UNDERSTANDING EVOLUTION WITH SIMULATIONS: THREE TALES ABOUT TREES

by

MURILLO FERNANDO RODRIGUES

A dissertation accepted and approved in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in Biology

Dissertation Committee:
Matt Streisfeld, Chair
Andrew Kern, Co-advisor
Peter Ralph, Co-advisor
Patrick Phillips, Core Member
Jayson Paulose, Institutional Representative

University of Oregon
Winter 2024

DISSERTATION ABSTRACT

Murillo Fernando Rodrigues

Doctor of Philosophy

Biology

Winter 2024

Title: Understanding Evolution with Simulations: Three Tales about Trees

Evolutionary processes impact patterns of genetic variation, so there is an opportunity to reverse this relationship and use genetic data to learn about past evolutionary events. Traditional evolutionary inference from genetic data is plagued by a few issues: (i) different processes can impact a particular feature in similar ways, making it difficult to disentangle them; (ii) there is a growing need for modeling interactions between processes; and (iii) many models do not make full use of genomic data and instead assume that loci are unlinked.

Simulation-based evolutionary inference can help alleviate many of these issues. It is now possible to simulate complex evolutionary scenarios, and these can be used to approximate analytically intractable likelihoods, for example by using supervised machine learning. The major downsides to using simulations is the computational cost, but recent advancements both in hardware and software have lessened this cost. In this dissertation, I pushed the boundaries on how simulation-based inference can be applied in evolutionary genetics.

To mitigate the costs associated with simulations, I made a few contributions to the `tskit` ecosystem of evolutionary simulation tools. First, I developed a way to partially parallelize the simulation of multiple populations to

make inference using multi-population genomic datasets more feasible. Second, I helped create standards for reproducible simulations with natural selection within the `stdpopsim` consortium. I implemented the ability to simulate using previously published distribution of fitness effects (DFEs) and to simulate selective sweeps. I demonstrate the utility of this tool by tackling the long-standing question of whether the power to detect sweeps varies along realistic chromosomes.

Next, I used simulations to better understand the behavior of a complex multi-population model. Species can be thought as semi-independent realizations of the same (or very similar) evolutionary process. Thus, by looking at multiple species at once it may be possible to better disentangle the processes that shape variation along genomes. Using simulations, I show that positive selection is necessary to explain the genetic data obtained from multiple great ape species. Further, I lay down a framework for leveraging multi-species information to better understand the effects of different processes on a group's evolutionary history.

Lastly, I present a new method that uses whole-genome genealogies for evolutionary inference. This data structure efficiently and sufficiently encodes evolutionary processes. I develop a machine-learning framework, tsNN, that takes whole-genome genealogies as inputs and is flexible enough to perform tasks at different scales (e.g., inferring mutation times, demographic parameters, etc.). I demonstrate that tsNN can learn to predict mutation times accurately, outperforming current likelihood-based methods. tsNN, represents an important step in genealogy-based evolutionary inference, but there still much work to be done in applying deep learning to gain new insights into past evolutionary events.

CURRICULUM VITAE

NAME OF AUTHOR: Murillo Fernando Rodrigues

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Universidade de São Paulo, São Paulo, SP, Brasil

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2024, University of Oregon
Master of Science, Genetics and Evolutionary Biology, 2018, Universidade de São Paulo
Bachelor of Science, Biology, 2015, Universidade de São Paulo

AREAS OF SPECIAL INTEREST:

Evolutionary Biology
Population Genetics
Computational Biology

GRANTS, AWARDS AND HONORS:

Harvey E. Lee Graduate Scholarship, University of Oregon, 2022-2023
Marthe E. Smith Memorial Science Scholarship, University of Oregon, 2022-2023
Hill Fund Graduate Award, University of Oregon, 2020-2021
Genetics Training Grant, University of Oregon, 2019-2021
Research Internship Abroad Fellowship, The São Paulo Research Foundation, 2017-2018
Master's Research Fellowship, The São Paulo Research Foundation, 2016-2018
Undergraduate Research Fellowship, The São Paulo Research Foundation, 2013-2014

PUBLICATIONS:

Rodrigues, M. F., Kern, A. D., & Ralph, P. L. (2024). Shared evolutionary processes shape landscapes of genomic variation in the great apes. *Genetics*, iyae006.

Estevez-Castro, C.F., Rodrigues, M.F., Babarit, A., ... & Olmo, R. P. (2024). Neofunctionalization driven by positive selection led to the retention of the loqs2 gene encoding an *Aedes* specific dsRNA binding protein. *BMC Biology* 22, s12915-024-01821-4.

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., ..., Rodrigues, M. F., ... & Gronau, I. (2023). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife*, 12, RP84874.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., ..., Rodrigues, M. F., ... & Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3), iyab229.

Rodrigues, M. F., Vibranovski, M. D., & Cogni, R. (2021). Clinal and seasonal changes are correlated in *Drosophila melanogaster* natural populations. *Evolution*, 75(8), 2042-2054.

Rodrigues, M. F., & Cogni, R. (2021). Genomic responses to climate change: Making the most of the *Drosophila* model. *Frontiers in Genetics*, 12, 676218.

Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLoS Biology*, 17(7), e3000391.

ACKNOWLEDGEMENTS

With some sense of accomplishment, I am delighted to present this dissertation, which is the culmination of over five years of dedication to expanding our understanding of biology. This journey would not have been possible without the support, guidance, and encouragement of many. I wish I could thank everyone involved personally, but my memory is not what it used to be when I first embarked on this quest... As I reflect on this long process, I would like to offer my gratitude to those who have been instrumental in shaping my academic growth and in keeping my mental sanity.

I am so fortunate to have found not one, but two incredibly kind, patient and generous advisors. Andy, thank you for your enthusiasm for science, for supporting my growth as a scientist, and for your encouragement through ups and downs. Peter, I deeply appreciate your generosity and patience, the valuable and thorough insights, and your steadfast support throughout my time as a PhD student. I am so incredibly thankful for all your effort in making me feel at home here in Oregon. Your dedication to nurturing sense of community and collegiality within our CoLab has been admirable.

