Comparative and population genomics analyses of TF-DNA interactions

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Manas Joshi
aus Pune, Indien

Köln, April 2023

Gutachter und Prüfer

1. Prof Dr Miltos Tsiantis
2. Prof Dr Thomas Wiehe

Vorsitzender der Prüfungskommission

Prof Dr Andreas Beyer

Beisitzender

Dr Stefan Laurent

Abstract

Molecular biology provides a unique insight into the workings of the evolutionary forces present in nature. As opposed to comparative anatomy, which relies on the structural makeup of species, molecular biology relies on the information contained within the biochemical makeup of the species. One of the salient features of a molecular evolution-based approach is that it relies on observations drawn from the changes occurring within the biomolecules for making inferences on the evolutionary forces in action. The overall makeup of such biomolecules is consistent across species, thus allowing for robust and comparable inferences across distantly related species. In many aspects, these biomolecules could be thought to carry the imprints of evolutionary forces. At the centre of these biomolecules is the DNA molecule, through which the necessary information on the species-specific traits is passed down from the parent generation to the offspring generation. DNA, in the form of genes, also codes for a specialized class of biomolecules, proteins, which are responsible for many functions within the cell, ranging from regulating pathways, aiding in response to pathogens, and controlling the expression of other genes. This transition of genes to proteins is tightly controlled by regulatory machinery that ensures the context-dependent activation of the genes and, consequently, the production of proteins. Hence, natural variants occurring within these regulatory elements could result in differential gene expression patterns and, potentially, alter the transition of genotype to phenotype. Given the central role of this regulatory machinery, it would be expected to be under a stronger influence of the evolutionary forces as compared to the genomic background.

This study focused on understanding the impact of a specific evolutionary force, natural selection, on the gene regulatory elements through the perspective of the genetic variants occurring within them. Natural selection could be perceived as a force that confers fitness advantages to individuals based on their genotypic and phenotypic makeup. This study was specifically aimed at understanding the action of natural selection on the regulatory transcription factor (TF)- DNA interactions. These regulatory interactions have two motifs: DNA-binding domains (DNABDs), occurring on the TFs, and the Transcription Factor binding sites (TFBSs), occurring on the DNA molecules. The central aim of the study was to elucidate the action of negative/purifying and positive selection acting on these domains through a comparative framework. This study combined population and comparative genomics approaches to quantify the intensity of natural selection acting across two different evolutionary time scales.

In the case of the DNABDs, we identified a signal of high constraint consistent across the evolutionary time scales and irrespective of the genomic control regions included in this study. This observation indicated

that DNABDs are under an increased intensity of purifying selection, which could be explained by the pleiotropic nature of the TFs. However, we do not observe similar trends when investigating the action of positive selection. Specifically, the intensity of positive selection was observed to be comparatively high for the DNABD regions only in certain populations of species with larger effective population sizes (N_e).

In the case of the TFBSs, given that they are primarily a part of the noncoding genome, we developed a summary statistic to quantify the intensity of natural selection that would also be comparable to the summary statistic from the coding regions. On comparing the summary statistics from the coding and noncoding regions, we identify the signal of a comparatively relaxed constraint acting on the TFBS regions compared to the DNABD and other control regions. In addition, we also highlight that, overall, the TFBS are under a reduced influence of positive selection. The signal of reduced constraint and a decreased intensity of positive selection was consistent across the two evolutionary time scales.

Overall, by exploring the intensities of selection on the DNABDs and the TFBSs, this study contributes to our understanding of the impact of natural selection acting on the regulatory elements across coding and noncoding regions.

Zusammenfassung

Die Molekularbiologie bietet einen einzigartigen Einblick in die Funktionsweise der in der Natur vorhandenen evolutionären Kräfte. Im Gegensatz zur vergleichenden Anatomie, die sich auf den strukturellen Aufbau der Arten stützt, stützt sich die Molekularbiologie auf die im biochemischen Aufbau der Arten enthaltenen Informationen. Eines der hervorstechenden Merkmale eines auf der molekularen Evolution basierenden Ansatzes ist, dass er sich auf Beobachtungen stützt, die aus den Veränderungen innerhalb der Biomoleküle gezogen werden, um Rückschlüsse auf die wirkenden evolutionären Kräfte zu ziehen. Der Gesamtaufbau (allgemeine Aufbau, grundlegende Aufbau) solcher Biomoleküle ist bei allen Arten gleich, so dass robuste und vergleichbare Schlussfolgerungen auch bei weit voneinander entfernten Arten möglich sind. In vielerlei Hinsicht könnte man annehmen, dass diese Biomoleküle Abdrücke evolutionärer Kräfte tragen. Im Zentrum dieser Biomoleküle steht das DNA-Molekül, durch das die notwendigen Informationen über die artspezifischen Merkmale von der Elterngeneration an die Nachkommengeneration weitergegeben werden. DNA kodiert in Form von Genen auch für eine spezielle Klasse von Biomolekülen, Proteine, die für zahlreiche Funktionen innerhalb der Zelle verantwortlich sind, von der Regulierung von Stoffwechselwegen über die Reaktion auf Krankheitserreger bis hin zur Expressionskontrolle anderer Gene. Die Umwandlung von Genen in Proteine wird streng von einem Regelungsapparat kontrolliert, der die kontextabhängige Aktivierung der Gene und folglich die Produktion von Proteinen gewährleistet. Natürliche Varianten innerhalb dieser regulatorischen Elemente könnten daher zu unterschiedlichen Genexpressionsmustern führen und möglicherweise den Übergang vom Genotyp zum Phänotyp verändern. In Anbetracht der zentralen Rolle dieses Regelungsapparats ist zu erwarten, dass er im Vergleich zum genomischen Hintergrund einem stärkeren Einfluss der evolutionären Kräfte ausgesetzt ist.

Diese Studie konzentrierte sich darauf, die Auswirkungen einer bestimmten evolutionären Kraft, der natürlichen Selektion, auf die genregulatorischen Elemente aus der Perspektive der in ihnen vorkommenden genetischen Varianten zu verstehen. Die natürliche Selektion kann als eine Kraft angesehen werden, die Individuen aufgrund ihrer genotypischen und phänotypischen Ausstattung Fitnessvorteile verschafft. Diese Studie zielte speziell darauf ab, die Wirkung der natürlichen Selektion auf die regulatorischen Interaktionen zwischen Transkriptionsfaktoren (TF) und DNA zu verstehen. Diese regulatorischen Interaktionen haben zwei Motive: DNABDs (DNA-binding domains), die auf den TFs vorkommen, und die TFBSs (Transcription Factor binding sites), die auf den DNA-Molekülen vorkommen. Das Hauptziel der Studie war es, die Wirkung negativer/reinigender und positiver Selektion, die auf diese Domänen einwirken, durch

einen vergleichendens Framework zu erhellen. In dieser Studie wurden Ansaetze der Populationsgenomik und der komparativen Genomik kombiniert, um die Intensität der natürlichen Selektion zu quantifizieren, die über zwei verschiedene evolutionäre Zeitskalen hinweg wirkt.

Im Fall der DNABDs konnten wir ein Signal für eine starke Einschränkung feststellen, welches über die evolutionären Zeitskalen hinweg und unabhängig von den in dieser Studie einbezogenen genomischen Kontrollregionen konsistent ist. Diese Beobachtung deutet darauf hin, dass DNABDs einer verstärkten reinigenden Selektion ausgesetzt sind, was durch die pleiotrope Natur der TFs erklärt werden könnte. Bei der Untersuchung der Wirkung positiver Selektion konnten wir jedoch keine ähnlichen Trends beobachten. Insbesondere wurde beobachtet, dass die Intensität der positiven Selektion für die DNABD-Regionen nur bei Arten mit größeren effektiven Populationsgrößen (N_e) verhältnismäßig hoch ist.

Da die TFBS in erster Linie Teil des nicht kodierenden Genoms sind, entwickelten wir eine zusammenfassende Statistik zur Quantifizierung der Intensität der natürlichen Selektion, die auch mit der zusammenfassenden Statistik der kodierenden Regionen vergleichbar sein sollte. Beim Vergleich der zusammenfassenden Statistiken der kodierenden und nicht-kodierenden Regionen stellen wir fest, dass in den TFBS-Regionen im Vergleich zu den DNABD- und anderen Kontrollregionen eine vergleichsweise geringere Einschränkung herrscht. Darüber hinaus stellen wir fest, dass die TFBS insgesamt einem geringeren Einfluss positiver Selektion ausgesetzt sind. Das Signal einer geringeren Einschränkung und einer geringeren Intensität der positiven Selektion war über die beiden evolutionären Zeitskalen hinweg konsistent.

Durch die Untersuchung der Intensität der Selektion auf die DNABDs und die TFBSs lieferte diese Studie ein tiefes Verständnis der Auswirkungen der natürlichen Selektion auf die regulatorischen Elemente in kodierenden und nicht-kodierenden Regionen.

Table of contents

General Introduction

Natural selection and Molecular evolution

One of the many staggering observations made in biology is the wide variety of phenotypic differences in the species inhabiting this planet. In some cases, subtle phenotypic differences could also be observed in individuals belonging to the same species but originating from different populations. Evolutionary biology sets out to answer two critical questions regarding these phenotypic differences: *Why do these differences occur? How do these differences occur?* The former question could be answered through the perspective of causality. Specifically, changes occurring within the species are somehow warranted by the changes occurring within the species' surroundings. These changes are mainly driven by the fitness cost attached to them. One of the predominant drivers of such phenotypic changes would be natural selection. This evolutionary force rewards beneficial phenotypes and penalizes detrimental phenotypes, i.e., individuals having beneficial phenotypes would have a reproductive advantage. To address the question of '*How do these differences occur?*', one needs to look closely at the background genotype and the transition from genotype to phenotype. In this aspect, molecular biology provides an interesting perspective in understanding the evolutionary forces that shape phenotypic differences. Biomolecules could be perceived as elements constructed from building blocks that are, overall, consistent across different species. Subtle differences introduced in these building blocks could affect the resulting phenotype. Hence, differences in the genotypes could help in understanding the observed phenotypic differences between species.

Based on the magnitude of information contained within them, DNA molecules could be perceived as an important class of biomolecules within the biochemical machinery of a cell. Primarily, these act as a blueprint containing information on the species-specific traits. Hence, these molecules are considered a key hereditary material responsible for transferring these species-specific attributes from a given generation to the next generation. The DNA molecules contain stretches of sequences, referred to as genes, that encode for a specialized class of biomolecules, proteins. These protein molecules play varying roles within the cell, ranging from assisting in biochemical pathways, aiding in host immune response, supporting the cell structure, etc.(Alberts et al. 2002). In addition, a specialized class of proteins, transcription factors (TFs), are also responsible for controlling the expression of other genes. The conversion of genes to proteins encapsulates the flow of information, which impacts multiple downstream biological processes. Hence, the expression of genes is kept under the control of regulatory machinery responsible for the context-dependent switching-on and -off of the expression of genes. The seminal work by Jacob and Monod (Jacob and Monod 1978) was one of the first to highlight the role of these gene regulatory elements (GREs) in controlling gene expression. King and Wilson (King and Wilson 1975) further highlighted that the variants occurring within such GREs could result in differential gene expression patterns and alter the transition of genotype to phenotype.

Impact of natural selection on the GREs

Given the critical role of GREs, they would be expected to be under the stringent control of natural selection. Several studies have investigated the impact of natural selection acting on the GREs through the perspective of genetic variants occurring within them. Many studies have investigated the impact of negative/purifying selection on the GREs. To exemplify, (Mu et al. 2011) highlighted that, in the case of *Homo sapiens*, the Transcription Factor binding sites (TFBS) are under a stronger influence of negative selection than the control regions, which did not exhibit TF binding activity. This study highlighted that TFBS showed a reduced diversity within populations and reduced "fixed" differences with a genetically neighbouring species, *Pan troglodytes*. A signal of reduced genetic diversity within the TFBS regions was also reported by (Vernot et al. 2012). Similar results of a reduced genetic diversity within the TFBS regions were also reported for *Saccharomyces cerevisiae* by (Connelly et al. 2013).

These studies suggest that the GREs host limited genetic variation. However, multiple studies have highlighted instances of selective sweeps wherein naturally occurring beneficial variants within GREs rapidly increased in frequency within a population of species and were ultimately fixed (Schlenke and Begun 2004; Chan et al. 2010; Enattah et al. 2002). In general, elucidating the action of positive selection on individual genomic elements is challenging because the proportion of naturally occurring beneficial mutations is mostly lower than that of deleterious mutations (Barghi, Hermisson, and Schlötterer 2020). Hence, most studies focus on aggregating the signal of positive selection from multiple loci. To exemplify, (Vernot et al. 2012) highlighted that the genes participating in the pigmentation pathway within the European population of *H. sapiens* display a signature of positive selection. Perdomo-Sabogal and Nowick (Perdomo-Sabogal, Nowick, and Enard 2019) highlighted that the *KRAB-ZNF* group of genes are under an increased intensity of positive selection within multiple *H. sapiens* populations. Studies have also shown that the GREs are under the shared influence of both positive and negative selection. Here, negative selection is responsible for maintaining the existing regulatory elements, and positive selection is responsible for the species-specific gain or loss of these regulatory elements (Haddrill, Bachtrog, and Andolfatto 2008; He et al. 2011).

Comparative and population genomics-based framework

Naturally occurring genomic variants originate from random genetic mutations. The overall evolutionary trajectory of species is decided by the fitness effect of such genomic variants in the context of the surrounding environment. Most of these variants do not directly influence the fitness of species, and these are referred to as neutral variants. The change in frequency of neutral variants is usually governed by random genetic drift, with exceptions occurring in scenarios when they are "linked" to non-neutral variants. However, a small proportion of the naturally occurring variants directly influence the fitness of species, referred to as non-neutral variants, and are subjected to selection. Variants detrimental to fitness are kept at lower frequencies within a population and are eventually "lost", whereas variants beneficial for fitness are maintained at higher frequencies within a population and eventually get "fixed". Hence, beneficial variants contribute less to within-species differences and more towards between-species differences. Comparative genomics studies use these between-species differences to highlight the action of selection on homologous regions between two or more species. This approach has been used extensively in the recent past to highlight genomic sequences potentially under the influence of selection across phylogeny. A possible caveat in such studies is that in terms of evolutionary time scale, the multi-species comparison highlights the signal of selection over comparatively longer time scales compared to within-species individual comparisons. Hence, the multi-species comparison approach will not be able to detect elements that have undergone selection in recent times. In the recent past, due to advances in sequencing technologies, performing ultra-deep sampling of populations within species has become feasible. Such an ultra-deep sampling approach enables the identification of the elements that have recently undergone selection and also helps identify population-specific genomic elements that could be under selection and potentially contribute to local adaptation. Hence, as has been pointed out by Lawrie and Petrov (Lawrie and Petrov 2014), complementing between-species variation data with the within-species population-specific variation data could prove to be a more potent approach in identifying functional elements that are under the action of selection.

Impact of the effective population size (N_e) on the intensity of selection

The action of selection strongly governs the rate of adaptation of species to environmental changes, wherein selection is expected to remove detrimental mutations and promote the beneficial mutations that would aid in adaptation. However, mutations that are not under the action of selection are expected to fluctuate in frequency randomly under the action of genetic drift. Hence, the forces of selection and genetic drift act in tandem in the natural populations. One of the factors hypothesized to influence the intensities of these two forces in the species-specific effective population size (N_e) (Eyre-Walker and Keightley 2007.; Galtier 2016; James, Castellano, and Eyre-Walker 2016). Specifically, an increase in N_e would result in an increase in the intensity of selection and a reduced intensity of random genetic drift. Several findings have corroborated this hypothesis. To exemplify, studies have shown that the proportion of naturally occurring beneficial mutations is lower in low N_e species (James, Castellano, and Eyre-Walker 2016; Eyre-Walker and Keightley 2007) and higher in large N_e species (Eyre-Walker and Keightley 2007; Andolfatto 2005). As stated in (Galtier 2016a), there are two main reasons for large N_e species exhibiting an increased intensity of adaptive evolution: the probability of beneficial mutations is directly proportional to the number of individuals within the population, and the chances of the fixation of a beneficial mutation are greater in high N_e species as compared to low N_e species where genetic drift is the predominant force. Consequently, the probability of purging deleterious variants would be higher in large N_e species.

Scope of this study

One of the central aims of this study was to identify the impact of natural selection on the GREs across different species. At the centre of this study were the regulatory TF-DNA interactions. We specifically focused on the interacting motifs in the TF-DNA interactions: DNA-binding domains (DNABDs), occurring on the TFs, and TFBS, occurring on the non-coding DNA. We use summary statistics to contrast the intensities of positive and negative/purifying selection acting on these GREs against different genomic control regions. Additionally, this study employed a comparative- and population-genomics approach to understand the impact of natural selection across different evolutionary time scales. This study spanned three species, namely – *Homo sapiens*, *Arabidopsis thaliana* and *Drosophila melanogaster*, and eight populations belonging to these species.

Chapter 1 – Elucidating the action of natural selection on the DNA-binding domains (DNABDs)

Abstract

Transcription Factors (TFs) are an essential element within the biochemical machinery of a cell. They are usually responsible for controlling expression patterns of multiple downstream effector gene(s). This pleiotropic characteristic makes them an essential element in the gene regulatory network of species. The function of gene regulation is primarily performed via a set of regulatory domains integrated within the protein structure of TFs. Given their direct involvement in gene regulation, this study is focused on understanding the action of natural selection acting on these domains through a comparative framework. Specifically, this study integrated data from multiple species and their respective populations to analyze the action of selection acting on these regulatory elements on two evolutionary timescales. To better quantify the action of selection acting on these regulatory domains, we compare this signal against the selective pressures acting on multiple genomic control regions. We note a consistent signal of high constraint acting on these regulatory sequences across all species and populations, as compared to and irrespective of the control regions. This observation suggests that these regions are predominantly under the action of selection which weeds out non-beneficial and potentially deleterious alleles. In addition to the action of purifying selection, we also note an excess of beneficial alleles in these regulatory regions in high N_e species as compared to the control regions. These observations combined suggest that the regulatory regions in TFs are under strict action of selection and are under the influence of both positive and negative selection.

Introduction

Gene regulation is an important process that contributes to the overall molecular evolution of species. This process mainly encapsulates the context-dependent switching-on and switching-off of the expression of genes that are relevant in terms of the effective phenotype. Gene regulatory factors (GRFs) are a class of genomic elements responsible for this vital process. These GRFs could be broadly classified into two categories based on their mode of action: cis–acting elements (CREs) and *trans*-acting elements (TREs).

CREs are mainly non-coding regions of the genome that regulate the gene expression of neighbouring (and, at times, distal) genes. On the other hand, TREs are usually factors derived from coding regions of the genome that participate in the regulatory action by binding directly to the CREs. These *cis* and *trans* elements are usually acting in conjunction with each other. Given the functional importance of these domains, these are expected to be under a comparatively stronger influence of natural selection.

This chapter will solely focus on elucidating and quantifying the effects of natural selection on the TREs, specifically focusing on an essential subset of TREs – Transcription Factors (TFs). TFs often serve various functions ranging from aiding cell differentiation and development (Ihn Lee and Young 2013) to controlling various biological pathways (Desvergne, Michalik, and Wahli 2006). TFs perform these functions by binding to the effector genes in a sequence-specific manner. Individual TFs have been documented to be regulating multiple target genes. Consequently, it is hypothesized that, given their pleiotropic nature, TFs could be experiencing stronger levels of purifying selection (Chesmore et al. 2016). On the other hand, mutations occurring within TFs could potentially promote or disrupt multiple regulatory interactions, which could aid in local adaptation. TFs undertake their regulatory function by binding to the effector gene(s) in a sequence-specific manner through specific functional domains within the TFs. Hence the TF-effector gene interactions have two interacting motifs: functional domains on TFs – DNA-binding domains (DNABD), and the stretch of sequence located in the proximity (in some instances also distally) of the effector genes – Transcription Factor Binding Sites (TFBS). In the context of their modes of action, TFBSs could be categorized as CREs. On the other hand, DNABDs are an essential element within TFs that facilitate the interaction of CREs and TREs. Given their direct involvement in gene regulation, functional genomics studies have attempted to characterize the action of selection acting specifically on these two motifs.

