$$

Hence, ρ is invariant (i.e., $\rho P = P$)

Since the chain is finite and irreducible, the stationary distribution is unique. Therefore, substituting and normalising, we get

$$\pi_j = \frac{\rho_j}{\sum_{k \in S} \rho_k} \quad \text{and hence} \quad \pi_i = \frac{\rho_i}{\mathbb{E}_i[T_i]} = \frac{1}{\mathbb{E}_i[T_i]}$$

□

Theorem 5. *If a finite Markov chain is irreducible and aperiodic, then there exists a unique stationary distribution π such that for all $i, j \in S$:*

$$\lim_{n \to \infty} (P^n)_{ij} = \pi_j$$

(In other words, if the Markov chain is ergodic on S, then the limiting probability distribution exists and is equal to the unique stationary distribution)

Proof. We consider two independent Markov chains, X_n and Y_n, both governed by the same transition matrix P. Define a new process $W_n = (X_n, Y_n)$ on the state space $S \times S$. The transition probability for this joint chain is:

$$P((i,k),(j,l)) = P_{ij} P_{kl}$$

If the original chain is irreducible and aperiodic, the joint chain (X_n, Y_n) is also irreducible.

Let T be the first time the two chains meet:

$$T = \min\{n \geq 0 : X_n = Y_n\}$$

Since the joint chain is irreducible on a finite state space, every state is visited infinitely often with probability 1. Therefore, the chains will eventually collide:

$$P(T < \infty) = 1$$

For $n > T$, we define a new process Z_n:

$$Z_n = \begin{cases} X_n & \text{if } n < T \\ Y_n & \text{if } n \geq T \end{cases}$$

Because of the Markov property, Z_n has the same transition probabilities as X_n. Now, consider the difference in probabilities for any state j:

$$|P(Z_n = j) - P(Y_n = j)|$$

Using the coupling inequality:

$$|P(Z_n = j) - P(Y_n = j)| \leq P(Z_n \neq Y_n) = P(T > n)$$

As $n \to \infty$, the probability $P(T > n) \to 0$ because the chains are guaranteed to meet. Therefore:

$$\lim_{n\to\infty} |P(Z_n = j) - P(Y_n = j)| = 0$$

If we set Y_0 to be distributed according to the stationary distribution π (which exists for irreducible finite chains), then Y_n follows π for all n. This proves that:

$$\lim_{n\to\infty} P(Z_n = j) = \pi_j$$

Now, WKT $P(Z_n = j) = P(X_n = j)$ from the construction of Z_n (Once the 2 chains meet, from thereon, both are equal adn have same transition probabilities).

$$P(X_n = j) = \sum_{i\in S} P(X_n = j \mid X_0 = i)P(X_0 = i)$$

$$P(X_n = j) = \sum_{i\in S} (P^n)_{ij} a_i \quad \text{where } a_i \text{ is the initial probability of starting in state } i$$

Here, $(P^n)_{ij}$ is nothing but an n-step transition ($P(X_n = j \mid X_0 = i)$. Since this holds

for any initial frequency a_i, we choose

$$a_i = \delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases}$$

So

$$P(X_n = j) = \sum_{i \in S} \delta_{ik} P_{ij}^n$$

$$P(X_n = j) = 1 \cdot P_{kj}^n$$

$$\lim_{n \to \infty} P_{kj}^n = \lim_{n \to \infty} P(X_n = j) = \pi_j$$

Since k was an arbitrary state, this holds for all $i, j \in S$. □

Remark 5. All of the above results for finite irreducible discrete-time Markov chains also hold for Continuous-time Markov chains (CTMCs) on a finite state space. In particular:

- *Recurrence and transience remain class properties.*
- *Every finite irreducible CTMC is positive recurrent.*
- *A finite irreducible CTMC has a unique stationary distribution π.*
- *If the CTMC is also aperiodic (automatic in continuous time), then the limiting probabilities exist and are equal to the unique stationary distribution:*

$$\lim_{t \to \infty} P_{ij}(t) = \pi_j, \quad i, j \in S.$$

These statements follow from the analogous DTMC results applied to the embedded jump chain of the CTMC, and will also be consistent with the generator Q when it is introduced below.