Next, I would like to thank my lab mates, collaborators and colleagues who have provided academic and emotional support throughout my journey. Thank you to my amazingly helpful, supportive and fun lab mates: Jeff Adrion, CJ Battey, Jared Galloway, Matt Lukac, Anastasia Teterina, Gabby Coffing, Victoria Caudill, Saurabh Belsare, Vince Buffalo, Gilia Patterson, Chris Smith, Clara Rehmann, Jordan Anderson, Nate Pope, Jiseon Min, Jordan Rodriguez, Georgia Tsambos, Silas Tittes, Scott Small and Bruce Edelman. Nate Pope, in particular, has been

extremely patient and kind in collaborating with me to develop tsNN, herein presented as my fourth chapter. I would like to thank my committee members, Matt Streisfeld, Kelley Harris, Jayson Paulose, and Patrick Phillips for all their support throughout these years. My cohort mates have been so wonderful through the years and helped me make it through: Sabrina Mostoufi, Sophia Frantz, Monika Ruwaimana, Lina Aoyama, Max Spencer, Matt Lukac and Kayla Evans. My friends and former roommates Zac Bush, Matt Lukac and Ethan Shaw made my graduate school experience so much more fun and enjoyable. Clara Rehmann, my lab mate and friend, has been so kind to offer me shelter in Eugene over the past few months. I am thankful to many other members of the Institute of Ecology and Evolution, including the behind the scenes staff that helped me with so many travel and purchase needs over these years; thank you Arlene Crain, Sara Nash, Maria Heider, and Leah Frazier. Thanks are also in order to my always supportive collaborators and members of the tskit ecosystem and the PopSim consortium.

Before my PhD, I was fortunate to encounter amazing individuals who played a pivotal role in shaping me as a scientist and biologist. I would like to thank my undergraduate thesis advisor, Fernando Marques, for helping me get my feet wet with academic research. Rodrigo Cogni, my master's advisor, who helped me make the jump into empirical population genetics. John Pool welcomed me into his lab and gave me incredible freedom to do whatever research I wanted. Paulo Inácio Prado, Alexandre Adalardo de Oliveira and Adriana Martini opened my eyes to the importance of theory, models and simulation in biology. The entire team behind the Atlantic Forest Field Ecology course (offered by the Institute of Biosciences at the Universidade de São Paulo) contributed significantly to my growth as a naturalist.

I want to wholeheartedly thank my family and close friends, whose unwavering encouragement and love have sustained me during the demanding years of my academic pursuit. I have grown so much alongside my long-time friends: Bunni, Carol, Isa, Gabriel and Nat. Looking forward to many more years of adventures and laughter together. My Brazilians in Eugene crew – Bia, Ciro, Giovani, Helena, Jack, Karl, Lívia, Suenia – have made my life in Eugene much more colorful. I would like to thank my siblings, Giovanna and João, for being a source of laughter, support and love. Thank you to my mom, Kelly, for always supporting me with anything I set my mind to. Her strength and determination are inspiring, and I would not have made it through without it. Thank you to my dad, Aguinaldo, for showing me the great things one can accomplish with hard work and grit. I am proud for what my parents accomplished in raising me and my siblings at such a young age with little family support and no college education. I also need to acknowledge my parents for partially supporting me financially throughout my entire academic career, as this journey would have been orders of magnitude harder without it.

Lastly, I want to express gratitude to my partner, Luiza, and my pets, Ollie and Milo, who bring me daily moments of happiness. Luiza, thank you for providing steadfast support and understanding over the past (almost) decade. Your love, caring, patience, and belief in me have been a great source of strength. I could not have done any of this without you by my side anchoring and championing me!

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

1.1 Decoding evolution: a lay-friendly prelude

Evolution can occur over incredibly large spatial scales over the course of thousands to millions of years. This makes Evolutionary Biology unlike other fields within Biology in that hypotheses can not always be subject to experimentation. Thus, we are left with the task of putting together what happened in the past based on information we have today. To do so we need a sizable amount of data, models and statistical and computational tools.

One way to make inferences about past evolutionary processes is by studying genetic variation within and between species. Evolutionary processes, such as population size changes, natural selection and mutation, impact genetic variation in many ways. For example, we expect the amount of variation within a species to increase with the mutation rate (i.e., the higher the mutation rate, the more variation a population will harbor). On the other hand, real populations lose variation every generation due to genetic drift (i.e., sampling error), and the magnitude of this loss is proportional to the population size.

We can write models in form of equations that describe how some of these evolutionary processes impact genetic variation. For example, π (a measure of genetic variation) depends on the size of the population (N) and the mutation rate (μ), such that $\pi \sim 4N\mu$. If we knew the mutation rate, we could rearrange the equation to estimate the size of a population $N = \frac{\pi}{4\mu}$ from genetic data! To arrive at this simple equation, we make the assumption that an individual is equally likely to mate with any other individual in the population (i.e., panmixia). It is easy to see how this may not always hold: for example, a tree is much more likely to mate

with its neighbor than a distant tree across the forest (because pollen cannot travel too far).

Adding biological realism to our equation would prove to be complicated. Alternatively, we can more easily simulate this process with a computer: we can create virtual individuals and tell a computer how they choose their mates, pass on genetic material to their offspring, and change over time. If we were to run virtual experiments like this many times, it would be possible to map the relationship between genetic variation (π) and our parameters of interest (population size N and mutation rate μ). This new map would be closer to the reality than our simplified mathematical model above, but this approach is not a panacea, as simulations have a high computational cost (i.e., they can take a long time to run).

For my dissertation, I have explored how we can incorporate computer simulations to help us understand evolution. To this end, we need computational tools that are fast and reproducible (so that other researchers can use them). So, in my second chapter, I discuss my contributions to tools that make these simulations run faster and more reproducible. Next, I sought to answer a long standing biological question: to what extent have humans and other apes been shaped by natural selection? In my third chapter, I use simulations to tackle this question and show that both positive and negative selection are necessary to explain the genetic data obtained from multiple individuals for all great ape species. Lastly, I wanted to build a proper statistical method that uses simulations to infer evolutionary processes from genetic data. In my fourth chapter, I present a new method for evolutionary inference that uses genealogical trees along chromosomes (instead of tables with samples and mutations). These trees capture the genetic data perfectly,

but they are much more scalable (use less space) and they organize the data in a way that could be more conducive to extracting relevant evolutionary information.

1.2 The current landscape in inference of evolutionary processes from genetic data

Many questions in evolutionary biology are of a historical nature, because evolution is restricted in space and time and it cannot be replicated (Cleland, 2002; Losos, 2009). An avenue for investigating past evolutionary events and processes lies in genetic data. As evolutionary processes leave footprints on the genetic composition of populations, the possibility arises to invert this relationship, that is to use genetic data to make inferences about past events. Indeed, much of evolutionary genetics is tasked with trying to infer past evolutionary events and processes from genetic data. Among the major questions explored in the field are: Is it possible to find regions of the genome that are constrained by natural selection (and thus may be of functional importance)? Can we infer the movement of individuals and changes in population sizes over time? What are the relative contributions of different evolutionary processes, such as demography and selection, in shaping patterns of genetic variation?