We perform a selection-based analysis of these domains from a comparative genomics perspective to gain insights into the action of selection acting on the regulatory DNABDs. Here, we infer the action of selection acting on the genomic regions using basic summary statistics widely used in evolutionary biology, namely – π_n/π_s, $\pi_{nonsense}/\pi_s$, K_n/K_s, $K_{nonsense}/K_s$ (here referred to as constraint ratios) and α (proportion of adaptive substitutions). In order to compare and contrast the action of selection acting on DNABDs, we select two genomic control classes, namely – non-DNABDs and WGS. The former is a class of functionally unannotated sequences within the coding sequences of TFs, while the latter is the entire set of protein-coding genes per species. The proximity of DNABD and non-DNABD, in terms of their genomic location, counters for differential recombination rates influencing the comparison of these two classes. On the other hand, comparing DNABD with WGS enables to perceive the signal of selection acting on the DNABDs in the context of the entire protein coding gene set per species. We employ a multi-species approach, which enables understanding if the perceived signal of natural selection is consistent across multiple species, irrespective of their differing evolutionary trajectories. For this purpose, the choice of species included in this study was made on two parameters – genetic dissimilarity amongst each other and data availability. Consequently, the species included in this study were - *Homo sapiens*, *Arabidopsis thaliana* and *Drosophila melanogaster*. In addition to a multi-species approach, we further investigated the intensity of the action of selection on different evolutionary timescales per species. Specifically, we employed an intra- and inter-species approach to ascertain if the observed action of selection is consistent across different species and on shorter and longer evolutionary timescales. Hence, we included a total of eight populations spanning the three species included in this study. To further validate if the action of selection differs based on differing local conditions, when possible, we chose populations located in differing geographical locations per species. We highlight an overall strong signal of purifying selection acting on DNABDs compared to the control regions across all species and evolutionary timescales. Given the functional importance of the DNABDs, this observation is consistent with the expectation. We further supplement this high constraint by highlighting the proportions of annotated deleterious variants within the DNABDs compared to the non-DNABDs for *H. sapiens*. On the scale of positive selection, we observe an overall high α for the DNABDs compared to both control regions for high N_e species (*A. thaliana* and *D. melanogaster*), with some exceptions. By observing the signals of purifying and positive selection in conjunction, we also highlight that the overall efficiency of selection acting on the genomic regions increases with an increase in the species-specific effective population size (N_e), which is often used as a proxy for the actual population size of species. This result is in agreement with observations from previously reported studies (Galtier 2016a; James, Castellano, and Eyre-Walker 2016). The correlation between the efficiency of selection and N_e was the poorest in the non-DNABD regions, which are the functionally unannotated class of the genomic region.

Materials and Methods

Dataset of Transcription Factors

The species-specific lists of Transcription Factors (TFs) were constructed based on two important criteria. First, every species-specific gene included in the dataset had to have a corresponding manually annotated UniProt (UniProt Consortium 2022) identifier (UniProtKB/SwissProt), hence genes having a computationally annotated identifier (UniProtKB/TrEMBL proteins) were automatically filtered out of the list of genes. For every gene, a transcript that had a corresponding UniProtKB/SwissProt identifier was chosen as the representative transcript. Next, we filtered for genes that have at least one regulatory domain. Specifically, for every UniProtKB/SwissProt gene, we extract information on all the annotated functional domains as per the ProRule (Sigrist et al. 2002) annotation schema. Following this, all the genes that had at least one functional domain, which was annotated with either of the following Gene Ontology (GO) terms – "DNA binding", "Transcriptional regulatory", or "Transcriptional activity", were retained. The species-specific lists of TFs used in this study are thus comprised of these retained genes. The regulatory genomic regions annotated with the above-mentioned GO terms were of interest in this study and constituted the "test regions" in the analyses. For this study, these regions were termed DNA-binding domains (DNABDs). Besides the domains identified with the above-mentioned GO terms, we also identified domains within the coding region of TFs that were annotated with other (non-DNA binding) GO terms. Hence, as "control regions", we used the coding region of TFs that were not functionally annotated with a binding domain of any type. For this study, these regions were termed non-DNA-binding domains (non-DNABDs). Using "control" and "test" regions within the same coding region also controls for heterogeneous recombination rates, influencing the signature of selection along chromosomes. The number of TFs per species and corresponding DNABDs included in this study are highlighted in **Table 1.1**.

Species	No. of TFs included in this study	No. of DNABDs identified
H. sapiens	886	1198
A. thaliana	861	1030
D. melanogaster	217	325

Table 1.1 - Summary of the number of species-specific TFs and the corresponding DNA-binding domains included in this study

Polymorphism and Divergence information

This study employed an intra- and inter-species approach (**Figure 1.1**). Specifically, to study the intra-species variation, the study comprised eight different populations for the three species included in this study. The information on the population-specific variants was obtained from the population-specific variant call format (*.vcf*) files. This information was used to capture the action of selection on a relatively shorter evolutionary timescale. On the other hand, we performed a per-gene transcript-specific orthology search to study inter-species variation using a reciprocal blast search approach. Specifically, we identified the orthologous gene in the outgroup species' genome for every transcript sequence using *blastn* (Altschul et al. 1990) based on a 60% identity filter (Uricchio, Petrov, and Enard 2019). The obtained transcripts of the ingroup and outgroup species were further aligned with MUSCLE (Edgar 2004), and this alignment was used to estimate divergence.

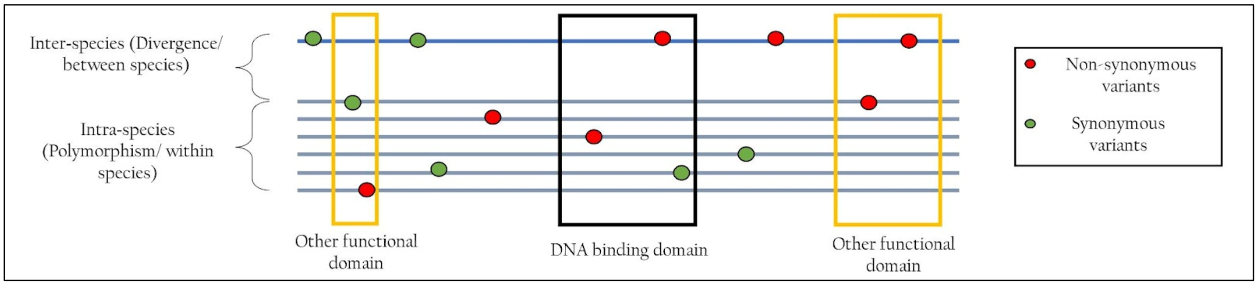


Figure 1.1 – Graphical representation of the construction of this study. First, we extract the coordinates of the annotated functional domains within the TFs. DNA-binding domains are the "test" set of regions. Functionally unannotated regions within the coding sequences, non-DNA-binding domains, are "control" set of regions. Next, we identify the genetic variants occurring within these two genomic regions on the divergence and polymorphism scales. Finally, we segregate variants on the basis of their impact on the encoded amino acids (non-synonymous variants: change the encoded amino acid; synonymous variants: do not change the encoded amino acid)

In the case of *Arabidopsis thaliana*, on the scale of shorter evolutionary time scales, this project focused on three populations, namely – Iberia (IB, n=45), North Sweden (NS, n=45) and Central Asia (CA, n=45). Information on the population-specific variations was obtained from the 1001 Genomes Project (Alonso-Blanco et al. 2016b). For divergence information, we performed transcript-specific *reciprocal blast* with the transcript sequences from the outgroup species *Arabidopsis lyrata* (Hu et al. 2011). In the case of *Homo sapiens*, on the scale of shorter evolutionary time scales, this project focused on three different populations, namely – Yoruba (YRI, n=105), Utah residents with European ancestry (CEU, n=96) and Southern Han Chinese (CHS, n=105). Information on the population-specific variations was obtained from Byrska-Bishop *et al.* (Byrska-Bishop et al. 2022) that was made available through the data repository of the 1000 Genomes project (Auton et al. 2015b). For divergence information, we performed transcript-specific *reciprocal blast* with the transcript sequences of the outgroup species, *Pan troglodytes* (Mikkelsen et al. 2005). Finally, in the case of *Drosophila melanogaster*, on the scale of shorter evolutionary timescales, this project focused on two populations – Zambia (ZAM, n=108) and Sweden (SWE, n=14). Information on the population-specific variations was obtained from (Kapopoulou et al. 2020) and DPGP3 (Lack et al. 2015). Similar to the previous two species, the divergence information was obtained by performing a *reciprocal blast* with the transcript sequences from the outgroup species, *Drosophila simulans* (Clark et al. 2007).

Constraint ratios

In coding regions of the genome, the majority of the nonsynonymous mutations (mutations changing the encoded amino acid) could be considered potentially deleterious due to their impact on the encoded amino acid; selection would be expected to work against such variants, and hence such variants would be maintained in lower frequencies within species. A small portion of the nonsynonymous variants could also be beneficial, selection would be expected to work in their favour, and these would increase in frequency and eventually get fixed within species. However, in natural populations, frequencies of nonsynonymous variants are not only explained by drift and selection but also by other external factors such as bottlenecks, changes in demography, etc. Functional genomics studies use synonymous variants (mutations not changing the encoded amino acid) as putative neutral sites to control for these external factors. This ratio of nonsynonymous to synonymous variants, here termed as constraint ratio, has been used extensively in functional genomics studies to highlight the action of selection on various coding region elements (Guéguen and Duret n.d.; Yang and Nielsen 2000; Choudhuri 2014). This study uses constraint ratios over two evolutionary time scales: intra-specific constraint ratio (π_n/π_s) and inter-specific constraint ratio (K_n/K_s). Comparing these constraint ratios across different genes/genomic regions could give an understanding of the underlying selective forces. A central factor influencing these ratios is the number of gene-specific sites where nonsynonymous and synonymous variants could occur. Different genes would have differing nonsynonymous and synonymous sites, primarily due to differences in gene lengths. This could be a potential caveat in comparing constraint ratios across different genes and genomic regions. To counter this caveat, we factor in the differing gene-specific number of nonsynonymous and synonymous sites during the construction of the constraint ratios. Specifically, we normalise the raw variant counts by the number of sites to obtain gene-specific π_n, π_s, K_n and K_s metrics.

An overall excess of the π_n/π_s constraint ratios as compared to K_n/K_s constraint ratios could be explained by purifying/negative selection. Here, selection maintains a high proportion of nonsynonymous variants at a lower frequency within the population, contributing more to within-species differences. At the same time, an excess of K_n/K_s as compared to π_n/π_s could be explained by positive selection. Here, beneficial alleles within species get “fixed” and contribute more towards between-species differences than within-species differences.

Accessing the clinically annotated variants for *H. sapiens*

Information on the clinically annotated variants within the *H. sapiens* genome was obtained from the ClinVar database (Landrum et al. 2018). Specifically, the clinical annotations were obtained through the data repository of ClinVar in a tab-delimited format (date of accession - 2021-10-16). This file was first filtered to contain annotations only for single nucleotide variants. Next, we subset the file to exclusively contain variants annotated within the coding region of the TFs included in our dataset. The variants were then segregated based on their consequences, i.e., benign and pathogenic variants, and their location, i.e., DNABD and non-DNABD. Fischer's exact test compared the proportions of pathogenic variants for the two genomic regions.

Correlating the efficiency of selection with the species-specific effective population size (N_e)

A drift-dependent variable could be used as a control to quantify the effect of change in species-specific Ne on the intensity of selection. Genetic diversity (π) (Nei and Li 1979) is a quantity directly dependent on the number of individuals within a population ($\pi \propto N_e*\mu$), where N_e is the effective population size and μ the mutation rate. However, according to Lewontin's paradox (Lewontin 1974), the magnitude of change in the neutral genetic diversity may not translate to a similar magnitude of change in N_e. This is mainly due to two reasons, firstly, neutral genetic diversity could also be affected by changes in mutation rates (μ), and secondly, neutral genetic diversity is correlated to a "mean" N_e and does not factor in recent changes in population due to bottlenecks (Galtier 2016a). The proportion of deleterious mutations segregating within species is another drift-dependent quantity, where a decrease in drift would limit this proportion. In coding regions, amino acid-changing nonsynonymous variants could be perceived as deleterious mutations due to their impact on the coding sequences and are used to quantify the change in drift caused due to a change in N_e. However, factors like changes in the demography, mutation rate (when studying distantly related species), bottlenecks etc., could influence the rate of occurrence of such deleterious variants (π_n). To control for such influences, synonymous sites have been used as neutral sites, the presumption being that external factors would influence both nonsynonymous and synonymous sites. Additionally, the rate of occurrence of these neutral variants (π_s) increases with an increase in the number of individuals within the populations. Concatenating the drift-dependent variable, which elucidates the action of drift and selection, along with a neutral variable, to control for factors obstructing the signal of selection, the ratio of nonsynonymous mutations to synonymous mutations (π_n/π_s) has been used to infer the effects of change in N_e on the genetic drift and efficiency of selection. In this study, we use the π_n/π_s ratio as a proxy for the efficiency of purifying selection. An increase in the efficiency of purifying selection would translate to an increase in the efficiency of "weeding out" non-beneficial and potentially deleterious, nonsynonymous variants, consequently lowering π_n/π_s. The rate of occurrence of the neutral variants, π_s, is taken as a proxy for N_e.

The measure of beneficial mutations under the influence of positive selection is given by α. This measure could be obtained by contrasting the proportions of "test" mutations segregating within species to the proportion of "test" mutations contributing to the between-species differences (explained in detail in the consecutive sections). Hence, α is used as a proxy for the efficiency of positive selection. Here too, the rate of occurrence of the neutral variants, π_s, is taken as a proxy for N_e.

Performing polymorphism- and divergence-based analysis over the whole gene set (WGS) per species

Species-specific whole gene sets (WGS) were used as an additional control to contrast the action of selection acting on the DNABDs. For *H. sapiens*, WGS consisted of all "protein-coding" genes within chromosomes 1 to 22. For every gene, the MANE Select transcript was chosen as the representative transcript. MANE transcripts are a product of the MANE project, a joint initiative from NCBI and EMBL-EBI, that aims at defining genome-wide representative transcripts for protein-coding genes in *H. sapiens* (Morales et al. 2022). Consequently, every gene that did not have an annotated MANE Select transcript was filtered out from the analysis. For *A. thaliana*, WGS consisted of all the "protein-coding" genes within chromosomes 1 to 5. For every gene, an Ensembl canonical transcript was chosen as the representative transcript. Consequently, every gene that did not have a corresponding Ensembl canonical transcript was filtered out from the analysis. Finally, in the case of *D. melanogaster*, WGS consisted of all the "protein-coding" genes within chromosomes X, 2L, 2R, 3L, 3R and 4. Similar to *A. thaliana,* an Ensembl canonical transcript was chosen as the representative transcript, and consequently, every gene that did not have a corresponding Ensembl canonical transcript was filtered out from the analysis. The species-specific outgroup species chosen for analysis over LGS analysis were similar to the ones noted in the TF analysis. The number of species-specific genes included in this study has been highlighted in **Table 1.2**.

Species	No. of genes included within WGS
H. sapiens	17751
A. thaliana	27419
D. melanogaster	17679

Table 1.2 - *Summary of the number of species-specific genes included within the WGS region*

Signal of positive selection - predicting the proportion of adaptive substitutions with *asymptoticMK*

The McDonald-Kreitmann test (McDonald and Kreitman 1991) is a widely used approach to detect the action of positive selection by inferring the proportion of potentially adaptive substitutions (α). In a nutshell, the test elucidates the signal of selection by comparing the π_n/π_s and K_n/K_s ratios exclusively. Rand and Kann (Rand and Kann 1996) introduced a similar metric to detect the direction of selection solely on π_n/π_s and K_n/K_s ratios, which they termed the Neutrality Index (NI). However, these approaches have been highlighted to have limitations due to some of their assumptions. Specifically, the assumptions of deleterious mutations being exclusively "lost" within species and beneficial mutations being exclusively "fixed" between species, where the latter is expected to not contribute to the within-species differences. However, slightly deleterious mutations segregating at low frequencies, and beneficial mutations about to reach fixation, would contribute to the within-species differences. Consequently, these assumptions could result in an underestimation of α (Haller and Messer 2017; Moutinho, Bataillon, and Dutheil 2019).

To counter these shortcomings, recently introduced methods, like *asymptoticMK* (Haller and Messer 2017), utilise information from the whole site frequency spectrum (SFS) and the divergence information. Similar to the traditional MK-test-like approaches, this tool requires input information for two types of sites, namely "test" and "control" sites, to counter external factors influencing the detection of selection. By utilising the SFS information, this tool estimates the strength of the positive signal by considering the presence of slightly deleterious alleles (occurring at lower frequencies in the SFS).

Alag – a tool for analysing the genetic variants occurring in the coding regions from a comparative and population genomics-based framework

To perform a selection-based study on the genomic regions of interest, we developed *Alag*. This tool utilises information on genomic variations on the level of polymorphism and divergence. Internally, *Alag* consists of six processing steps and a final output step where the summary statistics are calculated per gene. Primarily, *Alag* was developed for calculating the summary statistics for functional domains of interest. Here, the tool relies heavily on the annotation information on proteins, which in this case was obtained from UniProt (UniProt Consortium 2022). However, these summary statistics can also be calculated and compared for the entire coding regions of genes. *Alag* also has the functionality for requesting the required genomic data from constantly maintained datasets like Ensembl, UniProt (Cunningham et al. 2022; Bateman et al. 2017) etc. This tool has been described in more detail in **Chapter 3**.

Results and Discussions

Signal of high constraint acting on the DNA-binding domains as compared to the control regions across all populations from three species

The nonsynonymous polymorphism constraint ratio (π_n/π_s) depicts the proportion of nonsynonymous to synonymous variants occurring within specific genomic regions across a given population. A reduction in the proportion of nonsynonymous variants compared to synonymous ones is interpreted as a signal of high constraint (**Materials and Methods**). We calculated the mean of π_n/π_s ratio for the DNA-binding domain (DNABD) and non-DNA-binding domain regions (non-DNABD) across the eight populations (distributed across the three species) involved in this study. Here, non-DNABD regions, functionally unannotated regions within the coding region of (TFs), were used as a set of putative control regions. In addition to the non-DNABD control regions, we used the whole gene sets (WGS) per species as an additional control region. Using the WGS as a control enables comparing the constraint acting on the DNABD regions to the overall constraint acting on genes per species. On comparing the mean π_n/π_s ratios for DNABD, non-DNABD and WGS regions (**Table 1.3)** and their respective distributions (**Figure 1.2**), we report a consistent signal of high constraint acting on the DNABD regions across all the population. Compared to the other two species, in the case of *A. thaliana,* we also observed comparatively less constraint acting on the non-DNABD regions than the WGS and DNABD regions. In the case of the other two species, the constraint acting on the non-DNABD regions is comparable to that of the WGS regions.

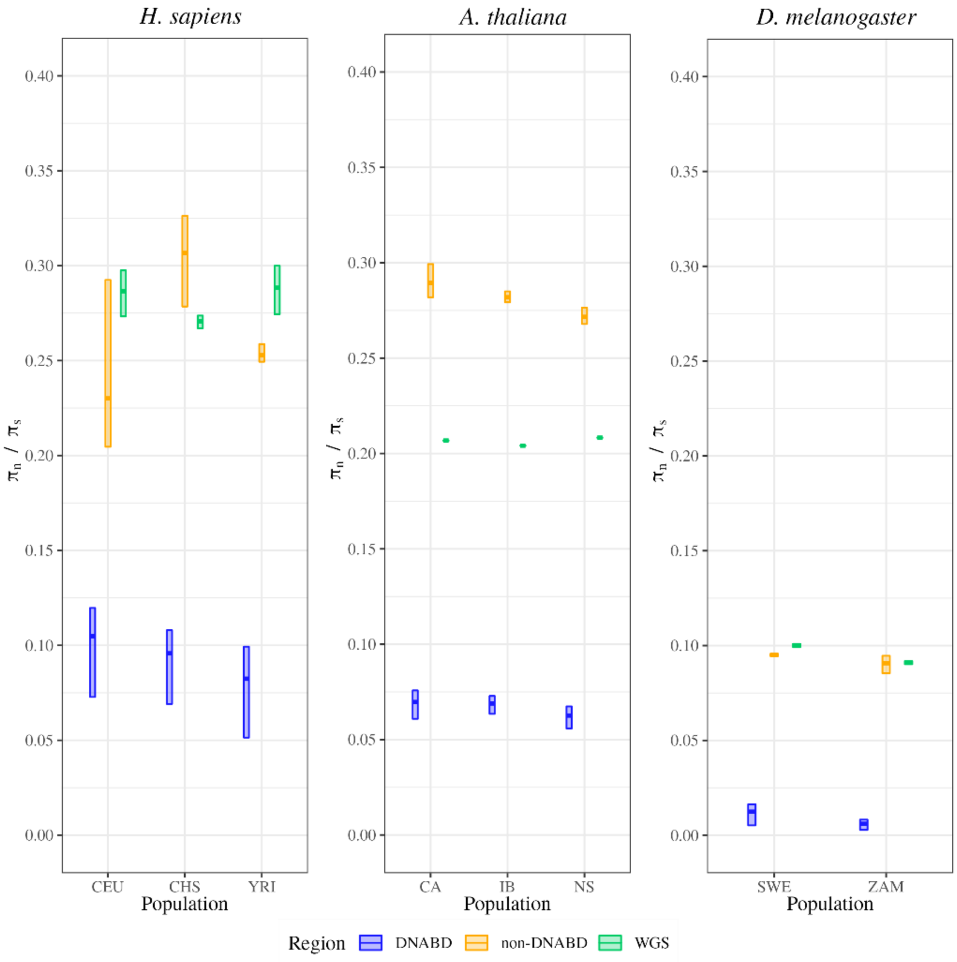


Figure 1.2 - Comparing the nonsynonymous polymorphism constraint ratios (π_n/π_s) across the three regions for the three species and eight populations. (Population codes are: YRI – Yoruba in Ibadan, CEU - Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)

Species	Population	DNABD region		non-DNABD region		WGS	
		π_n/π_s	$\pi_{nonsense}/\pi_s$	π_n/π_s	$\pi_{nonsense}/\pi_s$	π_n/π_s	$\pi_{nonsense}/\pi_s$
Homo sapiens	**YRI**	0.08241	0	0.25276	0.00023	0.28845	0.00161
	CEU	0.10841	0	0.23009	0.00048	0.28664	0.00394
	CHS	0.09582	0	0.30672	0.00043	0.27074	0.00395
Arabidopsis thaliana	**IB**	0.06893	0.00081	0.28187	0.00096	0.20408	0.00195
	NS	0.06261	0.00042	0.27161	0.00173	0.20823	0.00225
	CA	0.06970	0.00040	0.28941	0.00131	0.20679	0.00217
Drosophila melanogaster	**ZAM**	0.00605	0	0.09075	0.00006	0.09101	0.00045
	SWE	0.01236	0	0.09519	0	0.10007	0.00055

Table 1.3 - Comparing the mean estimates of the nonsynonymous and nonsense polymorphism constraint ratios of the three regions for the three species and eight populations. (Population codes are: YRI – Yoruba in Ibadan, CEU - Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)

In addition to the nonsynonymous mutations, we also collected information on nonsense mutations, a sub-category of nonsynonymous mutations that introduce a premature stop codon within the coding regions. We calculated the nonsense polymorphism constraint ratio ($\pi_{nonsense}/\pi_s$), which depicts the proportion of nonsense variants to synonymous variants occurring within specific genomic regions across a given population. On comparing the mean values of $\pi_{nonsense}/\pi_s$ ratios for the three genomic regions, we observe a similar signal of high constraint acting on the DNABD regions compared to the control regions (**Table 1.3)**. Interestingly, in the case of *H. sapiens* and *D. melanogaster*, we did not record any nonsense mutations within the DNABD regions across their respective populations.