Definition 6 (Generator Matrix)**.** *A Continuous-time Markov chain on a countable state space S can be parameterized either by holding rates $\{\lambda_i\}_{i \in S}$ and transition probabilities of the jump chain $\{P_{ij}^{jump}\}_{i,j \in S}$, or equivalently by a matrix $Q = (q_{ij})_{i,j \in S}$, called the* generator matrix *(or* rate matrix*), defined by*

$$q_{ij} = \begin{cases} \lambda_i P_{ij}^{jump}, & i \neq j, \\ -\lambda_i, & i = j, \end{cases} \qquad i, j \in S.$$

Properties. The generator matrix Q satisfies the following properties.

- $q_{ij} \geq 0$ for all $i \neq j$, and $q_{ii} \leq 0$ for all $i \in S$.

- Each row of Q sums to zero:

$$\sum_{j \in S} q_{ij} = 0, \qquad i \in S.$$

Proof. By definition,

$$\sum_{j \in S} q_{ij} = q_{ii} + \sum_{j \neq i} q_{ij} = -\lambda_i + \lambda_i \sum_{j \neq i} P_{ij}^{jump}.$$

If $\lambda_i > 0$, then $P_{ii}^{jump} = 0$ and $\sum_{j \neq i} P_{ij}^{jump} = 1$. If $\lambda_i = 0$, then $q_{ij} = 0$ for all j.

- The total rate of leaving state i is given by

$$\lambda_i = \sum_{j \neq i} q_{ij}.$$

Proof.

$$\sum_{j \neq i} q_{ij} = \sum_{j \neq i} \lambda_i P_{ij}^{jump} = \lambda_i \sum_{j \neq i} P_{ij}^{jump}.$$

If $\lambda_i > 0$, then $P_{ii}^{jump} = 0$ $\sum_{j \neq i} P_{ij}^{jump} = 1$, and the result follows. If $\lambda_i = 0$, LHS=RHS=0.

- For $i \neq j$, the transition probabilities of the embedded jump chain are given by

$$P_{ij}^{jump} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}}, \qquad \text{whenever } \lambda_i > 0.$$

- For small $\Delta t > 0$, the transition probabilities satisfy

$$P_{ii}(\Delta t) = 1 + q_{ii}\Delta t + o(\Delta t), \qquad P_{ij}(\Delta t) = q_{ij}\Delta t + o(\Delta t), \quad i \neq j.$$

Proof. If $X(0) = i$, the holding time in state i is exponentially distributed with rate λ_i. Hence,

$$\mathbb{P}(X(\Delta t) = i \mid X(0) = i) = \mathbb{P}(T_i > \Delta t) = e^{-\lambda_i \Delta t} = 1 - \lambda_i \Delta t + o(\Delta t),$$

which gives $P_{ii}(\Delta t) = 1 + q_{ii}\Delta t + o(\Delta t)$ since $q_{ii} = -\lambda_i$. For $i \neq j$,

$$\mathbb{P}(X(\Delta t) = j \mid X(0) = i) = \mathbb{P}(T_i \leq \Delta t)\, p_{ij} = (\lambda_i \Delta t + o(\Delta t))\, p_{ij} = q_{ij}\Delta t + o(\Delta t).$$

Equivalently,

$$Q = \lim_{\Delta t \to 0^+} \frac{P(\Delta t) - I}{\Delta t}.$$

Theorem 6 (Kolmogorov Forward and Backward Equations)**.** *Let $\{X(t)\}_{t\geq 0}$ be a time-homogeneous continuous-time Markov chain with generator matrix $Q = (q_{ij})$ and transition probabilities $P_{ij}(t)$. Then, for all $i, j \in S$,*

$$\frac{d}{dt}P_{ij}(t) = \sum_{k\in S} P_{ik}(t)\, q_{kj} \quad \textit{and} \quad \frac{d}{dt}P_{ij}(t) = \sum_{k\in S} q_{ik}\, P_{kj}(t)$$

(forward equations and backward equations)

$$\textit{Hence,} \quad P'(t) = P(t)Q \qquad \textit{and} \qquad P'(t) = QP(t),$$

with initial condition $P(0) = I$.