1.2.1 The impact of evolutionary processes on genetic variation.

To understand how it is possible to answer these questions with genetic data, it is necessary to map how different evolutionary processes affect genetic variation. At each position along a chromosome, there is a tree that describes how sampled individuals are related to each other. Due to linkage, neighboring trees are more like each other than distant trees. These trees or genealogies can fully encode the outcomes of evolutionary processes. Therefore, it is usually more intuitive to think about how these processes impact tree shapes.

The trees, along with mutations, can then be used to understand genetic variation itself.

Demographic processes directly impact the underlying genealogies because of the inverse relationship between population size and coalescence rate (coalescences are the merging of lineages backwards to a most recent common ancestor) (Wakely, 2016). For example, for a population that goes through a bottleneck (i.e., a sudden decrease in population size), the coalescences are expected to be concentrated during the bottleneck period. On the other hand, for a population that undergoes a size expansion, most of the coalescences will happen in the period preceding the expansion. These underlying trees yield different patterns of genetic variation in the extant population. One summary of genetic data that can be used to distinguish between these tree shapes and, in turn, between possible demographies is the site frequency spectrum (SFS). The SFS is the distribution of allele frequencies across sites in a sampled set of genomes. (For the sake of simplicity, I will focus on the unfolded SFS, where the frequency of the derived alleles have been measured; derived allele refers to the allele that is new to a population). A genealogy that has coalescences concentrated near the present (due to a recent bottleneck) will yield a SFS that is skewed towards the high frequency derived alleles. Conversely, a genealogy with coalescences happening earlier on will yield a SFS with an excess of low frequency derived alleles.

Natural selection can also impact genealogies and genetic variation. When a new (sufficiently) beneficial mutation arises in a population it rapidly increases in frequency, ultimately reaching fixation. This leads to many of the coalescences to occur during the fixation trajectory. If we were to sample the population right after fixation, the SFS would be skewed towards the high frequencies (Y. Kim, 2006),

similar to the effect of a population bottleneck. A strongly deleterious mutation, on the other hand, is rapidly purged from the population. Therefore, the shape of the underlying tree is less affected, though the tree will be shorter (Barton & Etheridge, 2004; S. Williamson & Orive, 2002).

Both beneficial and deleterious mutations lead to loss of genetic variation at the selected site (either due to fixation or purging), but nearby sites are also affected due to linkage. Linked mutations are carried along as a beneficial mutation increases in frequency, in a process called selective sweep (Coop & Ralph, 2012; Kaplan et al., 1989; Maynard Smith & Haigh, 1974). The further away from the focal site, the more likely it is for recombination to happen, allowing mutations to escape this "sweep". A deleterious mutation can also affect linked mutations in a similar way, and this process is called background selection (Charlesworth et al., 1993). Importantly, most of the models for selection assume strong selection, because it is complicated to deal with the interactions between multiple selected mutations in linkage disequilibrium (Y. Kim & Stephan, 2000).

1.2.2 Challenges in evolutionary inference. Although we have a good understanding of how different evolutionary processes impact variation, a few issues remain in the way of inversing this relationship to infer these processes from genetic data. First, different processes can impact genetic variation in similar ways, making it difficult to determine the specific processes that are consistent with data. Second, there is a growing need for modeling interactions between processes. Third, many models do not properly incorporate linkage information and treat whole genome data as unlinked single loci.

Summaries of genetic data may not contain enough information to distinguish between evolutionary scenarios. For example, both positive and negative

selection remove variation surrounding the selected site. Because the width of the effects of sweeps and background selection depends on the recombination rate, both these models predict relationships between (i) recombination rate and genetic diversity and (ii) density of functional sites and genetic diversity (Charlesworth et al., 1993; Kaplan et al., 1989; J. M. Smith & Haigh, 1974). Indeed, in many species of plants and animals it has been observed that regions of high recombination harbor more genetic variation (Corbett-Detig et al., 2015). Disentangling the relative contributions of sweeps and background selection to these patterns, however, is complicated (Andolfatto, 2001; Leffler et al., 2012). One alternative lies in going beyond single species metrics of variation, and to instead compare multiple species. Selection also directly impacts divergence between species: whereas positive selection increases fixation rates, negative selection decreases the rate of substitutions (Andolfatto, 2007; Cai et al., 2009; Macpherson et al., 2007) . It may also be possible to disentangle sweeps and background selection by aggregating multiple different measures of variation (Schrider, 2020).

Individually dissecting and modelling the effects of different processes on genetic variation is not enough. In the early days of evolutionary genetics, it was possible to survey just a few genetic markers, usually microsatellites. These markers are short tandem repeats (1-6bp) that are abundant in the genomes of many taxa. Microsatellites have a high mutation rate and are thought to be mostly neutral (Field & Wills, 1997). This allowed researchers to use these markers for understanding demographic processes almost independently of other processes, such as natural selection. There have been incredible advancements in our ability to obtain whole-genome data, and now it is feasible to sequence thousands of individuals for multiple species at once. More data opens up the

opportunities to investigate more complex questions (i.e., beyond the traditional studies of population structure), such as understanding constraint along genomes, recombination rate variation, among others. However, with whole-genome data the assumptions of neutrality, for example, are much less likely to hold. So there is a need for models that account for complex interactions between processes, which is not trivial to do. One interesting (and perhaps unintuitive) way that processes can interact is that demographic processes can exacerbate the effects of background selection (Torres et al., 2018, 2020). Inferences of background selection that do not take into account demography can lead to biased estimates of the rate of deleterious mutations and their fitness effects. Conversely, both positive and negative selection can bias demographic inferences in many ways (Ewing & Jensen, 2016; Johri et al., 2021; Schrider et al., 2016).

Taking full advantage of linkage information is complicated, and many current methods treat whole-genome data as a collection of unlinked loci. Much of demography inference relies on allele frequency data (i.e., the SFS) ignoring the information contained in patterns of linkage between sites (Gutenkunst et al., 2009; Schraiber & Akey, 2015). It is possible to incorporate linkage information by using the sequentially Markovian coalescent (SMC), a model which approximates the coalescent with recombination by disregarding long-range correlations between genealogies (Harris & Nielsen, 2013; Li & Durbin, 2011; Schraiber & Akey, 2015). Inference of selective sweeps usually proceeds by breaking up the chromosome into non-overlapping genomic windows of a particular size (DeGiorgio et al., 2016; Garud et al., 2015; Pavlidis et al., 2013), and thus it disregards linkage between nearby windows. This issue may be partially mitigated by training a classifier over larger regions, an approach which has been sucessfully implemented (Schrider &

Kern, 2016, 2017). The *a priori* choice of window sizes for calling sweeps remains an important consideration because the signatures of a sweep depend on the strength of a sweep and the local recombination rate (which determine the linked effects of a sweep). Choosing a particular window size can limit the power to detect sweeps of a specific strength (Caldas et al., 2022).