Signal of high constraint acting on the DNA-binding domains as compared to the control regions with the outgroup species

Similar to the polymorphism constraint ratios, we calculated the nonsynonymous divergence constraint ratios (K_n/K_s) from the variants collected between the ingroup and outgroup species. Here, K_n/K_s depicts the ratio of nonsynonymous to synonymous variants fixed between the two species. On comparing the mean values (**Table 1.4**) and their respective distributions (**Figure 1.3**) of the K_n/K_s ratios across the three genomic regions, we note a consistent signal of high constraint acting on the DNABD regions compared to the other two control regions. In addition to the nonsynonymous variants, we also calculated the proportions of nonsense variants for the divergent timescale ($K_{nonsense}/K_s$). On comparing the mean $K_{nonsense}/K_s$ constraint ratios, we note a similar signal of high constraint acting on the DNABD regions (**Table 1.4**)

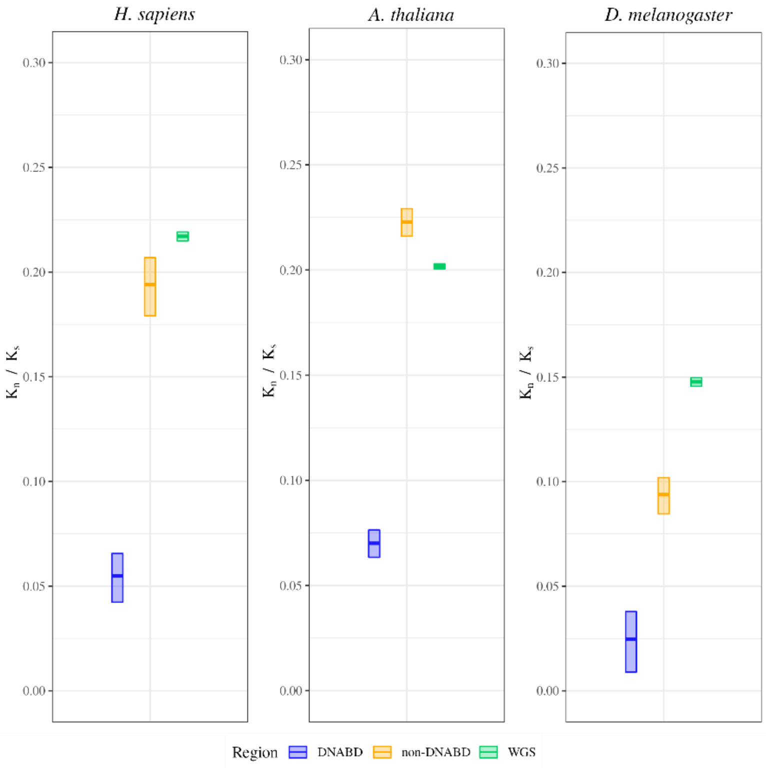


Figure 1.3 – Comparing the nonsynonymous divergence constraint ratios (K_n/K_s) across the three regions for the three species

Species	DNABD region		non-DNABD region		WGS region	
	K_n/K_s	$K_{nonsense}/K_s$	K_n/K_s	$K_{nonsense}/K_s$	K_n/K_s	$K_{nonsense}/K_s$
H. sapiens	0.05488	0	0.19400	0.00005	0.2171	0.00166
A. thaliana	0.07011	0.00017	0.22281	0.00144	0.20176	0.00481
D. melanogaster	0.02543	0.00021	0.09485	0.00091	0.1495	0.00181

Table 1.4 - *Comparing the mean estimates of the nonsynonymous and nonsense divergence constraint ratios of the three regions for the three species*

To summarize, we note a consistent signal of high constraint acting on the DNABD regions compared to the other two control regions. The signal of high constraint was observed across both evolutionary timescales. These observations suggest that the DNABD regions are under a comparatively higher intensity of purifying selection than the genomic background. This constraint could be attributed to their functional importance, specifically, their role in gene regulation.

DNA-binding domains host a comparatively higher proportion of deleterious variants in the case of *H. sapiens*

Given their central role in gene regulation, studies have highlighted the pleiotropic nature of TFs (Chesmore et al. 2016). TFs undertake this critical role of gene regulation mainly through regulatory DNA binding domain regions (defined as DNABD regions in this study). Hence, introducing a variant within these regulatory DNABD regions could disrupt multiple downstream gene regulatory interactions. To further validate this hypothesis, we compared the proportions of deleterious variants to the total number of annotated variants in the DNABD and non-DNABD regions. Specifically, we compared the proportions of pathogenic variants annotated by ClinVar (Landrum et al. 2018) to the total number of clinically annotated variants within H sapiens' DNABD and non-DNABD regio*ns*. ClinVar database catalogues clinically relevant variants within the human genome and annotates them based on their impact on fitness (pathogenic or benign variants). In addition to comparing the proportions of pathogenic variants, we further tested if the proportions for the two genomic regions were significantly different using the Fischer exact test. On comparing these proportions, we observe that the proportion of potentially pathogenic variants occurring within the DNABD regions is significantly higher than those occurring within the non-DNABD regions (p-value < 0.05, **Figure 1.4**). This observation concurs with the expectation that, given the pleiotropic nature of TFs, variants occurring within the regulatory DNABD regions are more likely to be deleterious for fitness than those falling within the control regions. These variants could disrupt multiple regulatory interactions within the species-specific gene regulatory networks.

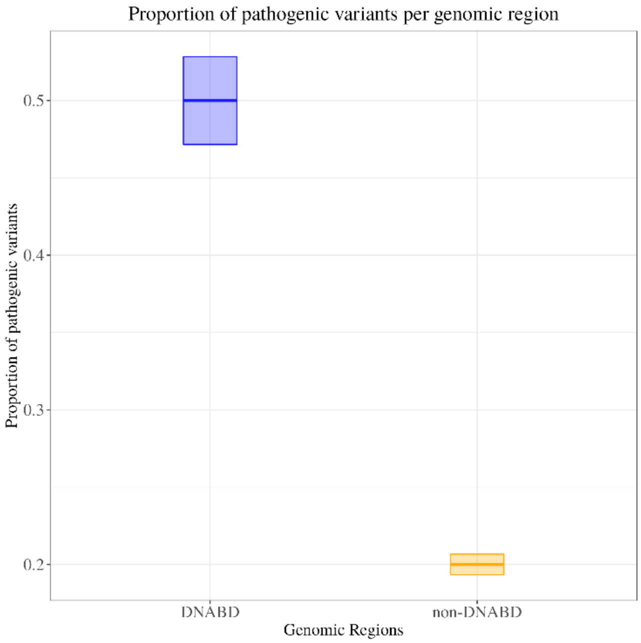


Figure 1.4 – Comparing the proportions of annotated pathogenic variants to the total number of annotated variants for the DNABD and non-DNABD regions in H. sapiens

Primary inferences on the action of positive selection by comparing π_n/π_s and K_n/K_s constraint ratios

Previous sections were focused on elucidating the action of negative selection (purifying selection) acting on the DNABD regions through a comparative framework. Subsequent sections will now focus on inferring the action of positive selection acting on the DNABD regions through a similar comparative framework. We first performed the standard McDonald-Kreitman test (McDonald and Kreitman 1991)(MK test) to obtain preliminary inferences on the action of positive selection on different genomic regions. The MK test makes inferences on positive selection by comparing the proportions of alleles segregating within a given population to the proportions of alleles fixed with the outgroup species.

Table 1.5 shows the estimated proportions of beneficial alleles (α) per genomic region for all three species and eight populations. We make an overall observation that the estimates of α increase across all genomic regions with an increase in the species-specific N_e (N_e *H. sapiens* < N_e *A. thaliana* < N_e *D. melanogaster*). For high N_e species (*A. thaliana* and *D. melanogaster*), we report an overall high estimate of α for the DNABD regions compared to both control regions. However, this was not observed to be the case for *H. sapiens*. Hence, for high N_e species, as compared to the control regions, selection works more efficiently within the DNABD regions in weeding-out non-beneficial alleles (potentially deleterious) and in fixing beneficial alleles (potentially adaptive).

Species	**Population**	**α estimates per region**		
		DNABD	**non-DNABD**	**WGS**
H. sapiens	**YRI**	-0.5	-0.30	-0.33
	CEU	-0.62	-0.13	-0.32
	CHS	-0.6	-0.51	-0.24
A. thaliana	**IB**	0.02	-0.26	-0.01
	NS	0.12	-0.21	-0.03
	CA	-0.01	-0.3	-0.01
D. melanogaster	**ZAM**	0.76	0.04	0.39
	SWE	0.95	-0.01	0.32

Table 1.5 – *Estimates of the genomic region-specific α for the three species and eight populations using the traditional MK test. (Population codes are: YRI – Yoruba in Ibadan, CEU - Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)*

Estimating the proportions of adaptive substitutions (α) with a hybrid of traditional MK test and *asymptoticMK*

The traditional MK-test offers an intuitive method of understanding the action of positive selection by directly comparing the π_n/π_s and K_n/K_s constraint ratios. However, many studies (Haller and Messer 2017; Moutinho, Bataillon, and Dutheil 2019) have highlighted potential shortcomings of the traditional MK test and similar approaches. Specifically, the traditional approach underestimates the α estimate due to slightly deleterious alleles segregating at comparatively lower frequencies within populations. To obtain a more accurate estimate of α, we also used *asymptoticMK* (Haller and Messer 2017). *asymptoticMK*, a proposed extension of the MK test, controls for the presence of slightly deleterious alleles segregating within populations by calculating SFS class-specific α (see **Materials and Methods**). The α estimates and the corresponding confidence intervals obtained from *asymptoticMK* for the different genomic regions across all the populations and species included within this study have been summarized in **Figure 1.5**.

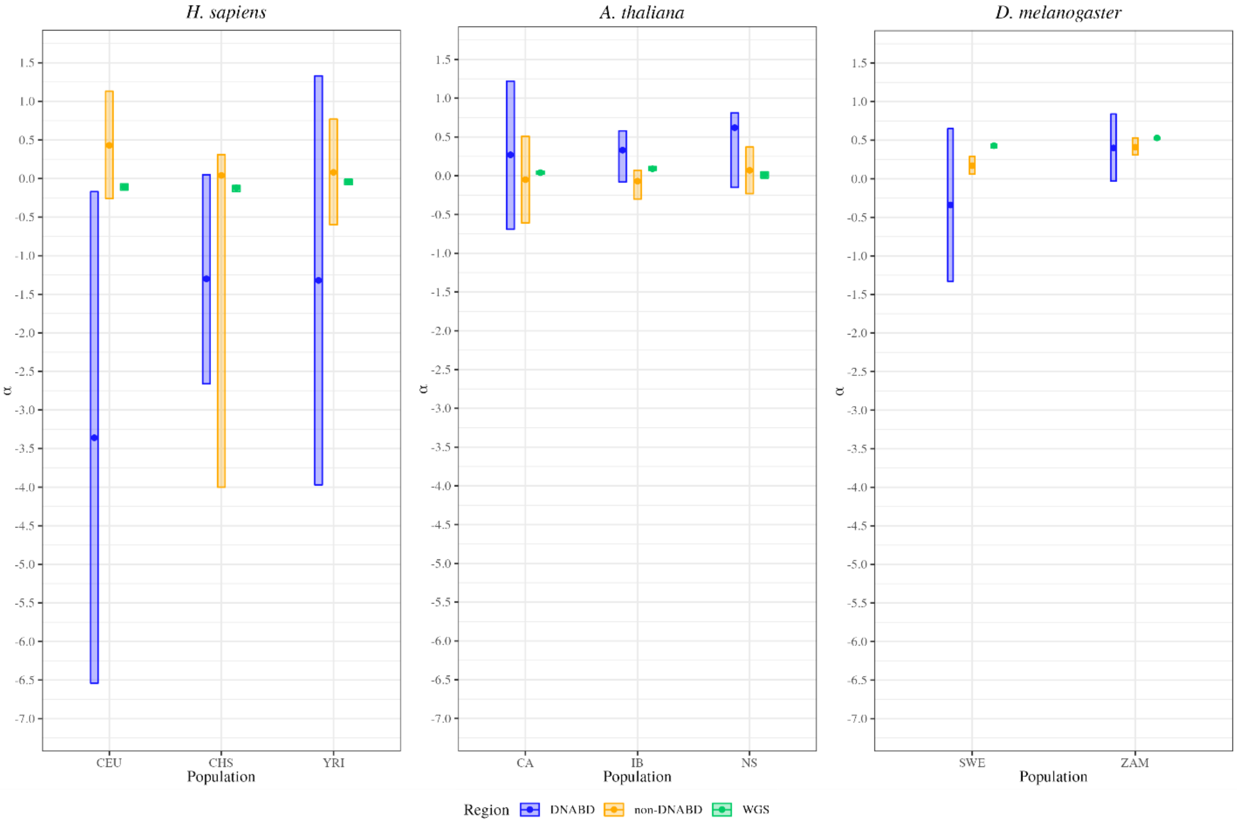


Figure 1.5 – Comparing the α estimates derived from asymptoticMK for the three genomic regions across the three species and eight populations. The bars indicate the 95% Confidence Interval deduced through bootstrapping. (*Population codes are:* YRI – Yoruba in Ibadan, *CEU – Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)*

On comparing the α for WGS regions per species, we observe an overall trend of an increase in the α estimates with an increase in the species-specific N_e. This observation seems to suggest an excess in the proportions of alleles under the influence of positive selection for species with large N_e. In contrast to the observations from **Table 1.5**, the mean estimates of α for DNABD regions were not found to be consistently higher than the other two control regions across all populations. However, the CIs around the estimates for DNABD and non-DNABD regions were relatively high compared to the WGS estimates. One of the main reasons for the high variance in these estimates could be the comparatively smaller number of variants within the DNABD and non-DNABD regions compared to the WGS regions. Consequently, this could result in highly varying α estimates.

Due to the fewer variants per frequency class for DNABD and non-DNABD regions across all populations, we pooled the variants and re-employed the traditional MK test approach but with a frequency cutoff to remove the influence of slightly deleterious variants. **Figure 1.6** depicts the SFS-class specific α estimate for the WGS region for species-specific ancestral populations. The convergence points of the asymptote curves in **Figure 1.6** are considered to be the predictor of mean α. On convergence, the influence of slightly deleterious alleles impacting the estimates of α would be negligible. Hence, we used these convergence points per species as a threshold frequency. Additionally, variants occurring in high frequencies could be mis-polarized. To avoid the disruption of the signal of adaptive substitution due to mis-polarization, we artificially remove alleles using a threshold frequency of 0.9. We removed all the alleles below the threshold frequency before re-calculating α for the DNABD and non-DNABD regions.

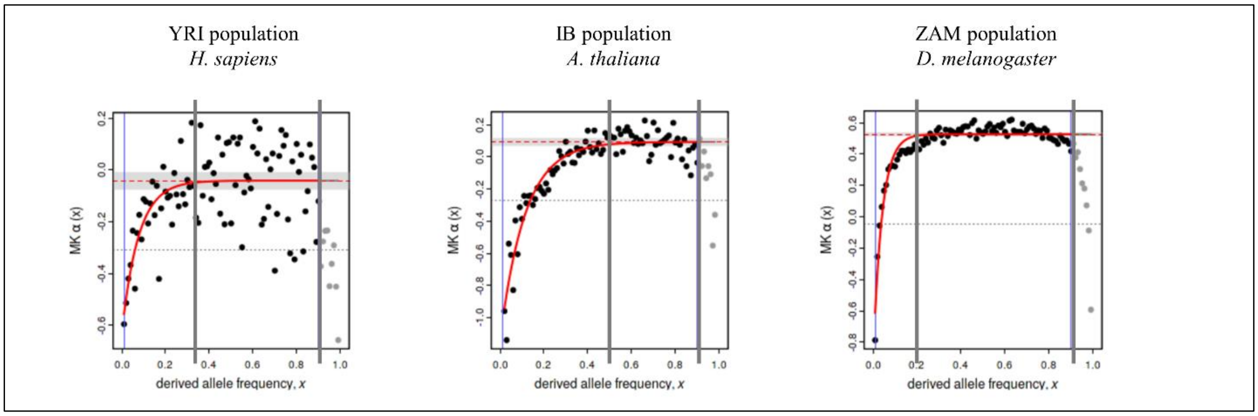


Figure 1.6 – asymptoticMK-based α estimates for the WGS regions of the ancestral populations within the three species. The asymptote is indicated with the red curve, the convergence points symbolize the α estimates, which are indicated with a horizontal red dotted line. The bootstraps are given with a horizontal grey patch around the α estimate. For DNABD and non-DNABD regions, the minimum and maximum frequency cutoffs are indicated with the two vertical grey lines

Table 1.6 depicts the α estimates from re-calculations for DNABD and non-DNABD regions and the mean α for the WGS regions from the *asymptoticMK* analysis. Concurring with results obtained from the traditional MK test approach (**Table 1.5**), here we observed consistently high α estimates for DNABD regions compared to the other two control regions for high N_e species (*A. thaliana* and *D. melanogaster*). However, a similar observation was not observed for the low N_e species (*H. sapiens*). This observation further suggests that for high N_e species, DNABD regions tend to collect a comparatively higher proportion of adaptive substitutions than the other two control regions.

Species	**Population**	**Frequency cutoffs**	**α estimate per genomic region**		
			DNABD	**non-DNABD**	**WGS**
H. sapiens	**YRI**	0.35-0.90	-0.07	0.23	-0.04
	CEU	0.35-0.90	-1.27	0.13	-0.11
	CHS	0.35-0.90	-0.45	0.09	-0.12
A. thaliana	**IB**	0.50-0.90	0.25	0.01	0.09
	NS	0.50-0.90	0.51	-0.07	0.01
	CA	0.50-0.90	0.01	-0.07	0.04
D. melanogaster	**ZAM**	0.20-0.90	0.85	0.38	0.53
	SWE	0.20-0.90	0.19	0.07	0.43

Table 1.6 – α estimates for DNABD and non-DNABD regions with the traditional MK-test using frequency cutoffs. The used frequency cutoffs per species are highlighted in the third column. The mean α estimates for the WGS region from asymptoticMK are highlighted in the last column. (Population codes are: YRI – Yoruba in Ibadan, CEU – Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)

Scaling of π_n/π_S and α with the species-specific effective population sizes (N_e)

In the previous sections, we observed correlation patterns between N_e and the action of selection, which could be perceived through summary statistics. These observations concur with the extensively studied effective population size hypothesis (Galtier 2016b; James, Castellano, and Eyre-Walker 2016). As described in the **Materials and Methods** section, the species-specific effective population size (N_e) is often used as a proxy to indicate the actual number of individuals within the population, whereas π_s is used as a predictor of N_e. Here, we investigate in detail whether an increase in the population size increases the efficiency of both positive and negative (purifying) selection.

The action of negative selection could be perceived through π_n/π_s. Natural selection would actively work in weeding-out non-beneficial and potentially deleterious nonsynonymous variants segregating within a population, thereby decreasing the π_n/π_s ratio. On plotting the genomic region-specific π_n/π_s against their respective π_s (**Figure 1.7**), we observe an overall inverse correlation between π_n/π_s and π_s for all the genomic regions. This inverse correlation concurs with the N_e hypothesis, suggesting that the efficiency of purifying selection increases with species-specific N_e.