Proof. We first prove the forward equations. Fix $i, j \in S$ and $t \geq 0$. By the Chapman–Kolmogorov equations,

$$P_{ij}(t+h) = \sum_{k\in S} P_{ik}(t)\, P_{kj}(h), \qquad h > 0.$$

Subtracting $P_{ij}(t)$ and dividing by h yields

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = \sum_{k\in S} P_{ik}(t)\, \frac{P_{kj}(h) - \delta_{kj}}{h},$$

where δ_{kj} is the Kronecker delta.

Taking the limit as $h \to 0^+$ and using the definition of the generator,

$$q_{kj} = \lim_{h\to 0^+} \frac{P_{kj}(h) - \delta_{kj}}{h},$$

we obtain

$$\frac{d}{dt}P_{ij}(t) = \sum_{k\in S} P_{ik}(t)\, q_{kj},$$

which proves the forward equations.

The backward equations are proved similarly. Using again the Chapman–Kolmogorov equations,

$$P_{ij}(t+h) = \sum_{k\in S} P_{ik}(h)\, P_{kj}(t).$$

Subtracting $P_{ij}(t)$, dividing by h, and letting $h \to 0^+$ gives

$$\frac{d}{dt}P_{ij}(t) = \sum_{k\in S} q_{ik}\, P_{kj}(t),$$

where we have again used the definition of the generator.

Writing these equations in matrix form yields

$$P'(t) = P(t)Q \qquad \text{and} \qquad P'(t) = QP(t),$$

with $P(0) = I$, completing the proof. □

Remark 6. The stationary distribution π described above satisfies $\pi Q = 0$. This is straightforward from the definition of the stationary distribution and the previous theorem 6.

A.2 What we need for DNA models

We now apply the general theory of Continuous-time Markov chains to the specific case of molecular evolution. We model the evolution of a single nucleotide site as a stochastic process on a finite state space $\mathcal{S}_{DNA} = \{A, C, G, T\}$.

Biologically, it is possible for any nucleotide base to mutate into any other base, given sufficient time (either directly or via intermediate steps). Consequently, the rate matrix for DNA evolution is strictly irreducible. Applying Theorem 3, since is finite and the chain is irreducible, the process is positive recurrent. This guarantees that a nucleotide site will return to any given state infinitely often over an infinite evolutionary timescale, with a finite expected return time $\mathbb{E}[T_i] < \infty$. Following Theorem 4 and Theorem 5, the irreducibility of this finite chain guarantees the existence of a unique stationary distribution $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ such that $\pi Q = 0$. In the biological context, this vector represents the equilibrium base frequencies of the genome. As $t \to \infty$, the probability of finding the site in state j converges to π_j regardless of the ancestral state: $\lim_{t\to\infty} P_{ij}(t) = \pi_j$

This convergence justifies the use of π as the prior probability of the root state in phylogenetic likelihood calculations, assuming the process has been running long enough to reach equilibrium. The evolutionary dynamics are fully characterized by the 4×4 generator matrix Q. While the general theory allows for arbitrary transition rates, biological models impose specific structures on. (For a detailed discussion on the specific biological assumptions, refer Section 1.2.2).