1.2.3 The promise of simulation-based evolutionary inference.

A computer simulation is a virtual model that mimics the behavior of a complex system with known rules. It is widely used in many disciplines to gain insights into systems that are too complex for analytical solutions. The major downsides to using simulations are the runtime and computational costs, but recent advances in computing have lessened these costs.

Simulations have played an integral part in evolutionary inference for the past five decades (Hoban et al., 2012). The main three uses for simulations in evolutionary genetics are: (i) to describe analytically intractable evolutionary models, (ii) to verify analytical solutions based on approximations, and (iii) to perform parameter estimation and/or model selection. For example, Ohta and Kimura (1974) used simulations to compare the distribution of allele frequencies yielded from the simple *infinite alleles model* with the analyitically intractable *stepwise mutation model*, which is more appropriate for allozyme data that was abundant at the time. On the other hand, Kimura (1980) used simulations to verify their analyitical solution of the average time to fixation for new beneficial alleles.

Using evolutionary simulations for model selection and parameter estimation has gained momentum in the recent past. As genomic data availability has increased, so has our ability to interrogate more complex questions from these data. However, as detailed in the earlier sections, writing down mathematical

models and likelihoods is unfeasible in many instances (Coop & Ralph, 2012)e.g.. Simulations can be used to generate data under complex evolutionary scenarios more easily. Then, replicates of these data can be generated under different evolutionary models (or over a parameter range) so that they can be used for inference. Advancements in computational power and algorithms have lessened the burden both in generating data and inferring parameters or doing model selection on them. For example, deep learning algorithms can handle complex inference tasks better than earlier approaches, such as Approximate Bayesian Computation (Sheehan & Song, 2016).

1.2.4 Outline of this dissertation. My overarching goal with this dissertation has been to push the boundaries of simulation-based inference in evolutionary genetics. Although I do not expect simulation-based inference to solve all the challenges in evolutionary inference, simulations hold promise in helping us model interactions between processes and in allowing us to explore the utility of metrics for which there is no theory yet.

An intermediate step to enable robust simulation-based inference is to develop and maintain scalable and reproducible tools for evolutionary simulations. In Chapter II, I present two of my contributions to the `tskit` ecosystem of evolutionary simulation tools. A drawback of simulations is the computational cost associated with them. So I worked out a way to parallelize multi-species simulations (either forward or backward-in-time). This allows for faster and less computationally intensive simulations, which helps enable inference from multi-species genomic datasets. Another challenge with evolutionary simulations has been reproducibility and compatibility (across tools). As part of the Stdpopsim consortium, we have been standardizing evolutionary models and simulations

(Adrion et al., 2020; Lauterbur et al., 2023). Using this new resource, I explore how the power to detect selective sweeps vary along chromosomes in humans. This illustrates how our tools enable the study of interactions between processes (in this case, recombination rate variation, background selection and selective sweeps). We found that the power to detect sweeps varies substantially along chromosomes mostly due to variation in recombination rates. Surprisingly, the variation in power along the chromosome is greater than across humans sampled from different parts of the world (which have drastically different demographic histories).

Simulations can help us understand the behavior of complex evolutionary models. In Chapter III, we use simulations to understand how the landscapes of diversity and divergence change over time under different evolutionary scenarios. By looking at multiple species at once, it is possible to better disentangle the processes that shape variation along genomes. Using a well-sampled set of great ape genomes, we find that landscapes of diversity (and divergence) are highly correlated over long time scales (about $60N$ generations, assuming $N = 10,000$). We tease apart how different processes, such as GC-biased gene conversion, mutation rate variation, positive and negative selection, can contribute to the observed correlations. Although many of these processes are able to produce patterns similar to the observed data, we find that positive selection is necessary to fully explain the data.

Beyond qualitative explorations of complex evolutionary models, simulations can be used for proper statistical inferences, such as model selection and parameter estimation. In Chapter IV, I present a new deep learning method for estimating parameters from genetic data using whole-genome genealogies. Whole-genome genealogies may be useful in evolutionary inference because this data structure

naturally encodes evolutionary processes (e.g., coalescences over time) and they can be more efficient than other representations (such as genotype matrices). We describe a neural network architecture that can be used to infer parameters at different scales using whole-genome genealogies. Next, we test the usefulness of the whole-genome genealogies for inference by estimating mutation times given a known demographic history. We find that our neural network performs well, even outperforming other approaches (that use genealogies or genotype matrices). Our proposed whole-genome genealogy inference framework is able to extract relevant evolutionary information from this data structure and it might be well suited for genome bank scale datasets.

Taken together, these chapters demonstrate how simulations can be employed to help us understand the role of different evolutionary processes in shaping genetic variation. I demonstrate how useful simulations are both in helping us describe intractable evolutionary models and in inferring parameters or performing model selection from genetic data. Simulations can enable us to answer complex questions that have tormented the field of evolutionary genetics for the past decades, and I anticipate that the contributions of simulation-based studies will grow even more as we overcome computational constraints.

CHAPTER II

ROBUST AND EFFICIENT TOOLS FOR EVOLUTIONARY SIMULATIONS

2.1 Introduction

As the availability of genetic data has increased, so has our ability to infer evolutionary processes at finer scales. An essential tool for evolutionary inference has been simulations, and many flexible and efficient evolutionary simulators have been developed over the past decade (Bulmer, 1976; Carvajal-Rodriguez, 2008; Hoban et al., 2012; Hudson, 1983; Hudson et al., 1987; Ohta & Kimura, 1974). Until recently, little had been done to integrate different tools and to ensure compatibility of inferred models across studies.

A unifying thread that has emerged is the tree sequence data structure (Kelleher et al., 2016a). The tree sequence provides a way of concisely representing correlated genealogies along chromosomes. Both msprime, an efficient coalescent simulator, and SLiM, the most widely used forward-in-time simulator, record the genealogical history of all samples in a population in the form of tree sequences (Haller & Messer, 2019; Haller et al., 2019; Kelleher et al., 2018). One of the major advantages of doing so in forward-in-time simulations is that it allows neutral mutations to be omitted. By definition, such mutations do not impact the underlying genealogies, but they pose immense computational demand (due to bookkeeping). Therefore, by omitting the neutral mutations from the forward step, it is possible to decrease computational resources dramatically. After completion of the forward simulation, it is possible to simply overlay the genealogies with the neutral mutations if necessary.

As inference becomes ever more complex, now enabled by powerful simulation tools, the need for easy and error-free dissemination of estimated models

has increased. Many researchers used to re-implement models independently, a process which is error-prone and results in duplication of effort. Indeed, a recent study described errors in the implementation of a demographic model in two published papers, which affected the interpretation of biological signals (Ragsdale et al., 2020).