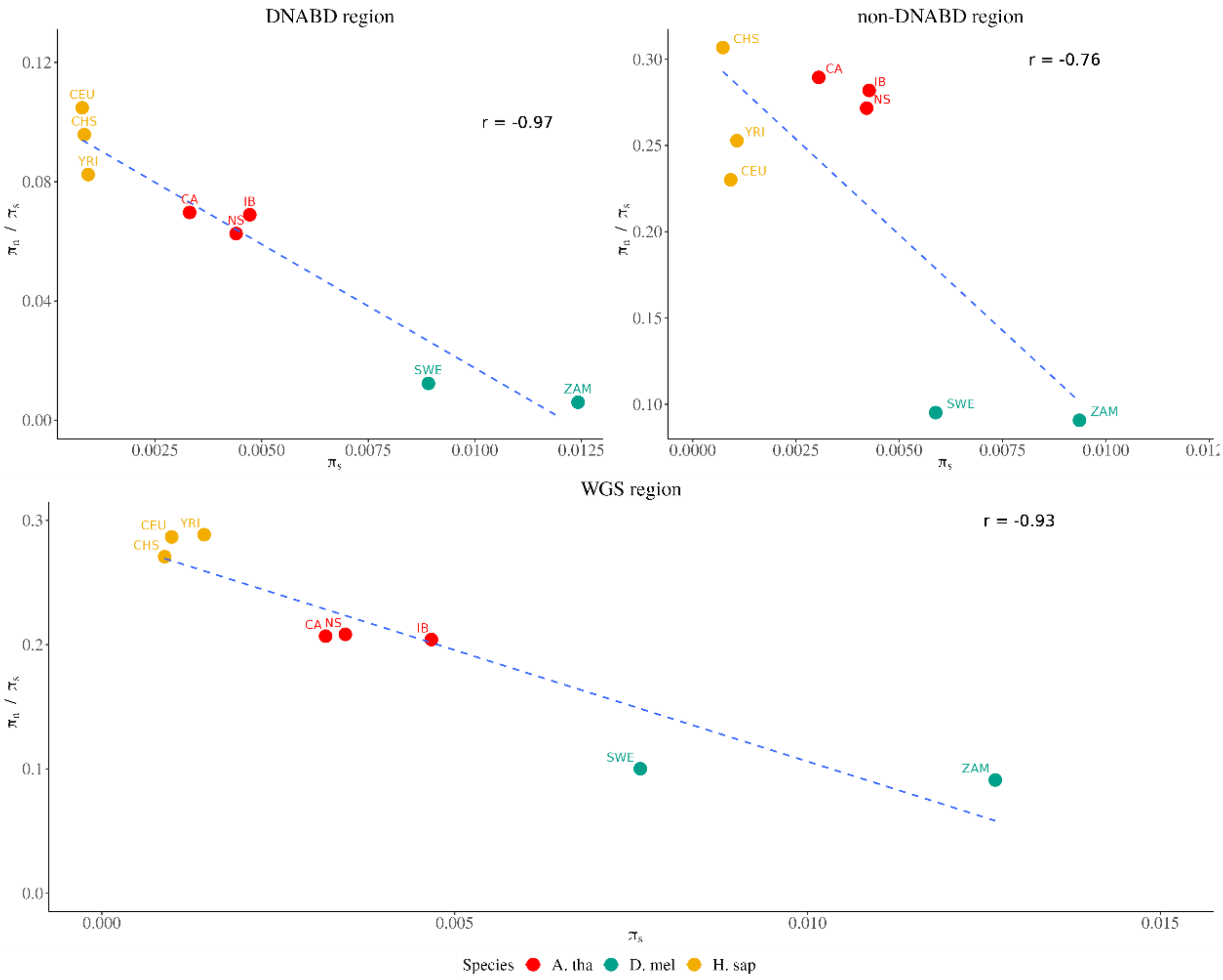


Figure 1.7 – Correlating the efficiency of purifying selection and the species-specific (N_e). Here, π_n/π_s is used as a proxy to quantify the efficiency of purifying selection, and π_s is used as a proxy for N_e. The correlation coefficients per region are noted in their respective panels. (Population codes are: YRI – Yoruba in Ibadan, *CEU – Utah residents with European ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia; Species codes are: A. tha – A. thaliana, D. mel – D. melanogaster, H. sap – H. sapiens)*

The action of positive selection could be perceived through α. In addition to suppressing non-beneficial and deleterious alleles, natural selection would be expected to drive the fixation of alleles bearing adaptive advantages. On plotting the genomic region-specific α against their respective π_s (**Figure 1.8**), we observe an overall direct correlation between π_n/π_s and π_s for all the genomic regions. This direct correlation also concurs with the N_e hypothesis, suggesting that the efficiency of positive selection increases with an increase in the species-specific N_e.

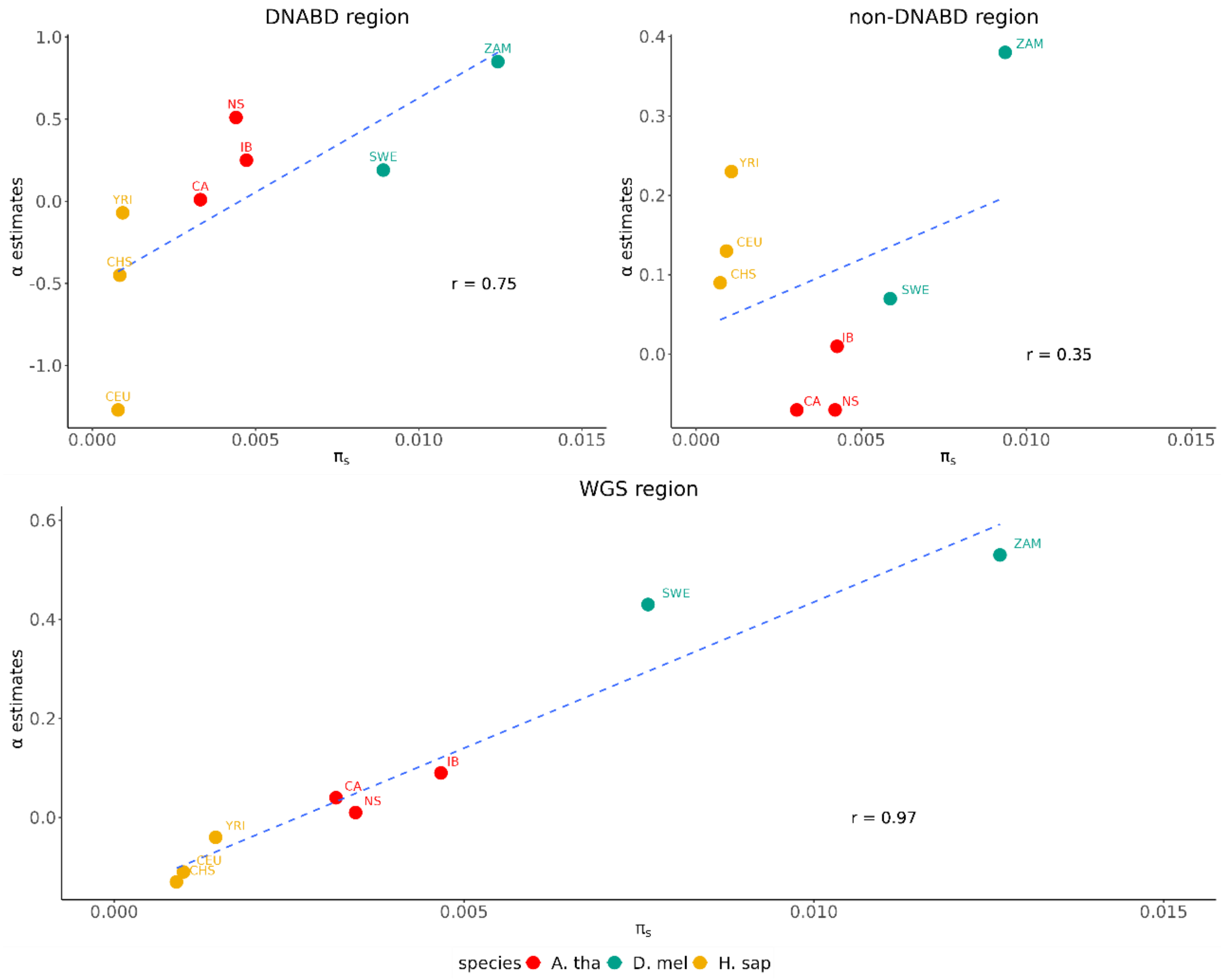


Figure 1.8 – Correlating the efficiency of positive selection and the species-specific (N_e). Here, α is used as a proxy to quantify the efficiency of positive selection, and π_s is used as a proxy for N_e. The correlation coefficients per region are noted in the respective panels. (Population codes are: YRI – Yoruba in Ibadan, *CEU – Utah residents with European Ancestry, CHS – Southern Han Chinese, CA – Central Asia, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia; Species codes are: A. tha – A. thaliana, D. mel – D. melanogaster, H. sap – H. sapiens)*

Hence, we observe an overall increase in the efficiency of positive and negative selection with an increase in N_e. Additionally, we note that the correlation coefficients for non-DNABD regions were consistently the lowest of the three regions in both cases. This observation further suggests that the scaling of selection efficiency with an increase in N_e differs for different subsets of genomic regions.

Chapter 2 – Elucidating the action of natural selection on the Transcription Factor Binding Sites (TFBS)

The Abstract and Introduction sections of this chapter are mainly taken from the following review:

Joshi, M., Kapopoulou, A., & Laurent, S. (2021). Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective. Frontiers in Genetics, 12(July), 1–10

Abstract

The unprecedented rise of high-throughput sequencing and assay technologies has provided a detailed insight into the non-coding sequences and their potential role as gene expression regulators. These regulatory non-coding sequences are often termed cis-regulatory elements (CREs). Genetic variants within CREs could potentially be associated with altered gene expression and phenotypic changes. Such variants are known to occur spontaneously and ultimately get fixed, due to selection and genetic drift, in natural populations and, in some cases, pave the way for speciation. Aiming to understand the impact of natural selection on these CREs, we aggregate information on Transcription Factor Binding Sites (TFBS) and variants occurring within these sequences on intra- and inter-species levels. By employing a Position Weight Matrix (PWM)-based scoring metric, we could obtain a class of non-neutral variants and construct summary statistics. On comparing the summary statistics for the DNABDs, non-DNABDs, WGSs (from Chapter 1) and CREs, we note a consistent signal of relaxed constraint acting on the CREs compared to the other three regions across both evolutionary timescales. This signal suggests that non-coding functional elements are under a comparatively relaxed constraint compared to the coding regions. We also highlight that CREs were under a comparatively relaxed constraint in the derived populations compared to the older populations across both evolutionary timescales. In addition to a comparatively relaxed constraint, we highlight a comparatively sparse action of positive selection acting on these CREs. Besides the Iberian population of *A. thaliana* and the Zambian population of *D. melanogaster*, we report overall negative estimates of α across CREs for all the other species and populations. On comparing the genomic region-specific summary statistics across different evolutionary timescales, we could highlight that the non-coding regions are under a poor intensity of selection as compared to the coding regions.

Introduction

The initial human genome sequencing project revealed that the proportion of the total genome translated into proteins is ~1.5% (I. H. G. S. Consortium 2001), while the remaining portion (~98.5%) consists of non-coding DNA. This significant proportion of non-coding DNA is a hallmark of the genomes of higher organisms (Li and Liu 2019). Evaluating the impact of genetic variation at the coding level is facilitated by a large number of annotated gene models and the simplicity of the genetic code for protein-coding DNA sequences. However, similar studies at the functional non-coding level have suffered from the comparatively sparse annotation and the complex and multifarious nature of the regulatory code. In this context, a vigorous debate unfolded as to the amount of functional information carried by the non-coding genome and eventually led to the broad acceptance that while essential, non-coding functional elements amount to a modest proportion of the total non-coding DNA (Doolittle 2013; Graur et al. 2013; Rands et al. 2014; Huang, Gulko, and Siepel 2017). In the last decade, advances in sequencing and assay technologies have contributed to the annotation of a large number of functional non-coding elements. For example, the ENCODE and modENCODE consortia (The modEncode Consortium 2011; The ENCODE Project Consortium 2012) used chromatin immunoprecipitation using sequencing (ChIP-seq) and ChIP-on-chip assays to gather a comprehensive catalogue of binding sites for a large number of Transcription Factors (TFs) in *H. sapiens*, *D. melanogaster*, and *C. elegans* based on genome-wide binding affinity profiles. The availability of such annotation data, along with genomic variation data, has enabled the exploration of non-coding regions for diversity-based signatures of functional constraint. On the other hand, variants occurring in these regions have also contributed to adaptive evolution (Zhen and Andolfatto 2012). Hence, analyzing the patterns of constraint and variation in CREs contributes to our understanding of between-species phenotypic differences and the process of adaptation.

Previous studies focused on elucidating the action of natural selection on the non-coding elements have mainly adopted two approaches in identifying and annotating the potentially functional non-coding elements: biochemical signature-based and conservation-based approaches. Annotating such elements is essential to quantify their exposure to natural selection. Here, biochemical signature-based approaches aim to annotate functional elements through a specific biochemical signature enabled by high-throughput sequencing techniques. These approaches use a biochemical signature as a proxy for functionality. One of the methods to identify potential regulatory elements is DNase-seq. It allows the identification of regions in the genome at which the chromosome has lost its condensed structure and is therefore susceptible to interactions with available TFs and cleavage by the DNase I nuclease. Such loci are DNase I hypersensitive sites (DHSs). They are localized by sequencing the DNA fragments cleaved by the nuclease and mapping them to the reference genome (Sullivan et al. 2015). Another method to assess genome-wide chromatin accessibility is the assay of transposase-accessible chromatin using sequencing (ATAC-seq), which is considered faster and more sensitive than DNase-seq (Buenrostro et al. 2016). Although loci identified by DNase-seq and ATAC-seq have been shown to be enriched in TF binding sites (TFBS) (Karabacak Calviello et al. 2019), these methods do not provide information about the nature of interacting TFs. On the other hand, ChIP-seq can be used to identify binding sites for a specific TF. This method allows the TF of interest to bind to its putative binding sites before the DNA is sheared by sonication. TF-DNA bound complexes are then extracted using a TF-specific antibody. DNA is dissociated from the TF, sequenced, and aligned to the reference genome to identify enriched regions (ChIP-seq peaks) (Park 2009). The approach of equating a biochemical signature to functionality has been extensively highlighted to result in false-positive annotations (Graur et al. 2013; Doolittle 2013). However, functional genomics studies often use these elements displaying biochemical signatures as starting points.

On the other hand, conservation-based approaches identify elements that are conserved across populations of single species or multiple species in a phylogeny. These approaches use phylogenetic or population-specific conservation as a proxy for functionality. The availability of whole-genome sequence data from multiple species has enabled the detection of non-coding genomic regions with extreme sequence conservation at various phylogenetic levels. Conservation in these regions is generally thought to be caused by the presence of functional non-coding elements exposed to similar levels of negative selection across a set of species (Sandelin et al. 2004; De La Calle-Mustienes et al. 2005; Pennacchio et al. 2006). Therefore, comparative genomic analysis of conserved elements is an efficient approach to detecting non-coding elements involved in regulating developmental pathways common to many higher organisms. With the advent of sequencing technologies, it has also become possible to perform deep sampling across populations for species. Several projects (Auton et al. 2015a; Alonso-Blanco et al. 2016a) have embarked on performing such species-specific deep-sampling and sequencing of a large number of individuals across various populations. Through access to these individual-specific sequences, conserved non-coding regions specific to populations could be identified. However, using conservation as a proxy for functional elements could also potentially dilute the signal of selection (Arbiza et al. 2013).

Here, we combine the biochemical- and conservation-based approaches to highlight the intensities of selection acting on the non-coding regions. Specifically, in this chapter, we leverage the available information from biochemical assays and TF binding models to deduce the signal of selection acting on the Transcription Factor Binding Sites (TFBS) through a comparative and population genomics-based approach on three species: *H. sapiens*, *A. thaliana* and *D. melanogaster*. Further, given the summary statistics for coding regions from the previous chapter, we compare the intensities of selection acting on these three species' non-coding and coding regions. First, we obtain a catalogue of non-coding regions within the genomes of these three species that showcase biochemical signatures through ReMap2022 (Hammal et al. 2021). Specifically, we use the species-specific cis-regulatory modules (CRMs) highlighted in ReMap2022, which are stretches of sequences that are binding hotspots for multiple TFs. Next, we obtain binding models of TFs, which have been annotated to bind within the CRMs, through JASPAR (Castro-Mondragon et al. 2022a) and PlantTFDB (Jin et al. 2017). We merge these two data sources to identify TF-specific binding site coordinates within CRM regions using TEMPLE (Litovchenko and Laurent 2016), a tool which aids in performing the analysis of genetic diversity within TFBS regions. On identifying the coordinates of TFBS regions, we further identify genetic variants occurring within them on two evolutionary timescales and obtain a class of affinity-disrupting variants. These are identified on the basis of a metric that we introduced, *ratio score*, which scores variants on the basis of their potential impact on the TFBS through the PWM models. These affinity-disrupting variants could be perceived as the equivalent of non-synonymous mutations within the coding regions. Finally, using synonymous variants within the coding regions as a putative neutral class, we construct constraint ratios for TFBS regions. Comparing the constraint ratios from coding regions and TFBS, we highlight an overall signal of less constraint acting on the TFBS. However, we also observe that in the case of *D. melanogaster,* the levels of constraint acting on the TFBS regions are comparable to those acting on some coding regions. Additionally, on comparing the TFBS-specific constraint ratios, we see a consistent signal of comparatively high constraints acting on the TFBS regions in the ancestral populations as compared to the derived populations. In addition to purifying selection, we also report a comparatively lower intensity of positive selection acting on the derived populations than the ancestral populations. We were able to highlight that the DNABD regions are consistently under a comparatively stronger influence of purifying and positive selection than the TFBS regions.

Materials and Methods

Access to species-specific candidate cis-regulatory elements through ReMap 2022

We retrieved information on the non-coding elements exhibiting biochemical signatures through ReMap 2022 (Hammal et al. 2021). ReMap 2022 is a database that catalogues information on genomic regions within the three species that exhibit biochemical signatures through DNA-binding sequencing experiments (ChIP-seq experiments). This database catalogues stretches of regions annotated to be the binding hotspots for multiple Transcriptional Regulators (TRs) and identifies them as cis-regulatory modules (CRMs). For our study, we use these species-specific CRMs as starting sets of coordinates which are further filtered. First, we filter out CRMs that overlap with the annotated coding regions of the species. With this filter, we ensure to include only non-coding regions within the analysis. Next, we retain CRMs within a 2kb area of the coding regions and have a specific-specific overlap with this area (**Table 2.1**). With this filter, we aimed to retain potential cis-acting regulatory elements in the vicinity of genes (for example – TFBS and neighbouring enhancers). Additionally, we filter to keep CRMs with a minimum threshold of overlap with the 2kb region in the vicinity of the coding region. Finally, we filter these regions further to contain CRMs that have been assigned a score of 30 or more by ReMap 2022. The score assignment for every CRM correlates with the number of TRs annotated to be binding within the CRM. **Table 2.1** highlights the starting and filtered number of CRMs per species.

Species	**Total CRMs reported in ReMap 2022**	**CRMs in the vicinity of coding regions (2kb)**	**The threshold for minimum overlap within the 2kb vicinity region**
H. sapiens	3,329,428	22,208	500 bp
A. thaliana	228,624	9,736	500 bp
D. melanogaster	591,693	12,217	250 bp

Table 2.1 - Summary of the number of CRMs retrieved and filtered in this study. From all the species-specific CRMs retrieved from ReMap2022, only those within a 2kb vicinity of a coding region and a minimum overlap were retained. The species-specific retained numbers of CRMs are highlighted in the third column, and the threshold for overlaps is highlighted in the last column

Access to Position Weight Matrix (PWM) information

On obtaining the filtered list of CRMs, we also obtained information on the annotated TRs to have a binding activity within the specific CRM coordinates. The total number of TRs annotated to be binding in the filtered set of CRM coordinates is highlighted in **Table 2.2**. To precisely identify the binding coordinates of every TRs in the CRMs, we retrieve PWM data for these TRs. We access information on PWMs for *A. thaliana* through PlantTFDB (Jin et al. 2017) and for *H. sapiens* and *D. melanogaster* through JASPAR (Castro-Mondragon et al. 2022b). PWM gives information on a consensus binding profile for DNA-binding proteins. We first retrieved all the available species-specific PWMs from the data sources. Further, we retain PWMs whose corresponding TRs have been annotated to bind in the filtered CRM coordinates. The retained number of TRs annotated to have a binding activity within the filtered CRM coordinates and that also have corresponding PWM information are highlighted in **Table 2.2**.