Appendix B

Supplementary Figures

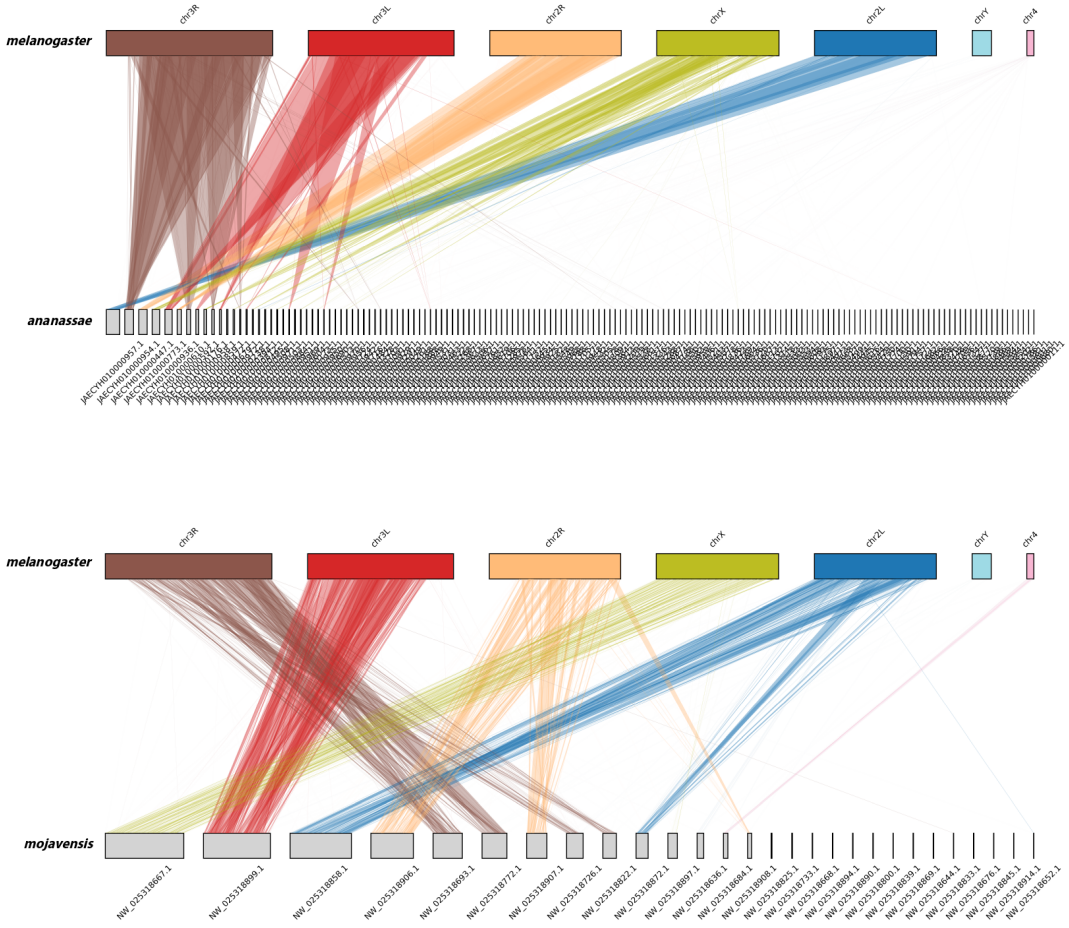

Figure B.1: Synteny Plots of *D. melanogaster* vs *D. ananassae* (Top) and *D. melanogaster* vs *D. mojavensis* (Bottom).

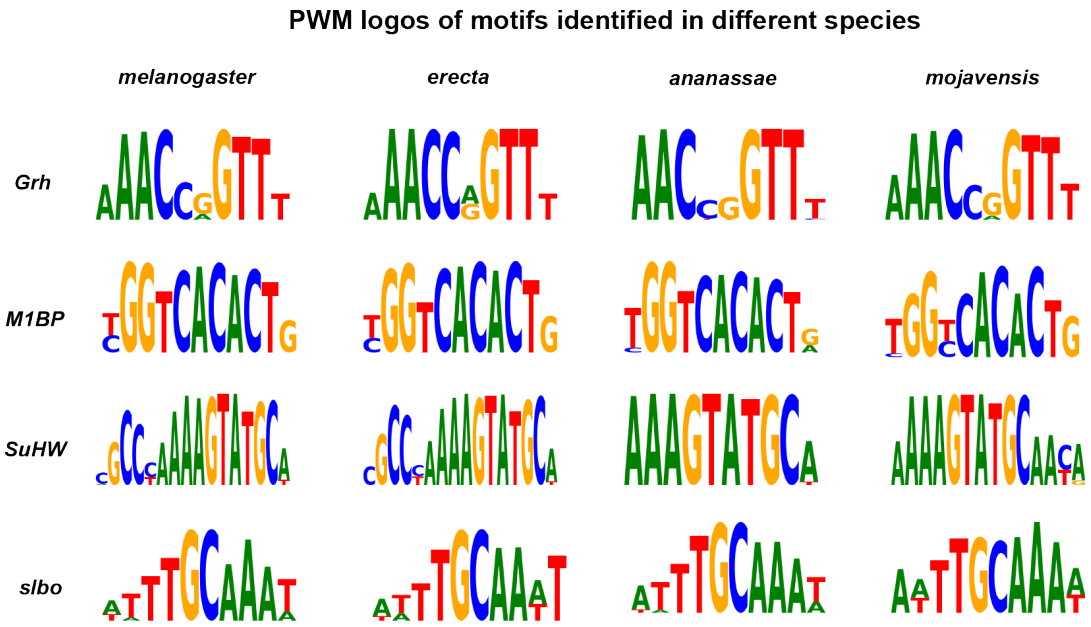

Figure B.2: PWM logos across species.