Stdpopsim is a community-driven open source package developed to minimize these issues (Adrion et al., 2020; Lauterbur et al., 2023). We organized a well-documented library with published simulation models from a range of organisms, which can be accessed with simple Python and command-line interfaces. The library contains demographic models, recombination maps and mutation rates for tens of species. All the models have to pass a quality control method, in which an independent party validates the model. Further, we provide a Python API to easily run simulations using the catalog, with msprime and SLiM as simulators on the backend, decreasing even further the barrier to reproducing simulations.

In this chapter, my contributions to the current ecosystem of evolutionary simulation tools in two vignettes. First, I describe a method I devised to make multi-population simulations faster with parallelization, which is enabled by tree sequences. Second, I report my main contribution to `stdpopsim`: the inclusion of selection models. Using this new feature, I investigate how the power to detect sweeps along more realistic chromosomes. Both of these contributions will appear (in some form) in future publications of the tree sequence toolkit library (`tskit`) and of the `stdpopsim` consortium.

2.2 A way to parallelize multi-population simulations with tree sequences

A major problem in simulation-based inference is the computational cost of simulations. In many cases, it is possible to minimize the time cost of simulations by running them in parallel, that is to divide up the simulation over many processes that can be executed concurrently over many CPUs (or cores). It is not always clear how to divide a simulation into sub-tasks, however.

In the case of evolutionary simulations, a natural way to break up a big simulation might be population splits. After a population splits into two (or more) subpopulations, their histories become independent (assuming no migration). Therefore, it is possible to parallelize a multi-population simulation by dividing the simulation over population splits. The history of populations A, B and C (shown in Figure 2.1) can be parallelized: any two branches stemming from the same node are independent.

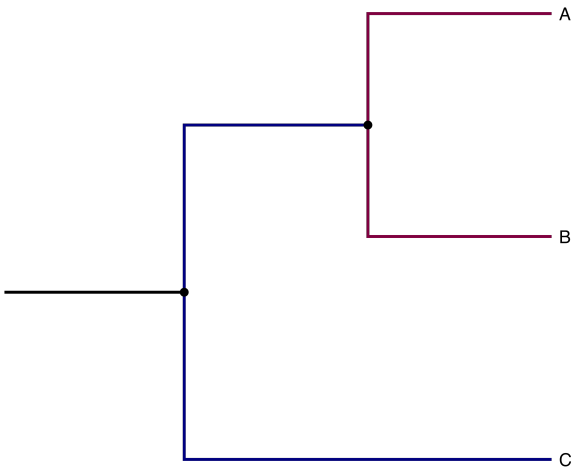


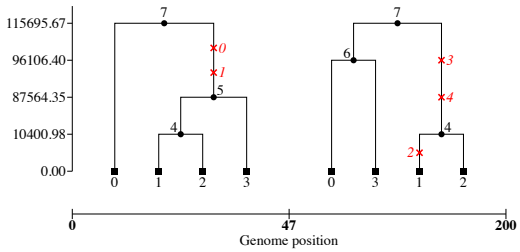
Figure 2.1. The relationships between populations A, B and C are depicted. Note how branches of the same color can be simulated in parallel if there is no migration.

To parallelize the simulations, we need to (i) simulate each branch of the tree independently and (ii) put together the bits that are simulated independently

at the end. There are many ways to parallelize the simulations, but see Rodrigues and Ralph (2021) for an idea. Below, I will describe the operation for putting together two (or more) tree sequences, which I implemented in `tskit` (see https://tskit.dev/tskit/docs/stable/python-api.html#tskit.TreeSequence.union for the documentation). A description of this operation will appear in an upcoming publication describing the tree sequence toolkit library (`tskit`).

The information underlying a tree sequence is stored as a collection of tables which define different features. This simple tabular format allows for rapid accessing of information. The main components are the Node, Edge, Site and Mutation tables. A haploid genome can be thought of as a node in a tree, which exists at a particular time. An edge defines genetic inheritance between two nodes, and it consists of a parent node, a child node, and the left and right coordinates (along a chromosome) over which the child genome inherited from the parent genome. A site is a location in the genome, and a mutation defines a change of state at a particular node. See Figure 2.2a for an example tree sequence and Figure 2.2b for the corresponding collection of tables. Note how information about the trees are completely independent from the notion of genetic variation (defined by the Site and Mutation tables).

After simulating two independent populations, one might want to obtain the node-wise union of these two tree sequences (e.g., for analyzing patterns of between population variation). This is almost as simple as concatenating the Node, Edge, Site and Mutation tables. However, nodes from the second tree sequence need to be re-enumerated, and this new numeric order needs to be propagated onto the other tables. I simulated the history of two populations independently using `msprime` (Baumdicker et al., 2022; Kelleher et al., 2016a) (Figures 2.3a and 2.3b). Then, I



(a) An example tree sequence.

Nodes

population	individual	time
0	0	0.00000000
0	1	0.00000000
0	2	0.00000000
0	3	0.00000000
0	-1	10400.97834031
0	-1	87564.35114390
0	-1	96106.40092689
0	-1	115695.67326147

Edges

left	right	parent	child
0	200	4	1
0	200	4	2
0	47	5	3
0	47	5	4
47	200	6	0
47	200	6	3
0	47	7	0
47	200	7	4
0	47	7	5
47	200	7	6

Mutations

site	node	time	derived_state
0	5	95764.69873733	C
1	5	98781.54806245	T
2	1	1359.48373362	G
3	4	81679.72997120	T
4	4	39253.99752936	G

Sites

position	ancestral_state
1	T
21	G
63	T
130	G
149	T

(b) The underlying collection of tables.

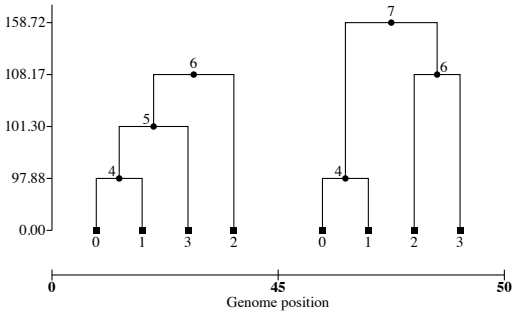
Figure 2.2. The relationship between four samples are depicted in a sequence of trees. Note how because of shared ancestry, mutations that are common to many of the samples need to be represented only once. However, in a matrix of variant sites against samples, these mutations would be repeated unnecessarily.

used the union operation to merge the two tree sequences (Figure 2.3c). Lastly, I added the pre-split history to the merged tree sequence so that all samples coalesce back-in-time (Figure 2.3d). See the code in Python to reproduce this analysis (Listing 2.1). By doing the simulation this way, we are able to run the simulation of each extant population in parallel. In the case of large simulations (as we will see in Section 2.4), this is essential so that we can distribute the memory usage over different computer nodes.