Species	Total number of TRs reported to bind in the CRMs	No. of TRs that exhibit binding activity in the 2kb vicinity region	No. of TRs with available PWM information
H. sapiens	1210	1207	421
A. thaliana	423	422	250
D. melanogaster	550	550	101

Table 2.2 - *Summary of the number of TRs annotated to be binding within the CRMs and a subset of those TRs on which PWM information could be retrieved via JASPAR (for H. sapiens and D. melanogaster) and PlantTFDB (for A. thaliana)*

Polymorphism and Divergence information

Intending to capture the signal of selection acting on different evolutionary time scales, we construct this study in an intra- and inter-species framework. In the case of *H. sapiens*, this study consisted of two populations, Yoruba (YRI) and Utah residents with European ancestry (CEU), with a sample size of 45 individuals per population. Population-specific polymorphism data were retrieved from Byrska-Bishop *et al.* (Byrska-Bishop et al. 2022), which was made available through the data repository of the 1000 Genomes project (Auton et al. 2015a). We used genome-wide alignments for divergence data from *H. sapiens* to the outgroup species, *Pan troglodytes*. Specifically, using the REST-API feature from Ensembl (Yates et al. 2014), we retrieved CRM region-specific alignment with the outgroup species. Consequently, the analysis filtered and excluded CRM regions that were not aligned with the outgroup in the context of genome-wide alignment. In the case of *A. thaliana*, this study consisted of two populations – Iberia (IB) and North Sweden (NS), with a sample size of 45 individuals per population. The population-specific polymorphism data were retrieved from the 1001 Genomes project (Alonso-Blanco et al. 2016b). Similar to humans, the divergence data for *A. thaliana* were also retrieved through the genome-wide alignment with the outgroup species, *A. lyrata*. We retrieved CRM region-specific alignments with the outgroup species through the REST-API feature of Ensembl (Yates et al. 2014). The non-aligned regions with the outgroup species were consequently filtered out. In the case of *D. melanogaster*, this study consisted of two populations, Zambia (ZAM) and Sweden (SWE), with a sample size of 30 and 14, respectively. The population-specific polymorphism data were retrieved from (Kapopoulou et al. 2020) and DPGP3 (Lack et al. 2015). Concerning the divergence information, the CRM regions were first aligned to the genome of the outgroup species, *D. simulans*. The resulting sequences were then re-aligned using MUSCLE (Edgar 2004).

Identifying TFBS regions within CRMs using TEMPLE

In order to precisely highlight the binding site of the TRs within a given stretch of the CRM region and to perform an overall analysis of the genetic diversity within the predicted TFBS region, we employ TEMPLE (Litovchenko and Laurent 2016). TEMPLE is a bioinformatics tool that predicts TFBS regions and performs genetic diversity analysis of these regions across different populations through a population genetics framework. This tool takes in three important input files:

- Sequence alignment file (across the population(s) and with the outgroup species) – To capture the genetic variants within TFBSs occurring across different populations, TEMPLE uses individual-specific sequence information. TEMPLE also facilitates the analysis of two populations in a single instance. Hence, we constructed a file containing sequence information per strain per population. In order to polarize the identified genetic variants, we also integrate the sequence information of the outgroup species, which was obtained using alignments (described in the previous section).
- Region file – This file acts as an annotation source that aids in identifying the CRM region that TEMPLE processes. Specifically, this file contains meta-information on the genomic location of the CRM region and identifies PWMs that are scanned in this CRM region.
- PWM file – TEMPLE uses PWM information to identify TFBS regions within the CRMs. This information is provided in the form of a count matrix. Specifically, the count matrices were constructed such that the sum of the counts of all four alleles for a single position would be 1000.

One of the relevant aspects of the results from TEMPLE is the reported TFBS regions and the population-specific variants occurring within these regions. TEMPLE generates a joint output file for the two populations involved in the analysis. We separate the output files per population to contain only the reported segregating variants occurring within the TFBS regions and filter out fixed variants.

Identifying the coding region nonsynonymous equivalent variants within TFBS

As highlighted before, a potential challenge in performing functional analysis of the non-coding regions is the poor availability of annotations. Given the information on the triplets of nucleotides and their corresponding amino acids, variants occurring within the coding regions could be segregated based on their impact on the encoded amino acid. In the case of TFBS, we employ a similar approach to identify variants that would potentially impact the TFBS region's affinity. Specifically, using the count matrix information (PWM) and variants within the TFBS identified by TEMPLE, we investigate the change in the position-specific count introduced due to the variant compared to the ancestral allele. To quantify this change in the position-specific counts, we devise a metric that we term the *ratio score*. This metric could be summarized as follows:

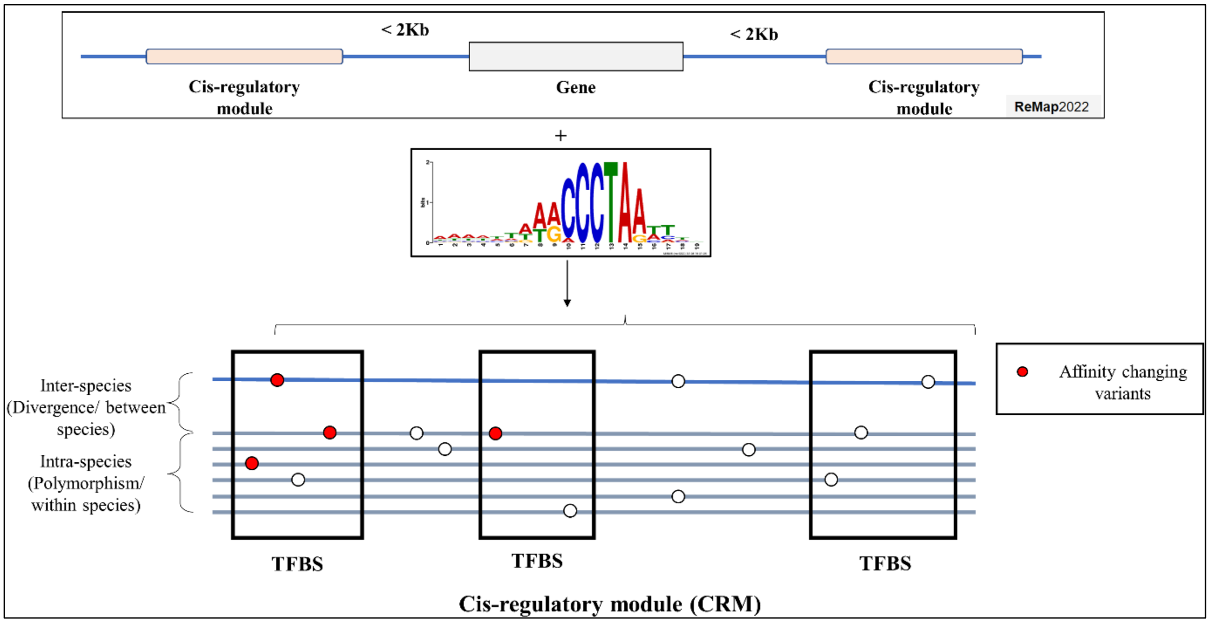
ratio score = absolute (count of the ancestral allele – count of the variant allele) / max (count of the ancestral allele, count of the variant allele)

This metric handles affinity-increasing and decreasing alleles similarly by measuring solely the magnitude of the potential change in the position-specific affinity introduced by the variant. We set two filters to identify nonsynonymous equivalent variants within the TFBS regions. First, the resulting variant had to have a *ratio score* of 0.6 or above. The PWM files used in this study are count-based. Specifically, the sum of the position-specific counts for all alleles is 1000. In some cases, both the reference and the alternate alleles could occur in small counts and result in a high *ratio score*. This would potentially be a false signal. To further ensure that alleles whose reference and alternate alleles occur in small counts in the PWM are not identified as nonsynonymous variants, we set a second filter of a minimum count for either of the ancestral or variant allele to be 400 or more.

Consequently, with the rigid thresholds and filters, the retained variants could potentially impact the affinity of TFBS. However, the filtered-out variants might not always be neutral to the binding affinity; hence, categorizing them into synonymous-like variants could be inaccurate (especially variants with a ratio-metric score of just under 0.6 and a maximum count of alleles just under 400). To avoid using a potentially non-neutral class as a control, we used the synonymous sites within the coding regions as control regions. The approach of employing synonymous sites from coding regions to identify the signal of signature in the non-coding regions has been used extensively (Kosakovsky Pond, Frost, and Muse 2005). This study uses the synonymous sites from the WGS (**from Chapter 1**) as control regions.

The overall construction of the study

The general workflow of this study is highlighted in **Figure 2.1**. We first access information on the species-specific CRM regions and the TRs that have an annotated binding activity within these regions through ReMap 2022 (Hammal et al. 2021). Next, we access information on the PWM models per TRs through JASPAR (Castro-Mondragon et al. 2022a) and PlantTFDB (Jin et al. 2017). These two sources of information are fed to TEMPLE (Litovchenko and Laurent 2016) for identifying the specific coordinates of TFBS regions. In addition to identifying the TFBS regions, TEMPLE also highlights the population-specific variants occurring within these regions. We identify the potential binding affinity disrupting variants using the ratio score metric. These variants are considered nonsynonymous equivalents for this study.



< 2Kb
< 2Kb
Cis-regulatory module
Gene
Cis-regulatory module
ReMap2022
+
Inter-species (Divergence/ between species)
Intra-species (Polymorphism/ within species)
Affinity changing variants
TFBS
TFBS
TFBS
Cis-regulatory module (CRM)

Figure 2.1 – Graphical representation of the construction of this study. We obtain information on the CRM region through ReMap2022. Next, we scan for TFBS regions within the CRM regions using the PWM models of the TFs annotated to have a binding affinity within these regions. Using TEMPLE, we identify the potential TFBS regions and the population-specific variants occurring within them. Finally, using the ratio score metric, we identify the nonsynonymous equivalent variants within the TFBS regions

templeRun – a wrapper around TEMPLE

This study spans multiple species and populations and aims to analyse a large number of CRM regions using a species-specific set of PWMs. Given the expansive nature of this study, and to aid in downstream analysis, we write a tool, *templeRun*, that acts as a wrapper around TEMPLE and enables customizing analysis as per our needs. Specifically, *templeRun* first constructs a *sequence file* per CRM region by building sequence information for every strain within the included two populations per species through a *vcf* (variant call format) file and retrieves information with the outgroup species by internally performing or requesting for alignments. Additionally, this wrapper also constructs the *region file* used by TEMPLE. Finally, *templeRun* internally executes TEMPLE per CRM region by importing the user-defined PWM file.

The outputs obtained from TEMPLE are then processed. Finally, *templeRun* also constructs the construct ratios, which could be used in inferring signatures of selection acting on the TFBS regions. The technical details of this wrapper are explained in more detail in **Chapter 3**.

Results and Discussions

Measuring levels of constraint on TFBS across the six populations

Using the ratio-score metric, we identified a class of affinity-changing variants occurring within the TFBSs. In order to compare the levels of constraints acting on the TFBS to those acting on the coding regions (**introduced in Chapter 1**), the affinity-changing variants in TFBS could be considered a nonsynonymous equivalent of the coding regions. We constructed a TFBS-specific nonsynonymous polymorphism constraint ratio (π_n/π_s), which could be considered comparable to the nonsynonymous constraint ratios of the coding regions. This study's coding and TFBS region-specific polymorphism constraint ratios are denoted by π_n/π_s. The observed mean π_n/π_s constraint ratios for coding and TFBS regions are highlighted in **Table 2.3**.

Species	**Population**	**π_n/π_s**			
		DNABD	**non-DNABD**	**WGS**	**TFBS**
Homo sapiens	**YRI**	0.08241	0.25276	0.28845	0.6502
	CEU	0.10841	0.23009	0.28864	1.3130
Arabidopsis thaliana	**IB**	0.06893	0.28187	0.20408	0.3275
	NS	0.06261	0.27161	0.20823	0.8066
Drosophila melanogaster	**ZAM**	0.00605	0.09075	0.09101	0.0889
	SWE	0.01236	0.09519	0.10007	0.2800

Table 2.3 - *Comparing the mean estimates of the nonsynonymous polymorphism constraint ratios of the coding (DNABD, non-DNABD and WGS) and the non-coding (TFBS) regions for the three species and six populations. (Population codes are: YRI – Yoruba in Ibadan, CEU -* Utah residents with European ancestry, *IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)*

In the case of species with low drift, *D. melanogaster*, we capture an interesting signal of a comparable constraint acting on the TFBS regions compared to the non-DNABD and WGS regions for the Zambian population. Some of the previous studies focused on elucidating the action of purifying selection on the non-coding regions have highlighted that the constraint acting on the non-coding region is comparatively less than the coding regions (Naidoo et al. 2018; Haddrill, Bachtrog, and Andolfatto 2008; Torgerson et al. 2009). Here we show that in the ancestral populations of species experiencing low drift, the level of constraint acting on the TFBS regions could be comparable to the level of coding regions. However, this signal fades in the ancestral populations of species with a comparatively higher drift (*H. sapiens* – YRI & *A. thaliana* – IB). Specifically, the magnitude of the difference between the mean π_n/π_s ratios for WGS and TFBS increases with an increase in drift.

The levels of constraint acting on the TFBS regions for the derived populations per species (*D. melanogaster* – SWE, *A. thaliana* – NS & *H. sapiens* – CEU) were noted to be consistently less than the coding regions. This observation agrees with previous studies suggesting that the non-coding regions are under comparatively less constraint than the coding regions (Naidoo et al. 2018; Haddrill, Bachtrog, and Andolfatto 2008; Torgerson et al. 2009). Additionally, the TFBS regions within the derived populations seem to be consistently under less constraint as compared to the TFBS regions within the ancestral populations. One of the possible reasons for less constraint could be explained by a comparatively higher influence of drift acting on the derived populations compared to the ancestral populations. This signal is further highlighted by comparing the overall distribution of the π_n/π_s constraint ratios for the derived and ancestral populations (**Figure 2.2**).

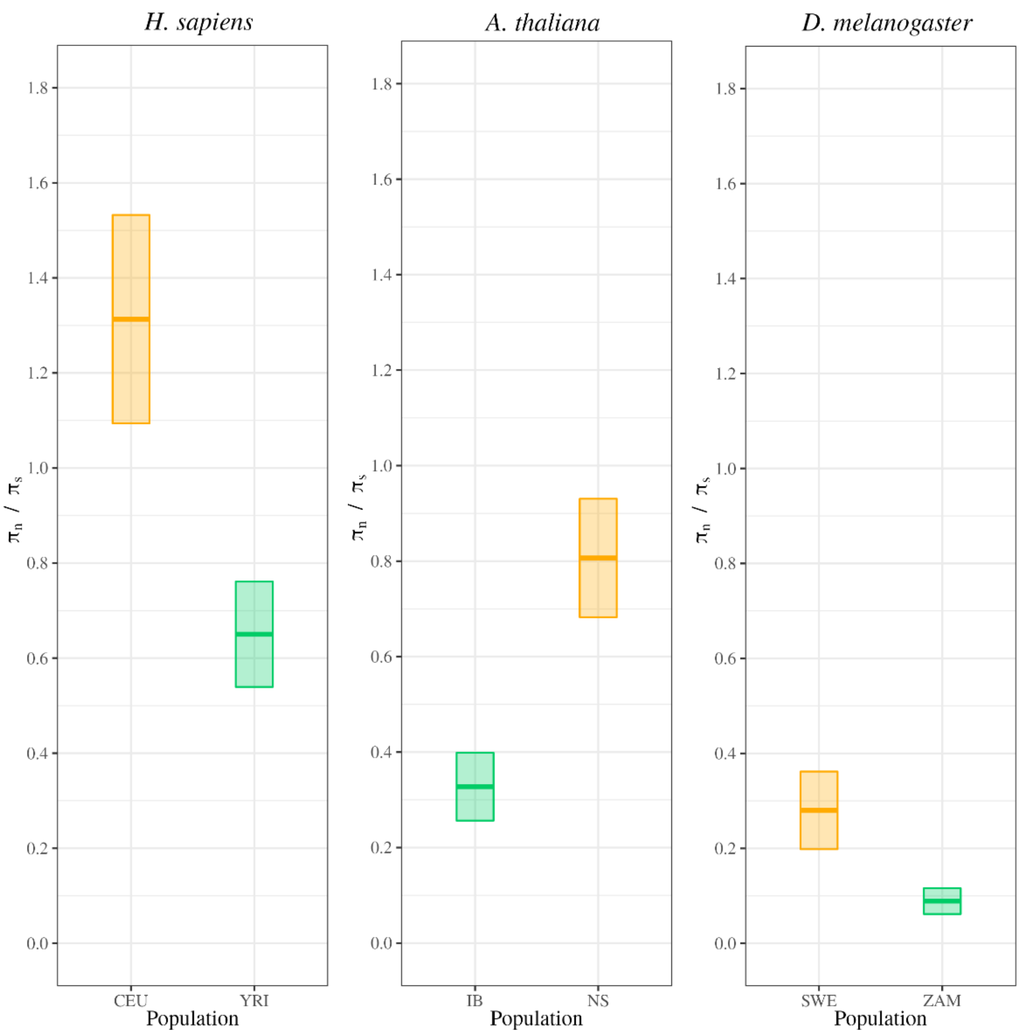


Figure 2.2 - Comparing the distribution of the nonsynonymous polymorphism constraint ratio (π_n/π_s) for the TFBS regions across the six populations. Here, the derived populations per species are highlighted in yellow, and the ancestral populations are highlighted in green. (Population codes are: YRI – Yoruba from Ibadan, CEU - Utah residents with European ancestry, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)

On comparing the means of the π_n/π_s ratios for the DNABD and TFBS regions (**Table 2.3**), we highlight a consistent signal of high constraint acting on the DNABD regions as compared to the TFBS regions. This observation suggests that, on the level of polymorphism, the regulatory domains occurring on the TFs are under a higher intensity of purifying selection than the regulatory domains occurring on the non-coding DNA.

Measuring levels of constraint on TFBS regions with the outgroup species

Similar to the π_n/π_s constraint ratios, we also constructed nonsynonymous divergence constraint ratios (K_n/K_s) for the TFBS regions using the variants collected between the ingroup and the outgroup species. The comparison of the mean of K_n/K_s constraint ratios for the coding (**from Chapter 1**) and TFBS regions is summarized in **Table 2.4**. In this study, the coding and TFBS region-specific divergence constraint ratios are denoted by K_n/K_s.

Species	**K_n/K_s**			
	DNABD	**non-DNABD**	**WGS**	**TFBS**
Homo sapiens	0.05488	0.19400	0.21710	0.6032
Arabidopsis thaliana	0.07011	0.22281	0.20176	0.2430
Drosophila melanogaster	0.02543	0.09485	0.14950	0.0687

Table 2.4 - Comparing the mean estimates of the nonsynonymous divergence constraint ratios of the coding (DNABD, non-DNABD and WGS) and the non-coding (TFBS) regions for the three species.

Along the similar lines of the observations from the π_n/π_s ratios comparison, for species with comparatively lower drift (*D. melanogaster*), here we note a signal of high constraint acting on the TFBS regions as compared to non-DNABD and WGS regions. The signal of a comparatively higher constraint acting on the TFBS regions is more pronounced on the divergence scale than on the polymorphism scale (**Table 2.3**). This signal of high constraint disappears for species with larger drift (*H. sapiens* & *A. thaliana*).

Overall, the level of constraint acting on the DNABD regions seemed to be larger than the TFBS regions. This signal was also observed on the scale of polymorphism data. Hence, this suggests that the regulatory domains occurring on the level of TFs are under a high constraint compared to the regulatory domains occurring on the level of non-coding DNA across both evolutionary time scales.

Estimating the proportion of adaptive substitutions (α) with a hybrid of traditional *MK* test and *asymptoticMK*

Previous sections focused mainly on elucidating the action of purifying selection acting on the TFBS regions. We were able to highlight a signal of a comparatively lower constraint acting on the TFBS regions across both evolutionary timescales compared to the coding regions. We highlight exceptions occurring in the case of species experiencing comparatively lower levels of drifts. To quantify the intensity of positive selection, we estimated TFBS-specific proportions of adaptive substitutions (α). Similar to the DNABD and non-DNABD regions, we employ a hybrid approach of employing the traditional MK test along with *asymptoticMK* (Haller and Messer 2017), an extension of the traditional MK test that incorporates intra-species allele frequency information to estimate α. We adopt this hybrid approach to counter the limited number of variants occurring within populations for TFBS regions (compared to the WGS regions). Similar to the population-specific DNABD and non-DNABD regions (**see Materials and Methods, Chapter 1**), we set minimum and maximum frequency cutoffs and pool the filtered variants to infer TFBS-specific α. The comparison of α for TFBS and all coding regions from **Chapter 1** is highlighted in **Table 2.5**.

Species	**Population**	**α estimates**			
		DNABD	**non-DNABD**	**WGS**	**TFBS**
Homo sapiens	**YRI**	-0.07	0.23	-0.04	-0.66
	CEU	-1.27	0.13	-0.11	-1.63
Arabidopsis thaliana	**IB**	0.25	0.01	0.09	0.39
	NS	0.51	-0.07	0.01	-0.96
Drosophila melanogaster	**ZAM**	0.85	0.38	0.53	0.51
	SWE	0.19	0.07	0.43	-0.88

Table 2.5 - α estimates the coding (DNABD and non-DNABD) and non-coding (TFBS) regions with the traditional MK-test using frequency cutoffs. The mean α estimates for the WGS region from asymptoticMK are also highlighted. (Population codes are: YRI – Yoruba in Ibadan, CEU - Utah residents with European ancestry, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia)

In the case of the ancestral population of *D. melanogaster* (ZAM), we obtain an α estimate that is higher than the non-DNABD region and slightly lower than the WGS region. Interestingly, for the ancestral population of *A. thaliana* (IB), we highlight an α estimate than all the coding regions included in this study suggesting the TFBS regions could harbour a comparatively higher proportion of beneficial mutations. The signal of comparable levels of α for the TFBS and coding region fades in the case of the ancestral population of *H. sapiens* (YRI). The α estimates for all the derived populations of the three species were consistently lower than those from the ancestral populations and their respective coding regions. This observation, in conjunction with the observations from the π_n/π_s (**Table 2.3**) and K_n/K_s (**Table 2.4**), suggests that the TFBS regions in the derived populations are under a stronger influence of drift.

The α estimates for the DNABD regions were consistently higher than those of the TFBS regions, suggesting that the regulatory domains on the TFs are under a comparatively stronger influence of both positive and negative selection than the regulatory domains occurring on the non-coding DNA.