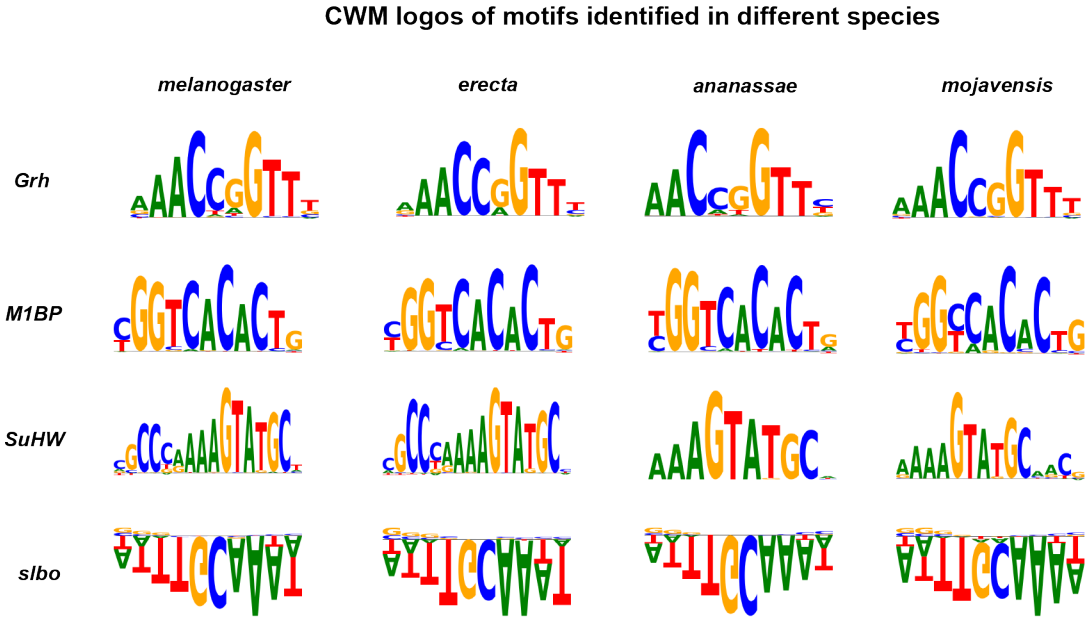

Figure B.3: CWM logos across species.