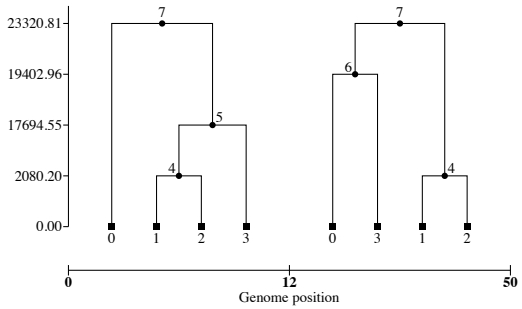
```
import msprime
import numpy as np
import tskit

# First, we build an msprime.Demography object with our two-population
    history that splits 100,000 units of time ago
```

```
demography = msprime.Demography()
demography.add_population(name="A", initial_size=10_000)
demography.add_population(name="B", initial_size=100)
demography.add_population(name="C", initial_size=100_000)
demography.add_population_split(time=100_000, derived=["A", "B"],
    ancestral="C")

# Now, we can simulate the histories of the two extant populations
    independently
ts1 = msprime.sim_ancestry(samples={"A":4}, ploidy=1, demography=
    demography, sequence_length=50, recombination_rate=1e-6, random_seed
    =123)
ts2 = msprime.sim_ancestry(samples={"B":4}, ploidy=1, demography=
    demography, sequence_length=50, recombination_rate=1e-6, random_seed
    =1)

# Then, we can union these two tree sequences.
# Note that because the two populations do not share any history,
# we specify a 'node_mapping' in which none of the nodes in 'ts2' have
    an equivalent in 'ts1'.
tsu = ts1.union(ts2, node_mapping=np.full(ts2.num_nodes, tskit.NULL),
    add_populations=False)

# Finally, we can add the pre-split history to the union'ed tree
    sequence with msprime.sim_ancestry.
tsu_recap = msprime.sim_ancestry(initial_state = tsu, demography=
    demography, random_seed=3)
```

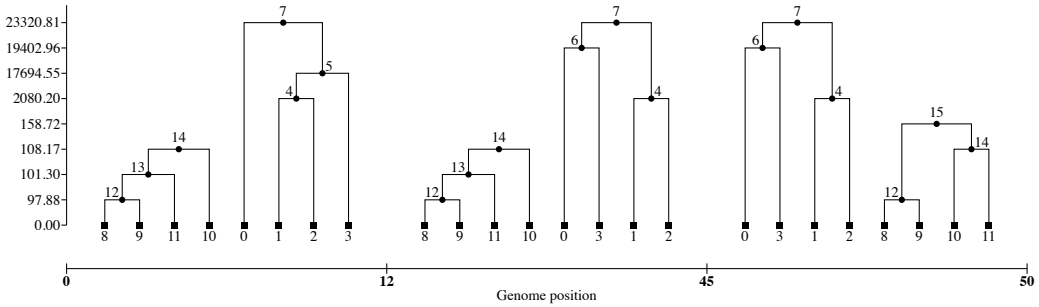
Listing 2.1 Python code to union two independently simulated populations using `msprime` and `tskit`.



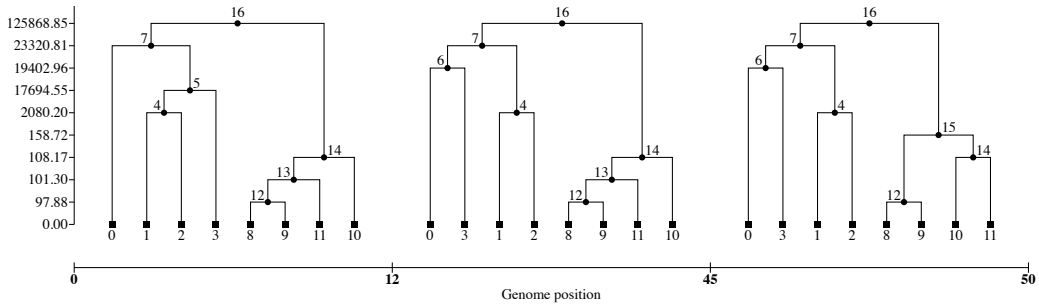
(a) Tree sequence of population 1.



(b) Tree sequence of population 2.



(c) Union of tree sequences 1 and 2.



(d) Adding the pre-split history to the union.

Figure 2.3. The union operation for tree sequences. (a) and (b) show the tree sequences for two independently simulated populations. (c) depicts the union of these tree sequences. (d) shows the merged tree sequences with added history for the ancestral population.

A trickier use case for the node-wise union operation is when there are two tree sequences which share part of their histories. That is, imagine you simulate the ancestor population of A and B, and then start two new independent simulations of A and B. In this case, it is necessary to defined which portions are not shared between the two tree sequences, so that the histories can be joined properly. To define the non-shared portions, it is necessary to specify which nodes in one tree sequence are equivalent to those in the other. Then, the portions that are not shared from one tree sequence are copied over onto the other. New nodes are given a new numeric identifier which are propagated to the other tables. Refer to Rodrigues and Ralph (2021) for a concrete example on how to union tree sequences with some shared history.

2.3 Towards more realistic and reproducible simulations: Introducing models of selection to Stdpopsim

Population genetics can aid in identifying the genetic basis of adaptation. As a new beneficial mutation increases in frequency due to positive selection, it wipes out genetic variation surrounding the selected site (Kaplan et al., 1989; J. M. Smith & Haigh, 1974). This genetic footprint can be used to infer past selection events in natural populations (Enard et al., 2014; Garud et al., 2015; Hernandez et al., 2011; Nielsen, 2000; Przeworski, 2002; Schrider & Kern, 2017; R. J. Williamson et al., 2014). Our ability to detect selective sweeps depends on many parameters, such as the recombination rate, the time since fixation, and the demographic history. The recombination rate determines the width of a selective signal (together with the strength of selection) (Kaplan et al., 1989). As time passes by, new mutations are accumulated restoring pre-sweep levels of genetic variation and erasing sweep signatures. Demographic events can leave footprints similar to a selective sweep,

confounding detection (Jensen et al., 2005; Przeworski, 2002). Conversely, sweep signatures can be erased by recent demographic events, such as bottlenecks.