Scaling of π_n/π_s and α with the species-specific effective population sizes (N_e)

This chapter contrasts the signatures of selection acting on the coding and TFBS regions through a comparative framework. We note that the differences in the intensities of selection acting on TFBS and coding regions reduce with an increase in the species-specific N_e.

Figure 2.3 depicts the correlation between population-specific π_n/π_s and π_s for coding and TFBS regions. Here, π_n/π_s is used as a proxy for the efficiency of selection to weed out non-beneficial and potentially deleterious nonsynonymous variants. Whereas π_s, a measure of the proportion of neutral mutations segregating within species, is used as a proxy for N_e for every genomic region. We observe a consistent negative correlation between π_n/π_s and π_s for all four genomic regions indicating that the intensity of purifying selection increases with an increase in N_e. The correlation between these two metrics seems to be lowest for non-DNABD regions, which are stretches of sequences within TFs that have not been functionally annotated. At the same time, the correlation is the strongest for the DNABD regions, which could be explained by their functional importance. These observations concur with the N_e hypothesis, suggesting that an increase in N_e enables purifying selection to act with better precision for weeding-out non-beneficial and deleterious alleles.

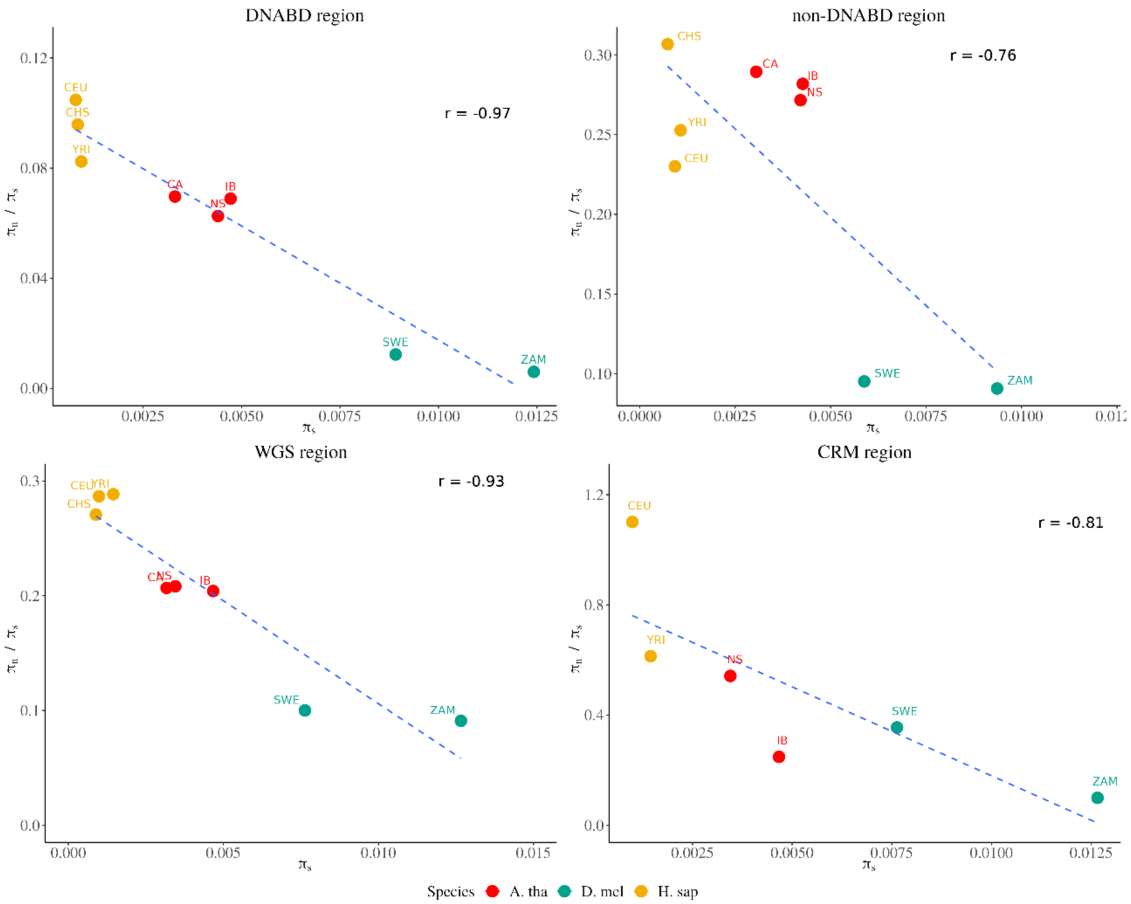


Figure 2.3 - Correlating the efficiency of purifying selection and the species-specific (N_e). Here, π_n/π_s is used as a proxy to quantify the efficiency of purifying selection, and π_s is used as a proxy for N_e. The correlation coefficients per region are noted in their respective panels. (Population codes are: YRI – Yoruba in Ibadan, *CEU – Utah residents with European ancestry, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia; Species codes are: A. tha – A. thaliana, D. mel – D. melanogaster, H. sap – H. sapiens)*

Figure 2.4 depicts the correlation between α and π_s for all the genomic regions. Here, α estimates the proportion of variants driven to fixation by positive selection. Hence, this quantity is a proxy for measuring the intensity of positive selection.

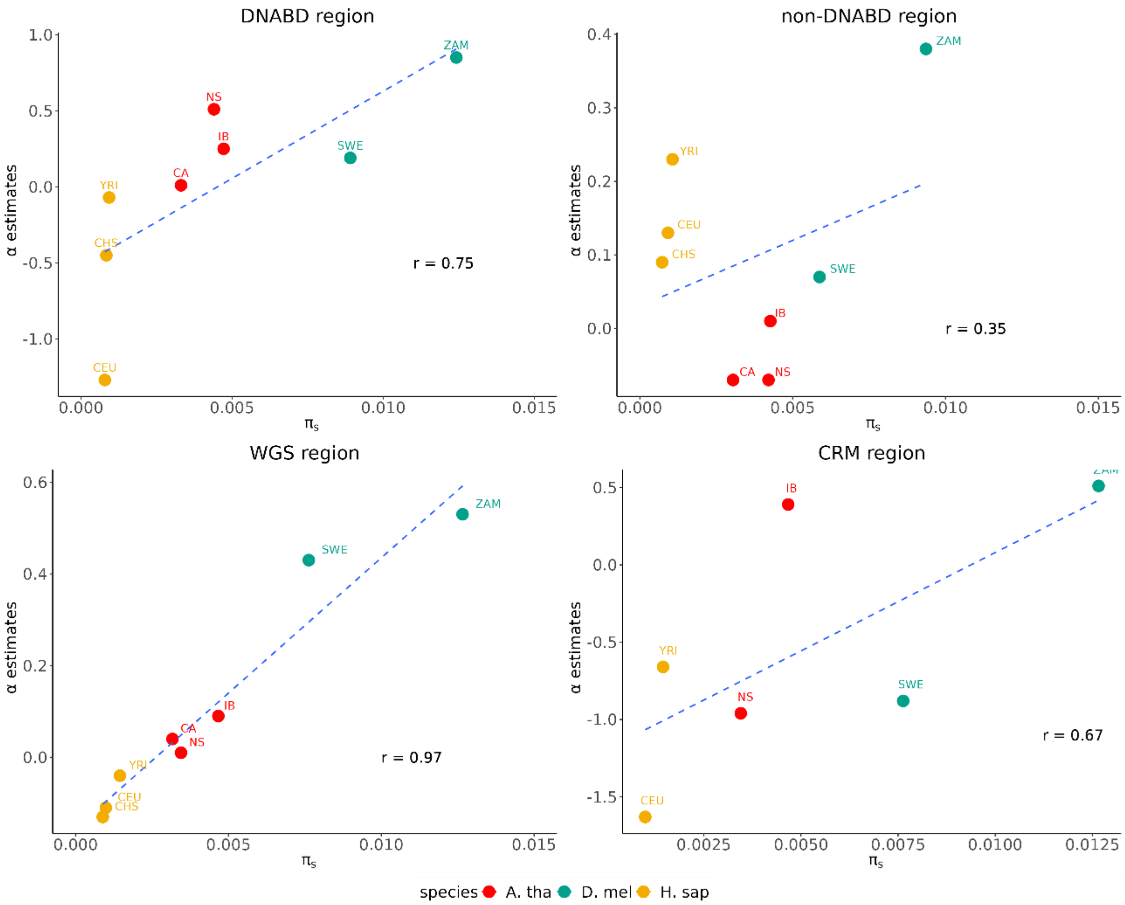


Figure 2.4 - Correlating the efficiency of positive selection and the species-specific (N_e). Here, α is used as a proxy to quantify the efficiency of purifying selection, and π_s is used as a proxy for N_e. The correlation coefficients per region are noted in their respective panels. (Population codes are: YRI – Yoruba in Ibadan, *CEU - Utah residents with European ancestry, IB – Iberia, NS – North Sweden, SWE – Sweden, ZAM – Zambia; Species codes are: A. tha – A. thaliana, D. mel – D. melanogaster, H. sap – H. sapiens)*

In addition to an increased action of purifying selection, there is an overall trend of an increased action of positive selection with an increase in N_e for all four genomic regions. Interestingly, the correlation between α and π_s is the strongest for WGA. This observation suggests that the precision of selection to drive alleles to fixation scales strongly with N_e for the overall gene sets per species as compared to the functional domains (DNABD and TFBS). Consistent with the signal from purifying selection, the correlation between an increase in the intensity of positive selection and N_e is the lowest for non-DNABD regions. Hence, the correlation of an increase in the intensities of purifying and positive selection with N_e is the poorest for the functionally non-annotated regions, non-DNABD.

Chapter 3 – Tools developed for performing analysis of genetic variants occurring within the regulatory coding and noncoding regions

(Tools described in this section are made available here – **gitlab.mpcdf.mpg.de/mjoshi/forthesis_tools/)**

Alag – a tool for performing comparative and population genomics-based analysis of genetic variants within coding regions and functional domains

Introduction

Molecular biology provides a perspective in understanding the evolutionary processes acting on the level of a single organism or phylogeny. Tracking the natural changes occurring within the biomolecules could be used to interpret the action of natural selection. Due to the sparse availability of sequencing data, traditional studies mainly focused on understanding the action of natural selection on the level of species. However, recent advents in sequencing technologies have enabled population-specific deep sampling. Hence the availability of a large number of individual-specific sequencing data per species has helped in understanding the influence of natural selection on the level of populations. The combination of species- and population-centred studies could provide a unique insight into understanding the impact of natural selection on specific genomic elements over two evolutionary time scales (Lawrie and Petrov 2014).

Performing a functional genomics-based study on coding regions is relatively straightforward, given the interpretable genetic code. Specifically, coding regions could be perceived as triplets of nucleotides (codons) that independently code for one amino acid. Hence, variants falling within these regions could be categorized based on their impact on the encoded amino acid (nonsynonymous and synonymous variants). Contrasting the proportions of these amino acid changing (nonsynonymous) to the neutral (synonymous) variants could be used to infer the intensities of selection acting on these genomic elements. Ratios of the "test" to "neutral" variants could be compared across various coding regions to contrast the proportions of constraints on these elements. Additionally, the signal of selection could also be deduced by comparing the proportions of "test" and "neutral" variants on the polymorphism and divergence levels.

Here, we propose *Alag*, a tool for performing functional genomics-based analysis of coding region elements. *Alag* is compatible with working across the longer (between-species) and shorter (within-species) timescales. Initially, *Alag* aimed to elucidate the impact of natural selection on specific classes of genomic regions, namely the DNA-binding domains (DNABD) occurring within the Transcription Factors (TFs). However, we further extend *Alag* to be compatible with performing analysis of the overall coding region sequences within species. Briefly, *Alag* takes in the following information: 1) Population-level variant annotation file of the ingroup species (in *.vcf* file), 2) Transcript sequences of both the ingroup and outgroup species (in *.fasta* format) and 3) Coordinates of the functional domains of interest (optional, in a tabular format). On processing the inputs through a plethora of functions, *Alag* estimates summary statistics that are commonly used in inferring the levels of natural selection acting on specific genes or functional domains. *Alag* also integrates the *asymptoticMK*-based approach (Haller and Messer 2017) to infer the action of positive selection. The following sections are focused on capturing the overall structure of *Alag*. For reproducibility, the code and example input files discussed in this chapter are available on the following GitLab page – (**gitlab.mpcdf.mpg.de/mjoshi/forthesis_tools/**).

Requirements and input files

Alag is majorly written in R (R Core Team 2022) and is compatible with versions *3.6.3* and above. It also integrates other tools and generates automatic system calls from within the scripts when needed. Following are the R packages and external tools used in *Alag* that are used in this example (R version *4.2.0*):

- *ape* (version *5.6-2*) – R package
- *stringr* (version *1.4.0*) – R package
- *Biostrings* (version *2.64.0*) – R package
- *plyr* (version *1.8.7*) – R package
- *dplyr* (version *1.0.9*) – R package
- *seqinr* (version *4.2-16*) – R package
- *parallel* (base package) – R package
- *biomaRt* (version *2.52.0*) – R package
- *ggplot2* (version *3.3.6*) – R package
- *vcftools* (version *0.1.14*)
- *bcftools* (version *1.9*)
- *blastn* (version *2.8.1*)
- *MUSCLE* (version *3.8.1551*)

In addition to the dependencies, *Alag* uses the following input files:

- List of gene identifiers to be used in the analysis – optional if the annotation file is already provided
- Annotation files for the ingroup and outgroup species (in *.gff* format)
- Transcript coding sequences (containing only CDS features) for the ingroup and outgroup species (in *.fa* format)
- A population-level variant annotation file of the ingroup species (in *.vcf* format) – optionally accession to individuals within specific populations
- Coordinates of the functional domains of interest per gene identifier (in a tabular format, optional and not included in this example)

Pre-processing

During the processing of genes through *Alag*, identifying the representative transcript per gene is an important task. Here, the choice of the representative transcript is context-dependent. For the analyses of the DNABD regions, the choice of transcript is based on the availability of the DNA-binding domain annotations per transcript. As stated in the **Materials and Methods section of Chapter 1**, annotations on DNA-binding domains are accessed through UniProt (UniProt Consortium 2022). Hence, the choice of the transcript was based on two filters – 1) The representative transcript should have an annotated *SwissProt* ID & 2) The corresponding *SwissProt* ID has an annotated DNA-binding domain. For the DNABD-based analyses, these criteria were constant across the three species included in our study. In the case of the complete gene analysis, the choice of the representative transcript was species-dependent. In the case of *A. thaliana* and *D. melanogaster*, the Ensembl canonical transcript was chosen as the representative transcript. On the other hand, for *H. sapiens*, MANE (Morales et al. 2022) transcript was chosen to be the representative transcript. The resulting conversion table between gene, transcript and peptide identifiers is stored in the input_files/ folder. The conversion tables are used on multiple instances during the processing.

The flow of information within *Alag*

Two central functions are involved in processing information through Alag: *key.R* and *main.R*. Here, *key.R* is a central point of importing required libraries, and functions, setting paths to the required input files and declaring constants. *Alag* is built for handling a large set of genes per analysis. Hence, the query gene sets are split into short batches by default. The total number of batches and the number of genes per batch is controlled with a combination of a *counter* and a *for-loop* combination, which could be changed when required. *key.R* internally executes *main.R* by supplying the list of gene identifiers to be processed. *main.R* is a hub of all the functions included in *Alag*. These functions are stored in the includes/ folder. *main.R* consists of seven interlinked steps that are executed sequentially. The output resulting from each step is stored in a batch-specific *backup* folder. In the scenarios of termination, this enables resuming the execution of the code from the point of halt. The following are the steps involved:

- Step 1 – Extracting the coding region coordinates for genes

  This step executes the *get_largest_transcript_cds_info* function. This function uses the identifier conversion table to extract the coordinates of genes included in a single batch from the ingroup annotation file.

- Step 2 – Identifying the potential ortholog with the outgroup species using a combination of a reciprocal blast heuristic and realignment – (Polymorphism scale information)

This step executes the *get_transcript_reciprocal_alignments* function. Using the information on the transcript sequences from the outgroup species, this function identifies the potential ortholog gene pairs with the outgroup species using a Reciprocal Best Hit Blast (RBGB) technique. Specifically, first, the transcript sequence from the gene in the ingroup is aligned against all the transcript sequences in the outgroup species (*forward blast*). From the resulting outputs, the transcript with the highest identity score is selected. Secondly, this output transcript from the outgroup species is aligned against all the transcript sequences from the ingroup species (*reverse blast*). Similar to the previous alignment, the choice of the top transcript is based on the identity score. An orthologous gene is identified on the following two conditions: 1) The gene identifier for the ingroup species from the *forward* and *reverse* blast is the same & 2) Both *forward* and *reverse* blasts have a minimum identity score of 60%. Genes that do not pass through these two were omitted from the further analyses. Following the outputs from the reciprocal blast, the resulting ingroup and outgroup transcripts are re-aligned using MUSCLE. Since the estimates made throughout this analysis rely on codons, and alignments through MUSCLE are not sensitive to the open reading frames (ORFs), this function further purifies the alignments by trimming the ends if necessary to retain the ORF.

- Step 3 – Extracting information on the divergent sites and calculating lengths to normalize the variants – (Polymorphism scale information)

This step executes the *get_divergent_sites_2* function. Using the ORF-sensitive alignments per gene generated from the previous step, this function first identifies the positions of divergent sites between the ingroup and outgroup species. These variant sites are first filtered to remove gaps. Next, this function estimates the background lengths on the potential nonsynonymous and synonymous sites within the alignments obtained from the previous step. Specifically, this function iterates over every codon, predicts the impact of mutations occurring in each position, and calculates the total number of nonsynonymous and synonymous sites per codon by dividing these lengths by 3. These per codon lengths are summed over for the aligned regions to obtain a single estimate of the background nonsynonymous and synonymous lengths.

- Step 4 – Calculating divergence statistics – (Polymorphism scale information)

  This step executes the *get_divergence_stats* function. This function combines information on the divergent sites (with the outgroup species obtained from the previous step) and the transcript sequence to categorize variants based on their impact on the encoded amino acid. Specifically, this function observes every variant in the context of the transcript sequence, identifies the codon affected by the variant and predicts the effect of the variant on the respective codon. This function uses the background length information (obtained from the previous step) to normalize the raw counts on obtaining the total number of nonsynonymous and synonymous variants. Consequently, this function calculates gene-specific ratios of nonsynonymous and synonymous variants (K_n and $K_{s,}$ respectively).

- Step 5 – Extracting information on the observed variants within populations – (Divergence scale information)

  This step executes the *get_frequencies_cds_for_aratha* function. First, from the population-specific variation data (supplied to Alag in *.vcf* file), this function extracts variants occurring within the coding sequence of the genes. The extracted variants are further filtered to retain only single nucleotide variants (SNVs). Using the transcript sequence, this function first identifies the codons that would be affected due to these variants and categorize them as either synonymous or nonsynonymous based on their impact on the corresponding codon. Next, the function polarises these variants to identify the ancestral allele state using the outgroup species. The information on the position-specific alleles of the outgroup species is obtained using the transcript alignments (generated in Step 2). Variant positions in a tri-allelic state (differing reference, alternate and outgroup alleles) are filtered out, and only variants in a bi-allelic state were retained. Finally, information on the frequency, polarization (derived or ancestral), effect on the codon (nonsynonymous or synonymous) and meta-information on the genomic position of the alleles are aggregated in a tabulated format.

- Step 6 – Calculating the polymorphism statistics – (Divergence scale information)

This step executes the *get_polymorphism_stats_per_gene2* function. Using the tabulated information on the variants occurring within a given population from the previous step, this function constructs the proportion of nonsynonymous and synonymous variants per gene. Specifically, using the frequency information of variants, this function first calculates the sum of diversity (π), separately for nonsynonymous and synonymous variants, per gene. Next, similar to the divergence statistics (calculated in Step 4), this function normalizes the proportion (diversity) of the two types of variants based on their background lengths. Consequently, this function calculates gene-specific ratios of nonsynonymous and synonymous variants (π_n and $\pi_{s,}$ respectively).

- Step 7 – Converging information from the two evolutionary timescales

This final step is executed within *main.R*. This step collects the divergence and polymorphism statistics (calculated in Step 4 and Step 6, respectively). It merges them to create a single output table. Each row represents a single gene, and the columns represent relevant summary statistics and corresponding meta-information on the gene identifiers.

Example outputs

The batch-specific outputs produced from *Alag* are stored in a *backup* folder. Additionally, on calculating summary statistics per gene, *Alag* creates a table summarizing this information for every batch. Some example outputs are stored in the example_outputs/ folder. The script for performing primary analysis of the outputs from *Alag* and the resulting outputs are stored in analysis/alag_analysis. In this example, we ran *Alag* over the whole gene set of *A. thaliana*. Here, we will discuss some of the outputs briefly:

- Constraint ratios

One of the central outputs from *Alag* is the estimation of the constraint ratios, which are the proportions of nonsynonymous (or similar) variants to synonymous variants. As mentioned before, these proportions are normalized using the background lengths. The resulting polymorphism and divergence constraint ratios for this example are highlighted in **Table 3.1(a) and (b)**, respectively:

π_n	π_s	$\pi_{nonsense}$	π_n/π_s	$\pi_{nonsense}/\pi_s$
0.0047	0.0009	9.13E-06	0.2041	0.0019

Table 3.1 (a) – Polymorphism constraint ratios

K_n	K_s	$K_{nonsense}$	K_n/K_s	$K_{nonsense}/K_s$
0.0277	0.1373	0.0006	0.2018	0.0048

Table 3.1 (b) – Divergence constraint ratios

On comparing the nonsynonymous polymorphism constraint rations (π_n/π_s) and the nonsynonymous divergence constraint ratios (K_n/K_s), it could be seen that π_n/π_s is slightly larger as compared to K_n/K_s. In addition to nonsynonymous mutations, *Alag* also reports the proportion of nonsense mutations. Nonsense mutations are nonsynonymous mutations that introduce a premature stop codon.