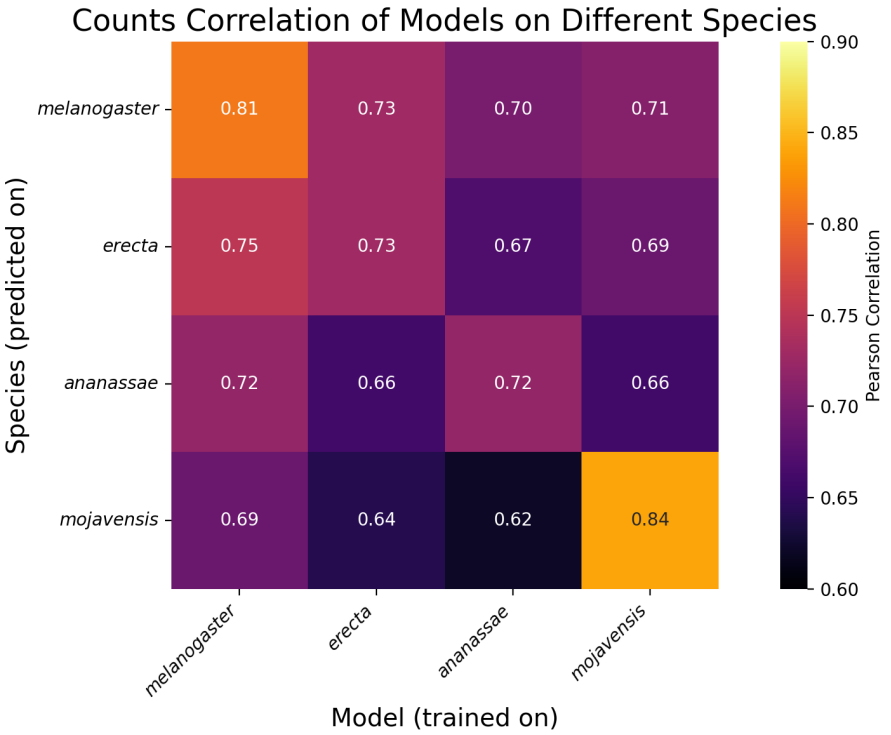

Figure B.4: Cross species prediction matrix - All species over all species

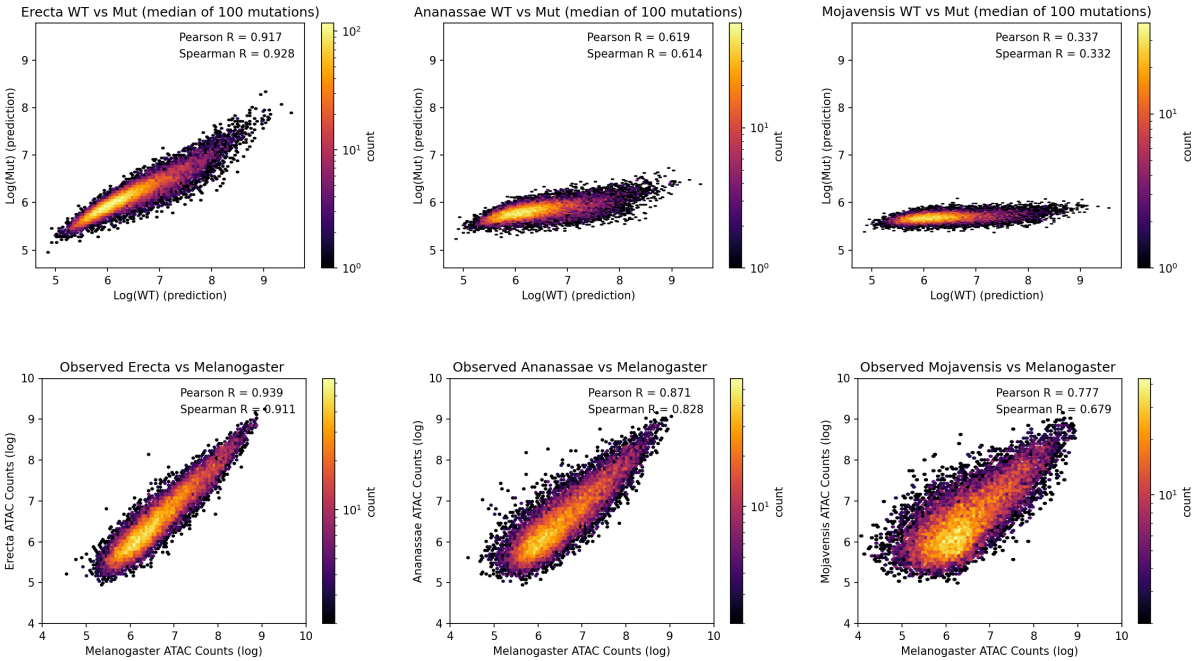

Figure B.5: Top: Correlation of ATAC signal of synthetically mutated sequences vs wildtype sequences (Predictions). Bottom: Correlation of ATAC counts of orthologous regions for different species vs *D. melanogaster*.