Another process that can impact our ability to detect sweeps is background selection. Background selection is the process whereby neutral genetic variation linked to deleterious mutations are lost (Charlesworth et al., 1993), and the extent of this effect is also modulated by strength of selection and recombination rate. Because of this relationship with recombination rate, background selection can confound the search for selective sweeps (Andolfatto, 2001). This process seems to be pervasive in multiple species; so much so that background selection may be considered a better null hypothesis in population genomic studies rather than neutrality (Comeron, 2017).

Background selection may be discernible from selective sweeps, specially when considering more realistic models (Schrider, 2020). Previous studies that found background selection can confound sweep calling assumed that the region constrained by selection is flanked by large streches of neutral sites, but the locations of exons and other constrained elements are usually scattered throughout the chromosome. In such cases, sweeps can be separated from constraint more easily.

To better understand the role of positive selection in shaping genetic variation across the genome, it is necessary to better delineate how different processes impact our ability to detect selective sweeps. I present below our efforts to include models of natural selection to `stdpopsim`, our community-driven library of evolutionary models. Colleagues and I included the ability to easily simulate background selection and selective sweeps using previously published distribution of fitness effects (DFEs). I then leveraged this new feature to understand how

power to detect selective sweeps varies over a realistic looking human chromosome, with recombination rate variation and a map of constraint taken from human annotations. These results will appear in an upcoming publication of the `stdpopsim` consortium.

2.3.1 Adding selection to simulations using Stdpopsim. There are two main ways of adding selection to a simulation in stdpopsim: (i) it is possible to specify a distribution of fitness effects (DFE) for new mutations that can occur across the genome or over some pre-specified regions; and (ii) we can introduce and track a single mutation to a population, as one would need to study selective sweeps. Beyond the machinery to actually perform these two kinds of simulations, we now added to Stdpopsim a library of previously published distributions of fitness effects and genomic annotations (e.g., exon and intron coordinates) which can be used to build realistic models with selection. See our Tutorial for more details on how to implement models with selection in stdpopsim: https://popsim-consortium.github.io/stdpopsim-docs/stable/tutorial.html#incorporating-selection.

2.3.2 Analysis of power to detect sweeps along realistic chromosomes. To demonstrate the utility of stdpopsim, we produced maps of power to detect sweeps along a realistic chromosome. We ran simulations of human chromosome 1 using the three population out-of-Africa model (identified as OutOfAfrica_3G09 in the stdpopsim library; Gutenkunst et al., 2009) and the genetic map estimated in Frazer et al. (2007) (HapMapII_GRCh38). We had three classes of simulations: (i) neutral (with the specified demography and genetic map), (ii) background selection, in which we applied the DFE estimated in B. Y. Kim et al. (2017) (Gamma_K17) to exon annotations (taken from Ensembl;

ensembl_havana_104_exons), and (iii) sweep with background selection, in which we simulated the hard sweep as described with the addition of constraint in exonic regions (same as for *ii*). For computational efficiency, we simulated a 5Mb region flanking 100 evenly distributed points along the chromosome 1.

When simulating truncated regions with selection, it is necessary to consider a boundary effect effect: a new deleterious mutation will be linked to fewer neutral mutations if it arises near the ends of a chromosome, so linked selection is not as efficient. Thus, we first established this effect in simulations in which new deleterious mutations can happen anywhere in the chromosome with the same DFE as in our main simulations, but with constant recombination rate. We found that by adding a buffer surrounding our focal simulated regions of 2.5cM it is possible to mitigate the edge effect (Figure 2.4). We rescaled the simulations to reduce computational cost by simulating smaller populations (factor of 2) while increasing times, mutation and recombination rates, and selection coefficients by the same amount (see Uricchio and Hernandez, 2014 and Adrion et al., 2020 for more details). In total, we produced 200 replicates for each class at each poisition, totalling 60,000 simulations.

We computed three statistics used for finding selective sweeps along the simulated regions: (i) nucleotide diversity, (ii) the probability of a sweep using a machine learning sweep classifier called diploSHIC (Kern & Schrider, 2018; Schrider & Kern, 2016), and (iii) the composite likelihood ratio (CLR) (Nielsen et al., 2005). Nucleotide diversity was computed directly from the underlying tree sequences using `tskit` (Ralph et al., 2020) over 10 equally sized and non-overlapping subwindows. diploSHIC was applied over 15 partially overlapping

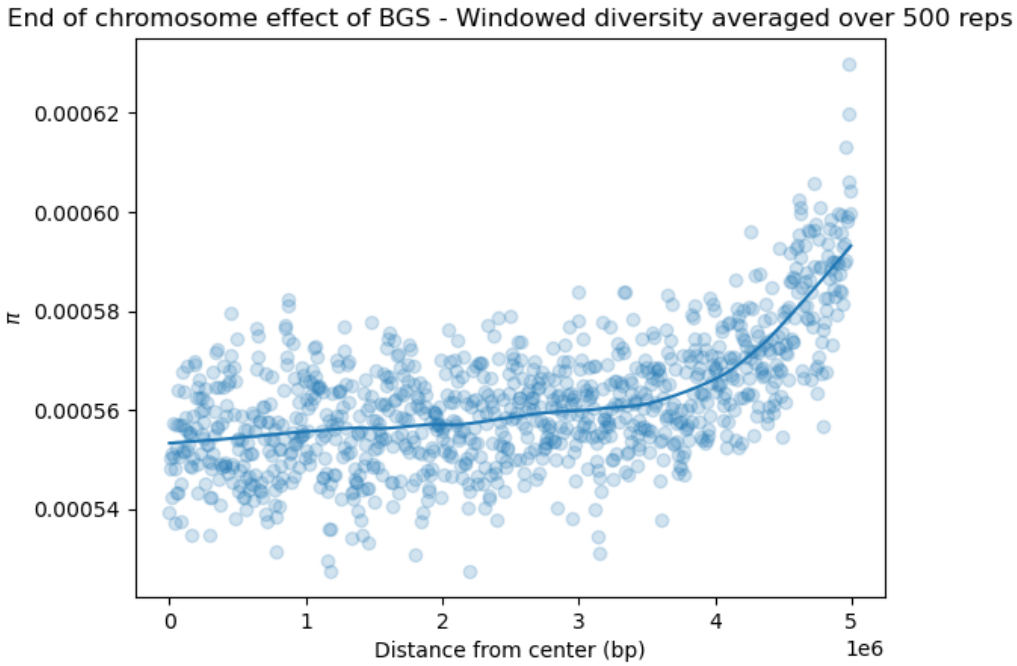


Figure 2.4. Linked selection is not as efficient near the ends of chromosomes. Note how the boundary effect plateaus around 2Mb from the ends (assuming a constant recombination rate).

subwindows. CLR was computed at 21 uniformly distributed points along the simulated region.