- Site frequency spectrum (SFS) plot

In order to understand the distribution of the allele frequencies across the population for the three types of variants, *Alag* constructs SFS. The SFS plot for the example is shown in **Figure 3.1**.

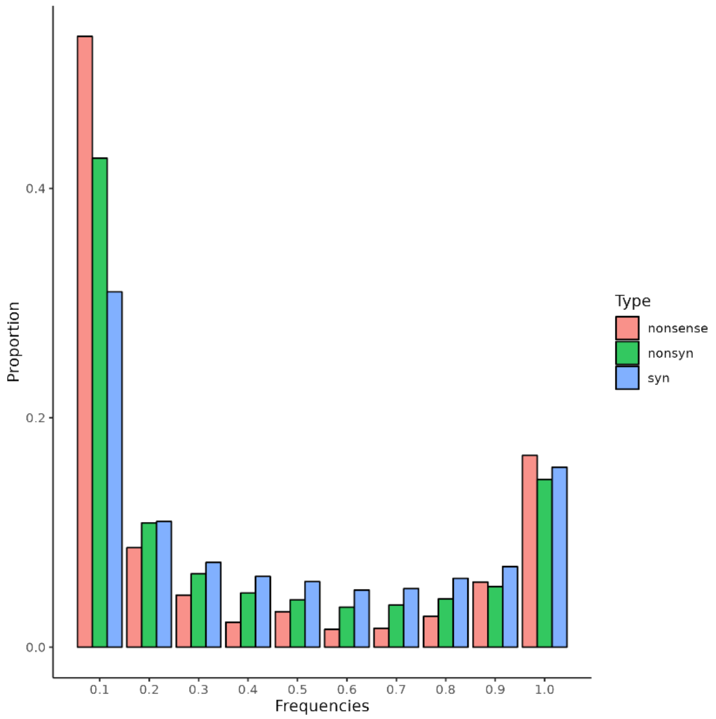


Figure 3.1 – Site frequency spectrum plot for comparing the class-specific frequencies of the three variants

Interestingly, it could be seen that the nonsense mutations harbour a comparatively higher proportion of variants in both low and high-class frequencies as compared to both nonsynonymous and synonymous variants.

- SFS class-specific α estimate

Inspired by the approach implemented in the *asymptoticMK* (Haller and Messer 2017) tool, *Alag* calculates α per frequency class within the SFS. These α estimates for the example are shown in **Fig 3.2**.

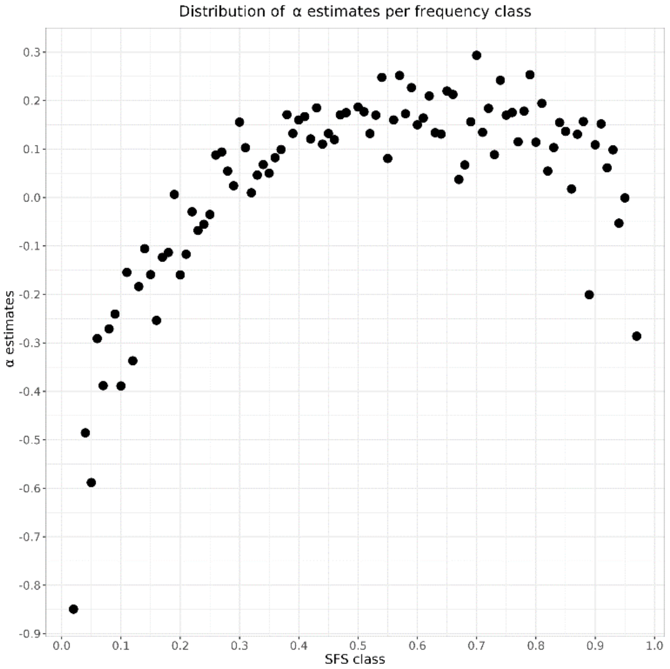


Figure 3.2 – SFS class-specific estimates of α. Here every point indicates an α estimate for the specific frequency class

Low-frequency classes are expected mainly to consist of slightly deleterious or deleterious mutations; hence their contribution towards α is negative. A negative α likely indicates an excess of variants segregating within a population (in this case for a specific SFS class) as compared to the fixed differences between two species. With an increase in frequency, the α estimates seem to increase. However, for higher frequencies, α estimates seem to be again lower, which could be caused due to mis-polarization.

templeRun – a wrapper around TEMPLE for automating the processing and analysis of genetic variants in the TFBS regions

Introduction

Given the central role of the gene regulatory elements within the overall biochemical machinery of a cell, variants occurring within these could be potentially responsible for differential expression patterns of the effector genes. Selection-based studies use the information on these genetic variants to infer the influence of evolutionary forces in action. Performing such studies on the coding region sequences is relatively straightforward due to a known genetic code. However, due to the absence of a similar genetic code for the noncoding elements, performing selection-based studies and inferring functional noncoding elements is challenging. Overall, previous studies have relied on two approaches for inferring the proportion of functional noncoding elements: biochemical signature- and conservation-based. The former relies on biochemical signatures resulting from biochemical assays (for example, ChIP-seq, ATAC-seq). It uses these as a proxy for inferring functionality. Recent advances in sequencing and assay technologies have resulted in an exponential increase in the availability of such data. However, deducing functionality from such signatures could result in false positives and inflate the proportions of the inferred functional noncoding elements (Graur et al. 2013; Doolittle 2013). The conservation-based approach identifies noncoding elements that are conserved across a population or on the level of a phylogeny. Hence, conservation is used as a proxy for functionality. However, this approach would potentially be unable to detect elements under positive selection (Ludwig et al. 2000; Dermitzakis and Clark 2002).

TEMPLE (Litovchenko and Laurent 2016) is a bioinformatics tool that uses the information from biochemical assay experiments and protein binding models to study the diversity within the Transcription Factor binding sites (TFBSs). This tool primarily predicts the exact intervals of the TFBS within a given stretch of sequence. This tool further highlights the population-specific variants using the individual-specific sequence information from multiple populations. Here we introduce *templeRun*, a wrapper around TEMPLE. On the level of data processing, this package mainly prepares the required input files and executes TEMPLE internally. On the level of output processing, this package collects the resulting information, processes them and finally calculates summary statistics (constraint ratios) which are used in making relevant inferences. The following sections are focused on capturing the overall structure of *templeRun*. For reproducibility, the code and example input files discussed in this chapter are available on the following GitLab page – (**gitlab.mpcdf.mpg.de/mjoshi/forthesis_tools/**)

Requirements and input files

templeRun is majorly written in R (R Core Team 2002) and is compatible with version *4.2.0* and above. This package internally generates system calls to external tools and incorporates those outputs. Following are the R packages and external tools used by *templeRun* that are used in this example (R version *4.2.0*):

- *httr* (version *1.4.3*) – R package
- *jsonlite* (version *1.8.0*) – R package
- *xml2* (version *1.3.3*) – R package
- *Biostrings* (version *2.64.0*) – R package
- *stringr* (version *1.4.0*) – R package
- *rlist* (version *0.4.6.2*) – R package
- *ggplot2* (version – *3.3.6*)
- *VCF-kit* (version *0.2.9*)
- *samtools* (version *1.6*)
- *vcftools* (version *0.1.14*)
- *TEMPLE* (version *1.0*)

In addition to the dependencies, *templeRun* uses the following input files:

- A population-level variant annotation file (in *.vcf* format)
- Whole genome sequence file of the ingroup (in *.fa* format)
- Count matrices for the TFs of interest in a single file (compatible with TEMPLE)
- Coordinates of noncoding regions (in a tabular format)
- Population-specific accessions (in a tabular format, one file per population)

The flow of information within *templeRun*

This package is centred around a single function – *temple_run.R*. This function serves two critical roles. First, this function is the central point for importing relevant libraries and functions, setting paths to the accessions, noncoding coordinates and binding motif files, and declaring constants. Secondly, this function compartmentalises the essential functions into a set of steps, processes individual noncoding coordinates through these steps sequentially and internally generates system calls for TEMPLE. Following are the steps in data processing through *templeRun*:

- Step 1 – Extracting the population-specific subset of variants for the given coordinates

  This step executes the *get_popspecific_vcf_subset* function. TEMPLE enables analysis of genetic diversity occurring on the TFBS on two populations in a single analysis. This function accesses the population-level variant annotation file (*.vcf* file), the population-specific accession information and the noncoding coordinates to generate a subset *vcf* file per population for the given coordinates. The population-specific subset of the *.vcf* is stored in backup_files/vcf_files/.

- Step 2 – Constructing the sequence input file for TEMPLE

  This step executes the *make_sequences_for_temple* function. This function utilises the population-specific *vcf* file generated in the previous step to identify the varying sites within the noncoding region of interest. It extracts the entire region's wild-type sequence from the ingroup species' reference genome. Next, it generates a sequence per individual accession using the wild-type sequence information and substituting the reported variant sites for the specific accession. Hence, this function constructs sequence per accessions for both populations. Finally, using the coordinates of the noncoding region, this function retrieves the sequence of the outgroup species from the whole genome alignments. The whole genome alignment is retrieved through the REST-API functionality of Ensembl (cite Ensembl). The outgroup sequence is merged with the sequence from the populations. This list of sequences is saved in the folder backup_files/sequence_files/ using a unique identifier that combines the coordinates of the noncoding sequence.

- Step 3 – Constructing the region input file for TEMPLE

  This step executes the *make_regionfile_for_temple* function. This function uses the coordinates for the noncoding regions to construct the required region file for TEMPLE. The region file is saved in the folder backup_files/region_files/ using a unique identifier that combines the coordinates of the noncoding sequence.

- Step 4 – Executing TEMPLE

  This final step is executed within *temple_run.R*. The required input files are first imported from their respective destination folder. Then a system call is generated to execute TEMPLE by supplying these files.

Example outputs

The output files from TEMPLE are first aggregated using this function analysis/outputs_aggregate/aggregate_mutationfile_outputs.R. Following aggregation, first, the population-specific variants are first split. A *ratio score* (**see Materials and Methods section for chapter 2**) is calculated per reported variant. Additionally, background lengths are calculated for all the positions within the reported PWMs using the ancestral allele information to normalise the proportion of reported variants. Finally, noncoding nonsynonymous variants are identified based on their ratio score metric and the counts of the reference and alternate alleles from the PWM. The functions executing these three steps are stored in the analysis/ folder. Some of the outputs from the example are discussed here:

- Constraint ratios

  On identifying the class nonsynonymous equivalent variants, and using the synonymous variants from the coding regions, *templeRun* constructs the constraint ratios and other summary statistics. In the case of this example, the polymorphism and divergence constraint ratios are highlighted in the **Table 3.2 (a and b)**

π_n (noncoding)	π_s (coding)	π_n/π_s
0.0011	0.0046	0.3275

Table 3.2 (a) – Polymorphism constraint ratios

K_n (noncoding)	K_s (coding)	K_n/K_s
0.0295	0.1373	0.2430

Table 3.2 (b) – Divergence constraint ratios

On comparing the π_n/π_s ratios for the two populations, it could be observed that the derived population NS (North Sweden) is under a comparatively relaxed constraint as compared to the older population IB (Iberia). The relaxed constraint could be explained due to a comparatively elevated π_n and comparatively lower π_s for the NS population.

- Site frequency spectrum (SFS) plot

In addition to the constraint ratios, *templeRun* also produces a plot depicting the distribution of nonsynonymous allele frequencies for the two populations. In the case of this example, the SFS plot is shown in **Figure 3.3**

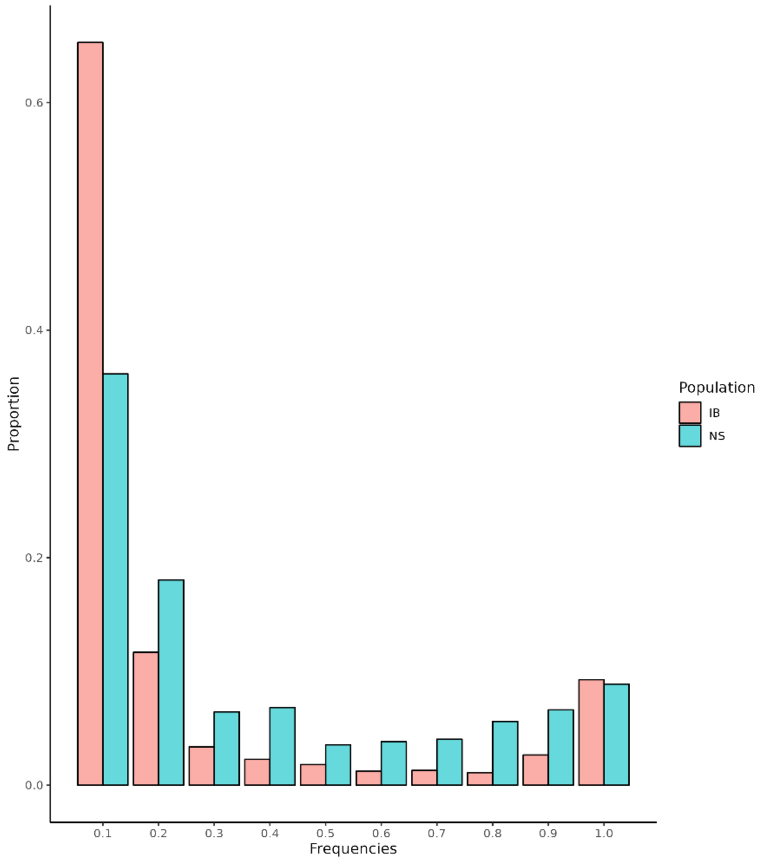


Figure 3.3 – Site frequency spectrum plot for comparing the population-specific proportions of nonsynonymous mutations in the TFBS regions

From the SFS plot, it could be observed that the IB population harbours a larger proportion of low-frequency nonsynonymous alleles than the NS population. This observation could explain the comparatively lower π_n/π_s estimates for the IB population as compared to the NS population. In addition, IB also harbours a slightly higher proportion of nonsynonymous alleles segregating in high frequency.

- Estimating α values for the TFBS regions

The raw number of variants collected from the TFBS regions is relatively less than the entire gene set (**described in chapter 3.1**). Hence, calculating a frequency class-specific α would not be feasible. To overcome this challenge of fewer variants, *templeRun* performs pooled α estimate for the TFBS regions using frequency cutoffs that are derived from the species-specific WGS analysis from **Chapter 1**. The population-specific α estimates for this example are shown in **Table 3.3**

α estimates	**IB population**	**NS population**
	0.39	-0.96

Table 3.3 – TFBS-specific *α estimates per population*

The estimates indicate a comparatively higher α estimate for the IB population than the NS population. Hence, the ancestral population (IB) is highlighted to be under higher constraints and a comparatively higher intensity of positive selection than the derived population (NS).

General Discussion

Given their central role in the transition of genotype to phenotype, GREs would be expected to be under a stronger selection influence than the overall genomic background. This study highlighted the intensity of natural selection acting on the domains participating in the regulatory TF-DNA interactions, explicitly focusing on the motifs participating in these interactions. Employing a population- and comparative genomics-based approach enabled us to test the signal of selection across the two evolutionary timescales.

Insights from the analysis of the DNA-binding domains

On the level of the TFs, we focused on the DNABDs, which are stretches of sequences that directly interact with the DNA molecules for gene regulation. We were able to highlight a consistent signal of high constraint acting on these motifs, suggesting an increased intensity of purifying selection. This signal was consistent across the non-DNABD and WGS control regions included in the study. Here, the non-DNABD control regions were used to counter differential recombination rates influencing the signature of selection. On the other hand, WGS control regions were used to contrast the signature of selection against the overall coding region average. The pleiotropic nature of the TFs could explain the high constraint. In the context of a gene regulatory network (GRN), TFs are often observed to control the expression of multiple target genes (Chesmore et al. 2016). This activity of controlling the gene expression patterns of multiple target genes is carried out via the DNABDs. Hence, introducing a variant within these regions could consequently impact multiple downstream regulatory interactions and be potentially detrimental to fitness. We supplement this hypothesis by employing available deleterious variants annotation data for *H. sapiens*, ClinVar (Landrum et al. 2018). Using the annotation data, we were able to show that the DNABD regions harbour a significantly higher proportion of “pathogenic” variants as compared to the non-DNABD regions. This finding further cemented the signal of high constraint.

Next, we investigated the action of positive selection using *asymptoticMK* (Haller and Messer 2017). In the case of WGS regions, the estimates for the proportion of adaptive substitutions (α) obtained per species were in agreement with some of the previously reported studies (Andolfatto 2005; Eyre-Walker and Keightley 2007; Moutinho, Bataillon, and Dutheil 2019). Interestingly, in the case of *H. sapiens,* the estimates of α for the WGS regions were consistently negative. This observation suggests that the nonsynonymous mutations occurring within populations are, overall, contributing more towards the within-species differences. The DNABD and non-DNABD regions are shorter in terms of their genomic lengths than the WGS regions. Hence the count of variants obtained in these regions is also relatively smaller. This consequently resulted in high variances in the α estimates obtained from the *asymptoticMK*-based approach. To counter this, we employed a hybrid approach of the traditional MK-test (McDonald and Kreitman 1991) along with *asymptoticMK*. For species with larger N_e, namely *A. thaliana* and *D. melanogaster*, we report a comparatively higher α for the DNABD regions than the control regions in certain populations. Interestingly, the α estimates for all the *H. sapiens* populations were consistently negative. This suggests that, in the case of *H. sapiens*, variants occurring within the DNABD regions actively contribute towards the differences within-species differences compared to the between-species differences.

Insights from the analysis of the Transcription Factor Binding Sites

On the level of the noncoding DNA, we focused on the TFBS regions, which are stretches of sequences to which the TFs are annotated to bind to initiate the transcription process. One central aim of this study was to compare the intensities of selection acting on DNABD and TFBS regions. Given that the TFBS regions are usually a part of the noncoding genome, we developed a metric (*ratio score*) which enabled us to compare the intensity of selection acting on the regulatory elements in the noncoding regions to the elements occurring in the coding regions. We proposed a method of identifying a “nonsynonymous” equivalent class of variants within the TFBS regions, which are identified depending on their potential impact on the binding affinity. He *et al.* (2011) (He et al. 2011) have also used a similar approach. In this study, we extend this approach across multiple species and populations. To control for demographic factors influencing the detected signals, we used the synonymous sites occurring within the coding regions as the putative neutral sites. Previous studies have highlighted that the overall levels of constraint acting on the noncoding regions are lower than those acting on the coding regions (Naidoo et al. 2018; Haddrill, Bachtrog, and Andolfatto 2008; Torgerson et al. 2009). However, on the polymorphism scale, we observe comparable levels of constraint acting on the TFBS compared to the non-DNABD and WGS regions for the ancestral population (ZAM) of species with comparatively low drift, *D. melanogaster*. This signal is further accentuated on the divergence scale, where we identify higher levels of constraint acting on the TFBS regions compared to the non-DNABD and WGS regions. The signal of comparable levels of constraint on the TFBS regions seems to fade for species with comparatively higher levels of drift (*A. thaliana* and *H. sapiens*). We also capture a consistent signal of a comparatively lower constraint acting on the TFBS in the species-specific derived populations compared to the ancestral populations.

Next, we calculated the TFBS region-specific α estimates per population. We highlight that, overall, the TFBS are under a lower intensity of positive selection compared to the coding regions, with an exception occurring in the case of the Iberian population of *A. thaliana*. We also highlight lower α estimates for the species-specific derived populations than their ancestral populations.

We identified a consistent signal of high constraint acting on the DNABD regions compared to the TFBS regions. We also report that the DNABD regions are under a comparatively stronger influence of positive selection than the TFBS regions. To summarize, we identified that the regulatory regions occurring on TFs are under a comparatively higher intensity of purifying and positive selection than the regulatory TFBS regions.

Insights from correlating the efficiency of selection with N_e

We also investigated the correlation between the species-specific N_e and the efficiency of selection. We used the species-specific proportion of neutral variants, π_s, as a proxy for N_e (Galtier 2016b; James, Castellano, and Eyre-Walker 2016). Additionally, we used the polymorphism constraint ratio (π_n/π_s) and α as the proxies for testing the efficiencies of purifying and positive selection, respectively. In the case of purifying selection, we observed an overall negative correlation between π_n/π_s and π_s, suggesting an increased efficiency of purifying selection with an increased N_e. Here, we observed the strongest correlation for the DNABD regions. In the case of positive selection, we observed an overall positive correlation between α and π_s, suggesting an increased efficiency of positive selection with an increased N_e. Here, we observed the strongest correlation for the WGS regions. We observed the poorest correlation for functionally unannotated non-DNABD regions in both cases.