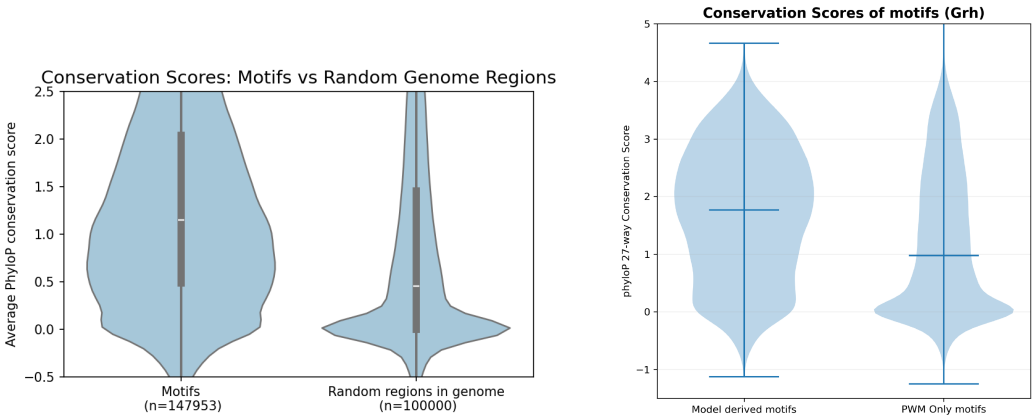

Figure B.6: Left: Conservation scores of motifs. Right: Conservation scores of model derived *Grh* motif instances vs PWM matched *Grh* motif instances.

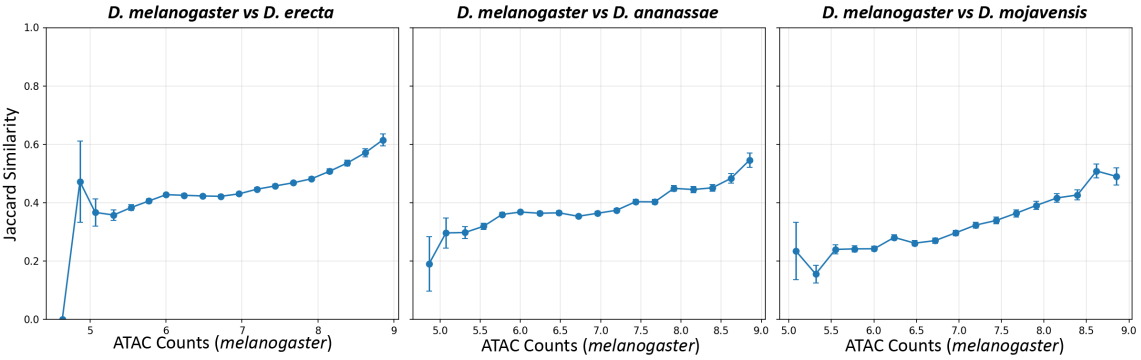

Figure B.7: Jaccard similarity vs ATAC counts across species.

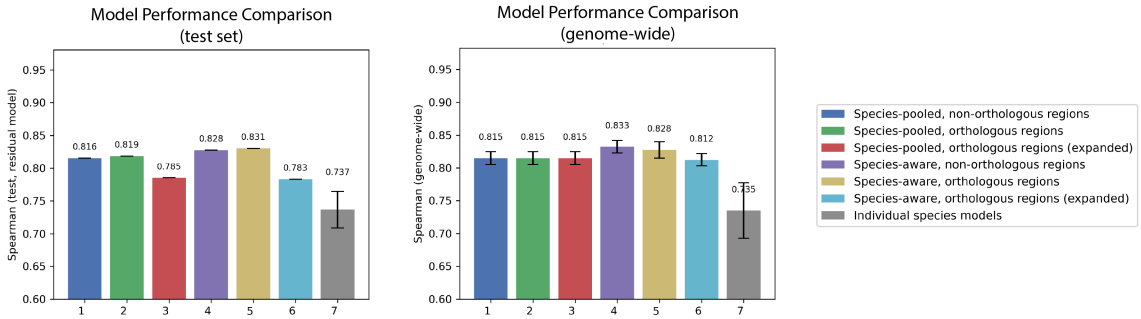

Figure B.8: Comparison of multi species model performance - Spearman correlation.

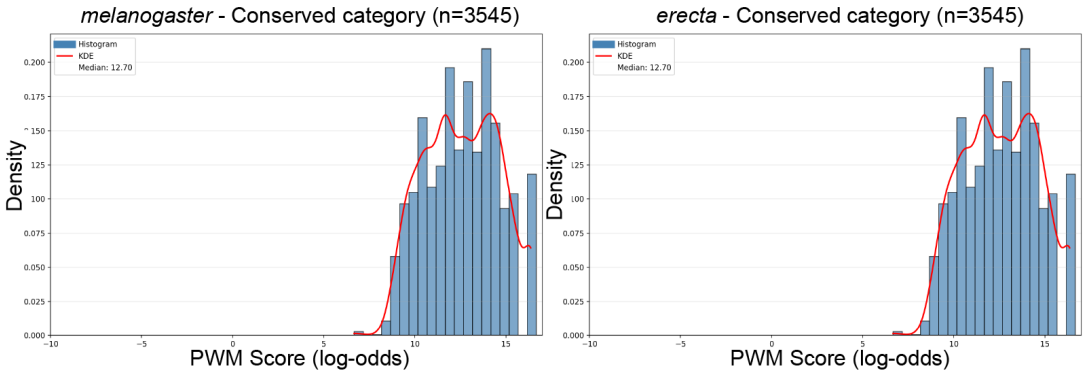

Figure B.9: PWM distribution of *Grh* instances in the Conserved category.

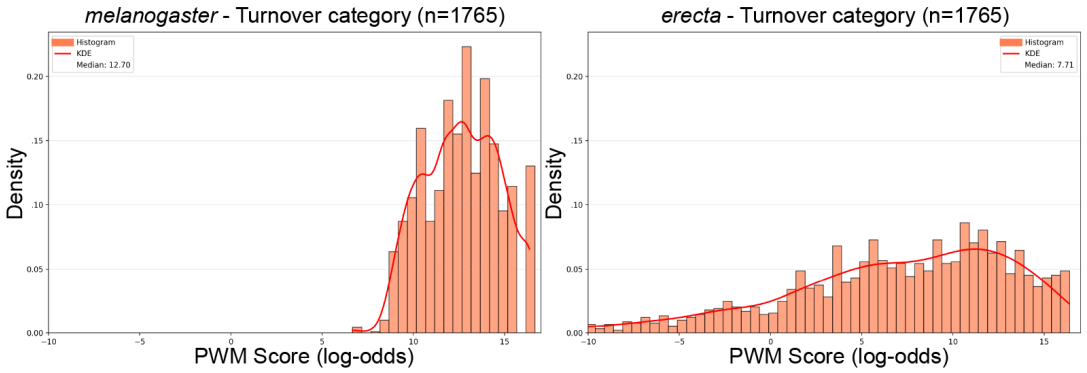

Figure B.10: PWM distribution of *Grh* instances in the Turnover category.

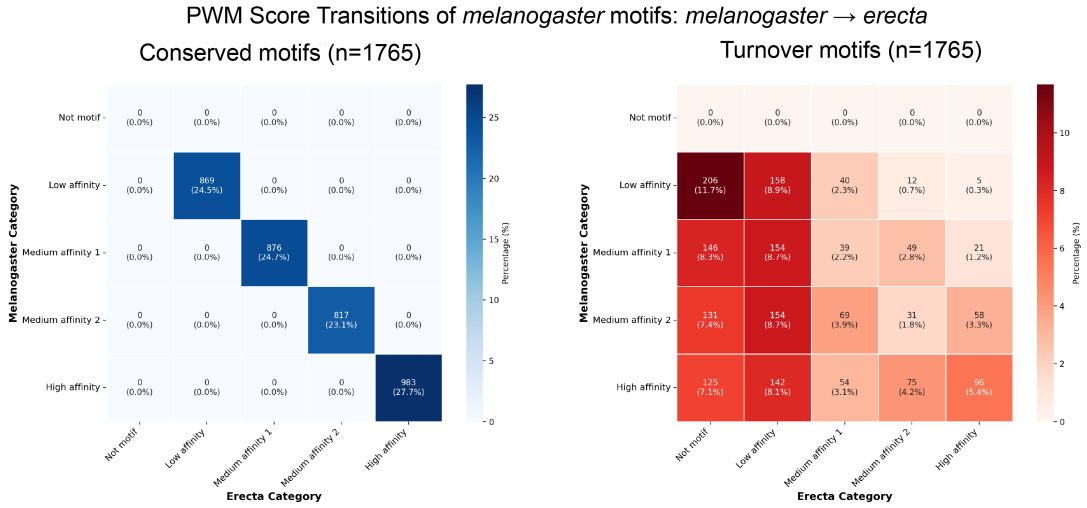

Figure B.11: Transition probability matrix for *Grh* motif instances in Conserved and Turnover category