Abbreviations and summary statistics

TF – Transcription Factors

DNABD – DNA binding domains

TFBS – Transcription Factor binding sites

GREs – Gene regulatory elements

CREs – cis-regulatory elements

TREs – trans-regulatory elements

CRM – cis-regulatory modules, hotspot for the binding activity of multiple TFs

non-DNABD – functionally unannotated regions within the Transcription Factors

WGS – Whole gene sets

YRI – Yoruba in Ibadan, Nigeria

CEU – Utah residents (CEPH) with Northern and Western European ancestry

CHS – Southern Han Chinese

IB – Iberia

NS – North Sweden

CA – Central Asia

ZAM – Zambia

SWE – Sweden

π – measure of diversity

π_n – proportion of nonsynonymous variants within species

π_s – proportions of synonymous variants within species

$\pi_{nonsense}$ – proportions of nonsense variants within species

π_n/π_s – nonsynonymous polymorphism constraint ratio

$\pi_{nonsense}/\pi_s$ – nonsense polymorphism constraint ratio

K_n – proportion of nonsynonymous variants with the outgroup

K_s – proportions of synonymous variants with the outgroup

$K_{nonsense}$ – proportions of nonsense variants with the outgroup

K_n/K_s – nonsynonymous divergence constraint ratio

$K_{nonsense}/K_s$ – nonsense divergence constraint ratio

MK test – McDonald-Kretimann test

α – proportion of adaptive substitutions

N_e – species-specific effective population size

μ – mutation rate

List of figures

List of Tables

Bibliography

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. "From DNA to RNA." https://www.ncbi.nlm.nih.gov/books/NBK26887/.

Alonso-Blanco, Carlos, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M M. Borgwardt, Jun Cao, et al. 2016a. "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis Thaliana." *Cell* 166 (2): 481–91. https://doi.org/10.1016/j.cell.2016.05.063.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andolfatto, Peter. 2005. "Adaptive Evolution of Non-Coding DNA in Drosophila." *Nature* 437 (7062): 1149–52. https://doi.org/10.1038/nature04107.

Arbiza, Leonardo, Ilan Gronau, Bulent A Aksoy, Melissa J Hubisz, Brad Gulko, Alon Keinan, and Adam Siepel. 2013. "Genome-Wide Inference of Natural Selection on Human Transcription Factor Binding Sites." *Nat Genet* 45 (7): 723–29. https://doi.org/10.1038/ng.2658.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, et al. 2015a. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Barghi, Neda, Joachim Hermisson, and Christian Schlötterer. 2020. "Polygenic Adaptation: A Unifying Framework to Understand Positive Selection." *Nature Reviews Genetics* 21 (12): 769–81. https://doi.org/10.1038/s41576-020-0250-z.

Bateman, Alex, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, et al. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1): D158–69. https://doi.org/10.1093/NAR/GKW1099.

Buenrostro, Jason, Beijing Wu, Howard Chang, and William Greenleaf. 2016. "ATAC-Seq Method." *Current Protocols in Molecular Biology* 2015: 1–10. https://doi.org/10.1002/0471142727.mb2129s109.ATAC-seq.

Byrska-Bishop, Marta, Uday S. Evani, Xuefang Zhao, Anna O. Basile, Haley J. Abel, Allison A. Regier, André Corvelo, et al. 2022. "High-Coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios." *Cell* 185 (18): 3426-3440.e19. https://doi.org/10.1016/J.CELL.2022.08.004.

Castro-Mondragon, Jaime A., Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, et al. 2022a. "JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 50 (D1): D165–73. https://doi.org/10.1093/NAR/GKAB1113.

Chan, Yingguang Frank, Melissa E. Marks, Felicity C. Jones, Guadalupe Villarreal, Michael D. Shapiro, Shannon D. Brady, Audrey M. Southwick, et al. 2010. "Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitxl Enhancer." *Science* 327 (5963): 302–5. https://doi.org/10.1126/science.1182213.

Chesmore, Kevin N., Jacquelaine Bartlett, Chao Cheng, and Scott M. Williams. 2016. “Complex Patterns of Association between Pleiotropy and Transcription Factor Evolution” 8 (10): 3159–70. https://doi.org/10.1093/GBE/EVW228.

Choudhuri, Supratim. 2014. “Fundamentals of Molecular Evolution.” *Bioinformatics for Beginners*, 27–53. https://doi.org/10.1016/B978-0-12-410471-6.00002-5.

Clark, Andrew G., Michael B. Eisen, Douglas R. Smith, Casey M. Bergman, Brian Oliver, Therese A. Markow, Thomas C. Kaufman, et al. 2007. “Evolution of Genes and Genomes on the Drosophila Phylogeny.” *Nature 2007 450:7167* 450 (7167): 203–18. https://doi.org/10.1038/nature06341.

Connelly, Caitlin F., Daniel A. Skelly, Maitreya J. Dunham, and Joshua M. Akey. 2013. “Population Genomics and Transcriptional Consequences of Regulatory Motif Variation in Globally Diverse Saccharomyces Cerevisiae Strains.” *Molecular Biology and Evolution* 30 (7): 1605–13. https://doi.org/10.1093/molbev/mst073.

Consortium, International Human Genome Sequencing. 2001. “Initial Sequencing and Analysis of the Human Genome International Human Genome Sequencing Consortium*.” *MacMillan Magazines Ltd*, 2001.

Consortium, The, Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, et al. 2011. “Identification of Functional Elements and Regulatory Circuits by Drosophila ModENCODE.” *Science* 330 (6012): 1787–97. https://doi.org/10.1126/science.1198374.Identification.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. 2022. “Ensembl 2022.” *Database Issue Nucleic Acids Research* 50: 989. https://doi.org/10.1093/nar/gkab1049.

Dermitzakis, Emmanouil T., and Andrew G. Clark. 2002. “Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover.” *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a004169.

Desvergne, Béatrice, Liliane Michalik, and Walter Wahli. 2006. “Transcriptional Regulation of Metabolism.” *Physiological Reviews* 86 (2): 465–514. https://doi.org/10.1152/PHYSREV.00025.2005.

Doolittle, W. Ford. 2013. “Is Junk DNA Bunk? A Critique of ENCODE.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (14): 5294–5300. https://doi.org/10.1073/pnas.1221376110.

Edgar, Robert C. n.d. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” Accessed February 9, 2023. https://doi.org/10.1093/nar/gkh340.

Enattah, Nabil Sabri, Timo Sahi, Erkki Savilahti, Joseph D. Terwilliger, Leena Peltonen, and Irma Järvelä. 2002. “Identification of a Variant Associated with Adult-Type Hypolactasia.” *Nature Genetics*. https://doi.org/10.1038/ng826.

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

Eyre-Walker, Adam, and Peter D Keightley. n.d. “Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change.” Accessed February 9, 2023. https://doi.org/10.1093/molbev/msp119.

Galtier, Nicolas. 2016a. “Adaptive Protein Evolution in Animals and the Effective Population Size

Hypothesis." *PLOS Genetics* 12 (1): e1005774. https://doi.org/10.1371/JOURNAL.PGEN.1005774.

Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B.R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. 2013. "On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of Encode." *Genome Biology and Evolution* 5 (3): 578–90. https://doi.org/10.1093/gbe/evt028.

Guéguen, Laurent, and Laurent Duret. n.d. "Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition." Accessed February 9, 2023. https://doi.org/10.1093/molbev/msx308.

Haddrill, Penelope R., Doris Bachtrog, and Peter Andolfatto. 2008. "Positive and Negative Selection on Noncoding DNA in Drosophila Simulans." *Molecular Biology and Evolution* 25 (9): 1825–34. https://doi.org/10.1093/molbev/msn125.

Haller, Benjamin C., and Philipp W. Messer. 2017. "AsymptoticMK: A Web-Based Tool for the Asymptotic McDonald-Kreitman Test." *G3: Genes, Genomes, Genetics* 7 (5): 1569–75. https://doi.org/10.1534/g3.117.039693.

Hammal, Fayrouz, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, and Benoit Ballester. 2021. "ReMap 2022: A Database of Human, Mouse, Drosophila and Arabidopsis Regulatory Regions from an Integrative Analysis of DNA-Binding Sequencing Experiments." *Nucleic Acids Research*, November. https://doi.org/10.1093/NAR/GKAB996.

He, Bin Z., Alisha K. Holloway, Sebastian J. Maerkl, and Martin Kreitman. 2011. "Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules." *PLoS Genetics* 7 (4). https://doi.org/10.1371/journal.pgen.1002053.

Hu, Tina T., Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan Fang Cheng, Richard M. Clark, Noah Fahlgren, et al. 2011. "The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change." *Nature Genetics* 43 (5): 476–83. https://doi.org/10.1038/NG.807.

Huang, Yi Fei, Brad Gulko, and Adam Siepel. 2017. "Fast, Scalable Prediction of Deleterious Noncoding Variants from Functional and Population Genomic Data." *Nature Genetics* 49 (4): 618–24. https://doi.org/10.1038/ng.3810.

Ihn Lee, Tong, and Richard A Young. 2013. "Leading Edge Review Transcriptional Regulation and Its Misregulation in Disease." https://doi.org/10.1016/j.cell.2013.02.014.

Jacob, FRANÇOIS, and JACQUES Monod. 1978. *Genetic Regulatory Mechanisms in the Synthesis of Proteins*. *Selected Papers in Molecular Biology by Jacques Monod*. Vol. 3. Academic Press. https://doi.org/10.1016/b978-0-12-460482-7.50042-7.

James, J, D Castellano, and A Eyre-Walker. 2016. "DNA Sequence Diversity and the Efficiency of Natural Selection in Animal Mitochondrial DNA." *Heredity* 118: 88–95. https://doi.org/10.1038/hdy.2016.108.

Jin, Jinpu, Feng Tian, De-Chang Yang, Yu-Qi Meng, Lei Kong, Jingchu Luo, and Ge Gao. 2017. "PlantTFDB 4.0: Toward a Central Hub for Transcription Factors and Regulatory Interactions in Plants" 45. https://doi.org/10.1093/nar/gkw982.

Kapopoulou, Adamandia, Martin Kapun, Bjorn Pieper, Pavlos Pavlidis, Ricardo Wilches, Pablo Duchen, Wolfgang Stephan, and Stefan Laurent. 2020. "Demographic Analyses of a New Sample of Haploid Genomes from a Swedish Population of Drosophila Melanogaster" 10 (1): 1–8. https://pubmed.ncbi.nlm.nih.gov/33376238/.

Karabacak Calviello, Asllhan, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. 2019. “Reproducible Inference of Transcription Factor Footprints in ATAC-Seq and DNase-Seq Datasets Using Protocol-Specific Bias Modeling.” *Genome Biology* 20 (1): 42. https://doi.org/10.1186/s13059-019-1654-y.

King, MC, and AC Wilson. 1975. “Evolution at Two Levels in Humans and Chimpanzees.” *Science* 188 (4184): 107–16. https://doi.org/10.1126/SCIENCE.1090005.

Kosakovsky Pond, Sergei L., Simon D.W. Frost, and Spencer V. Muse. 2005. “HyPhy: Hypothesis Testing Using Phylogenies.” *Bioinformatics* 21 (5): 676–79. https://doi.org/10.1093/bioinformatics/bti079.

La Calle-Mustienes, Elisa De, Cármen Gloria Feijóo, Miguel Manzanares, Juan J. Tena, Elisa Rodríguez-Seguel, Annalisa Letizia, Miguel L. Allende, and José Luis Gómez-Skarmeta. 2005. “A Functional Survey of the Enhancer Activity of Conserved Non-Coding Sequences from Vertebrate Iroquois Cluster Gene Deserts.” *Genome Research* 15 (8): 1061–72. https://doi.org/10.1101/gr.4004805.

Lack, Justin B., Charis M. Cardeno, Marc W. Crepeau, William Taylor, Russell B. Corbett-Detig, Kristian A. Stevens, Charles H. Langley, and John E. Pool. 2015. “The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila Melanogaster Genomes, Including 197 from a Single Ancestral Range Population.” *Genetics* 199 (4): 1229–41. https://doi.org/10.1534/GENETICS.115.174664.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67. https://doi.org/10.1093/NAR/GKX1153.

Lawrie, David S., and Dmitri A. Petrov. 2014. *No Title*. Vol. 30. https://pubmed.ncbi.nlm.nih.gov/24656563/.

Lewontin, R. C. n.d. *THE GENETIC BASIS OF EVOLUTIONARY CHANGE* .

Li, Jing, and Changning Liu. 2019. “Coding or Noncoding, the Converging Concepts of RNAs.” *Frontiers in Genetics* 10 (MAY): 1–10. https://doi.org/10.3389/fgene.2019.00496.

Litovchenko, Maria, and Stefan Laurent. 2016. “TEMPLE: Analysing Population Genetic Variation at Transcription Factor Binding Sites.” *Molecular Ecology Resources* 16 (6): 1428–34. https://doi.org/10.1111/1755-0998.12535.

Ludwig, Michael Z., Casey Bergman, Nipam H. Patel, and Martin KreLtman. 2000. “Evidence for Stabilizing Selection in a Eukaryotic Enhancer Element.” *Nature* 403 (6769): 564–67. https://doi.org/10.1038/35000615.

McDonald, John H., and Martin Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

Mikkelsen, Tarjei S., Ladeana W. Hillier, Evan E. Eichler, Michael C. Zody, David B. Jaffe, Shiaw Pyng Yang, Wolfgang Enard, et al. 2005. “Initial Sequence of the Chimpanzee Genome and Comparison with the Human Genome.” *Nature 2005 437:7055* 437 (7055): 69–87. https://doi.org/10.1038/nature04072.

Morales, Joannella, Shashikant Pujar, Jane E. Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, et al. 2022. “A Joint NCBI and EMBL-EBI Transcript Set for Clinical Genomics and Research.” *Nature 2022 604:7905* 604 (7905): 310–15. https://doi.org/10.1038/s41586-022-04558-8.

Moutinho, Ana Filipa, Thomas Bataillon, and Julien Y. Dutheil. 2019. “Variation of the Adaptive Substitution Rate between Species and within Genomes.” *Evolutionary Ecology 2019 34:3* 34 (3): 315–38. https://doi.org/10.1007/S10682-019-10026-Z.

Mu, Xinmeng Jasmine, Zhi John Lu, Yong Kong, Hugo Y.K. Lam, and Mark B. Gerstein. 2011. “Analysis of Genomic Variation in Non-Coding Elements Using Population-Scale Sequencing Data from the 1000 Genomes Project.” *Nucleic Acids Research* 39 (16): 7058–76. https://doi.org/10.1093/nar/gkr342.

Naidoo, Thijessen, Per Sjödin, Carina Schlebusch, and Mattias Jakobsson. 2018. “Patterns of Variation in Cis-Regulatory Regions: Examining Evidence of Purifying Selection.” *BMC Genomics* 19 (1): 1–14. https://doi.org/10.1186/s12864-017-4422-y.

Nei, Masatoshi, and Wen-Hsiung Li. 1979. “Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases (Molecular Evolution/Mitochondrial DNA/Nucleotide Diversity).” *Genetics* 76 (10): 5269–73.

Park, Peter J. 2009. “ChIP-Seq: Advantages and Challenges of a Maturing Technology.” *Nature Reviews Genetics* 10 (10): 669–80. https://doi.org/10.1038/nrg2641.

Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. “In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences.” *Nature* 444 (7118): 499–502. https://doi.org/10.1038/nature05295.

Perdomo-Sabogal, Álvaro, Katja Nowick, and David Enard. 2019. “Genetic Variation in Human Gene Regulatory Factors Uncovers Regulatory Roles in Local Adaptation and Disease.” *Genome Biology and Evolution* 11 (8): 2178–93. https://doi.org/10.1093/gbe/evz131.

Rand, David M., and Lisa M. Kann. 1996. “Excess Amino Acid Polymorphism in Mitochondrial DNA: Contrasts among Genes from Drosophila, Mice, and Humans.” *Molecular Biology and Evolution* 13 (6): 735–48. https://doi.org/10.1093/oxfordjournals.molbev.a025634.

Rands, Chris M., Stephen Meader, Chris P. Ponting, and Gerton Lunter. 2014. “8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage.” *PLoS Genetics* 10 (7). https://doi.org/10.1371/journal.pgen.1004525.

Sandelin, Albin, Peter Bailey, Sara Bruce, Pär G. Engström, Joanna M. Klos, Wyeth W. Wasserman, Johan Ericson, and Boris Lenhard. 2004. “Arrays of Ultraconserved Non-Coding Regions Span the Loci of Key Developmental Genes in Vertebrate Genomes.” *BMC Genomics* 5: 1–9. https://doi.org/10.1186/1471-2164-5-99.

Schlenke, Todd A., and David J. Begun. 2004. “Strong Selective Sweep Associated with a Transposon Insertion in Drosophila Simulans.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (6): 1626–31. https://doi.org/10.1073/pnas.0303793101.

Sigrist, Christian J.A., Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. 2002. “PROSITE: A Documented Database Using Patterns and Profiles as Motif Descriptors.” *Briefings in Bioinformatics* 3 (3): 265–74. https://doi.org/10.1093/BIB/3.3.265.

Sullivan, Alessandra M., Kerry L. Bubb, Richard Sandstrom, John A. Stamatoyannopoulos, and Christine Queitsch. 2015. “DNase I Hypersensitivity Mapping, Genomic Footprinting, and Transcription Factor Networks in Plants.” *Current Plant Biology* 3–4: 40–47. https://doi.org/10.1016/j.cpb.2015.10.001.

The ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human

Genome.” *Nature* 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Torgerson, Dara G., Adam R. Boyko, Ryan D. Hernandez, Amit Indap, Xiaolan Hu, Thomas J. White, John J. Sninsky, et al. 2009. “Evolutionary Processes Acting on Candidate Cis-Regulatory Regions in Humans Inferred from Patterns of Polymorphism and Divergence.” *PLoS Genetics* 5 (8). https://doi.org/10.1371/journal.pgen.1000592.

“UniProt: The Universal Protein Knowledgebase in 2023 The UniProt Consortium.” 2022. *Nucleic Acids Research* 51: 523–31. https://doi.org/10.1093/nar/gkac1052.

Uricchio, Lawrence H, Dmitri A Petrov, and David Enard. n.d. “Exploiting Selection at Linked Sites to Infer the Rate and Strength of Adaptation.” *Nature Ecology & Evolution*. Accessed February 9, 2023. https://doi.org/10.1038/s41559-019-0890-6.

Vernot, Benjamin, Andrew B. Stergachis, Matthew T. Maurano, Jeff Vierstra, Shane Neph, Robert E. Thurman, John A. Stamatoyannopoulos, and Joshua M. Akey. 2012. “Personal and Population Genomics of Human Regulatory Variation.” *Genome Research* 22 (9): 1689–97. https://doi.org/10.1101/gr.134890.111.

Yang, Ziheng, and Rasmus Nielsen. 2000. “Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models.” *Mol. Biol. Evol* 17 (1): 32–43. https://academic.oup.com/mbe/article/17/1/32/975527.

Yates, Andrew, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R.S. Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. 2014. “The Ensembl REST API: Ensembl Data for Any Language.” *Bioinformatics (Oxford, England)* 31 (1): 143–45. https://doi.org/10.1093/BIOINFORMATICS/BTU613.

Zhen, Ying, and Peter Andolfatto. 2012. “Methods to Detect Selection on Noncoding DNA.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-61779-585-5_6.

Acknowledgements

This thesis would not have reached this point without the contribution of many people. I would first like to thank my direct supervisor, Dr Stefan Laurent, for introducing me to the field of Population Genetics and for being patient as I explored it. He has always been open to new ideas I could bring to the table. I would also like to thank Prof Dr Miltos Tsiantis for supporting my project and for hosting me in the department. I would also like to thank Prof Dr Katja Nowick and Dr Ruben Garrido-Oter, my TAC members, for their valuable suggestions in steering this project. I would also like to thank Prof Dr Miltos Tsiantis, Prof Dr Thomas Wiehe and Prof Dr Andreas Beyer for agreeing to be a part of my thesis examination committee.

The three and half years of stay were truly made memorable by the current and former lab members of *grp_laurent*– Danijel, Danijela, Maria, Purva, Rachita, Stefan and Yasir. I want to thank all for the scientific discussions. I thank Danijela for helping me with the *D. melanogaster* dataset. I would also like to thank Maria for using *Alag* and giving her valuable suggestions for improving the pipeline and writing suggestions. In addition to the group members, I would also like to thank all the members of our department for their support throughout my PhD. I want to thank our graduate school office and our PhD coordinator, Dr Stephan Wagner, who provided a great deal of support to all the PhD students. I would also like to thank our institute's administrative and IT staff for providing constant support. The support provided by the technical experts from Ensembl and UniProt has made a significant contribution to this project, and I would like to thank them too.

This PhD would particularly not have been possible without three individuals. My parents, who always supported me in pursuing my career. Their sacrifices are invaluable. Madhura, my partner, for believing in me and being someone I can rely on. Also, a special thank you to all my friends here and in India.

Finally, I would like to thank you, the *reader*.

List of publications

Review paper:

Joshi, M., Kapopoulou, A., & Laurent, S. (2021). Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective. Frontiers in Genetics, 12(July), 1–10

Research paper:

Joshi, M., Laurent, S. (2023). Comparative and population genomics analysis of regulatory domains participating in the TF-DNA interactions (in preparation)