We then classified the central subwindow as a sweep or not sweep. To do so, we empirically determined the distribution of each statistic under a null model and defined the top 5% quantile as a cutoff to call a sweep. We considered two null models: neutral and background selection (as explained before). We computed the power to detect a sweep at each one of the 100 locations along chromosome 1, which is the proportion of simulations with a sweep that were correctly classified as a sweep.

There is significant amount of variation in the power to detect sweeps along human chromosome 1 (Figure 2.5). Sweeps are easier to detect in YRI than in CEU, most likely because of the higher effective population size of YRI. Variation in power varies across statistics, with nucleotide diversity yielding the greatest variation in power. The power varies tremendously when using nucleotide diversity to call sweeps, ranging from 0% to 100%. diploSHIC has good power overall, reaching close to 100% at many locations with YRI demography, but it falters at around 15% of the locations where it dips below 75%. CLR, on the other hand, maintains a good average power of around 80% and it does not vary as much.

Using background selection as a null model decreases power for both CLR and Diversity. The effects of background selection on these two statistics are similar to a sweep, so using this null model leads to more stringent cutoffs to call sweeps. diploSHIC behaves differently: using background selection as a null model increases power. This indicates that background selection does not confound sweep detection using diploSHIC, as has been hypothesized before (Schrider, 2020). It

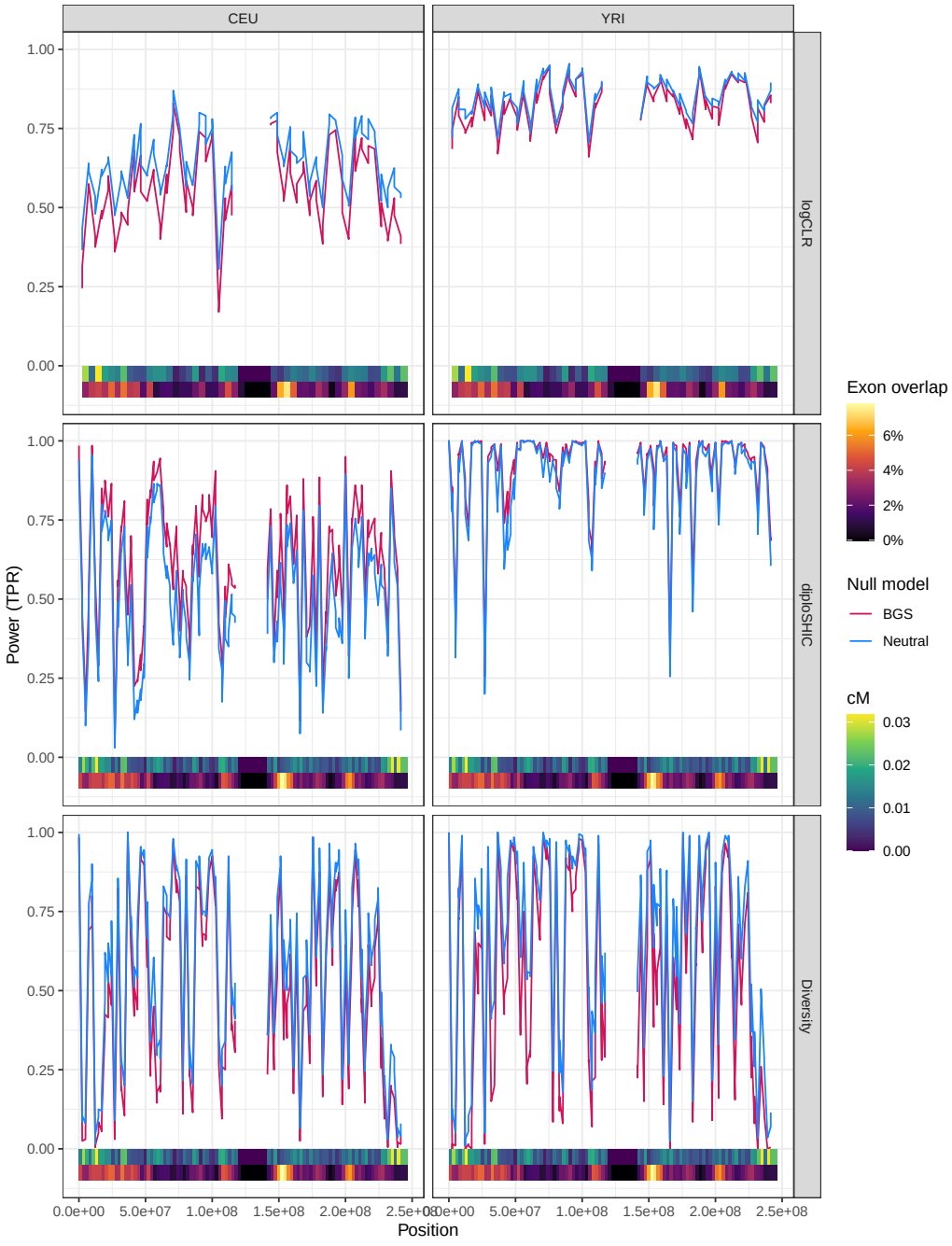


Figure 2.5. There is variation in the power to detect sweeps along human chromosome 1 for all three statistics considered (CLR, diploSHIC probability of a sweep, and nucleotide diversity). Note that power is computed by assuming either a neutral (blue) or background selection (red) null model.

is encouraging that a composite statistic such as diploSHIC can discern between modes of selection, which has been a major problem in population genetics.

To gain a better understanding at which genomic features affect variation in power, we plotted power against recombination rates and percentage of exon overlap. We found that sweeps placed regions of higher recombination are generally harder to detect (Figure 2.6). diploSHIC is not as affected by recombination rates, probably because it can distinguish between locations directly affected by sweeps and locations linked to sweeps.

This decrease in power with high recombination rates is caused by two things: (i) the effect of recombination rate on the variance of statistics under the null, and (ii) the effect of recombination rate on the mean of the statistics under a sweep. With lower recombination rates, we expect the variance of statistics to increase (due to coalescent noise). Indeed, both nucleotide diversity and diploSHIC probability of a sweep vary more in lower recombination regions (Figure 2.7). CLR is not as affected, and this is probably why power does not vary as much along the chromosome. In general, we expect regions of lower recombination to be more affected by a sweep, because it is harder for linked variation to escape a sweep. This is true for both nucleotide diversity and CLR, and we see that the median value for these statistics decrease with higher recombination rates (Figure 2.8). However, the probability of sweep computed by diploSHIC increases with higher recombination rates in our sweep simulations. Because diploSHIC models the difference between regions directly affected by a sweep and regions linked to a sweep, it is able to distinguish between classes better in regions of higher recombination rate.

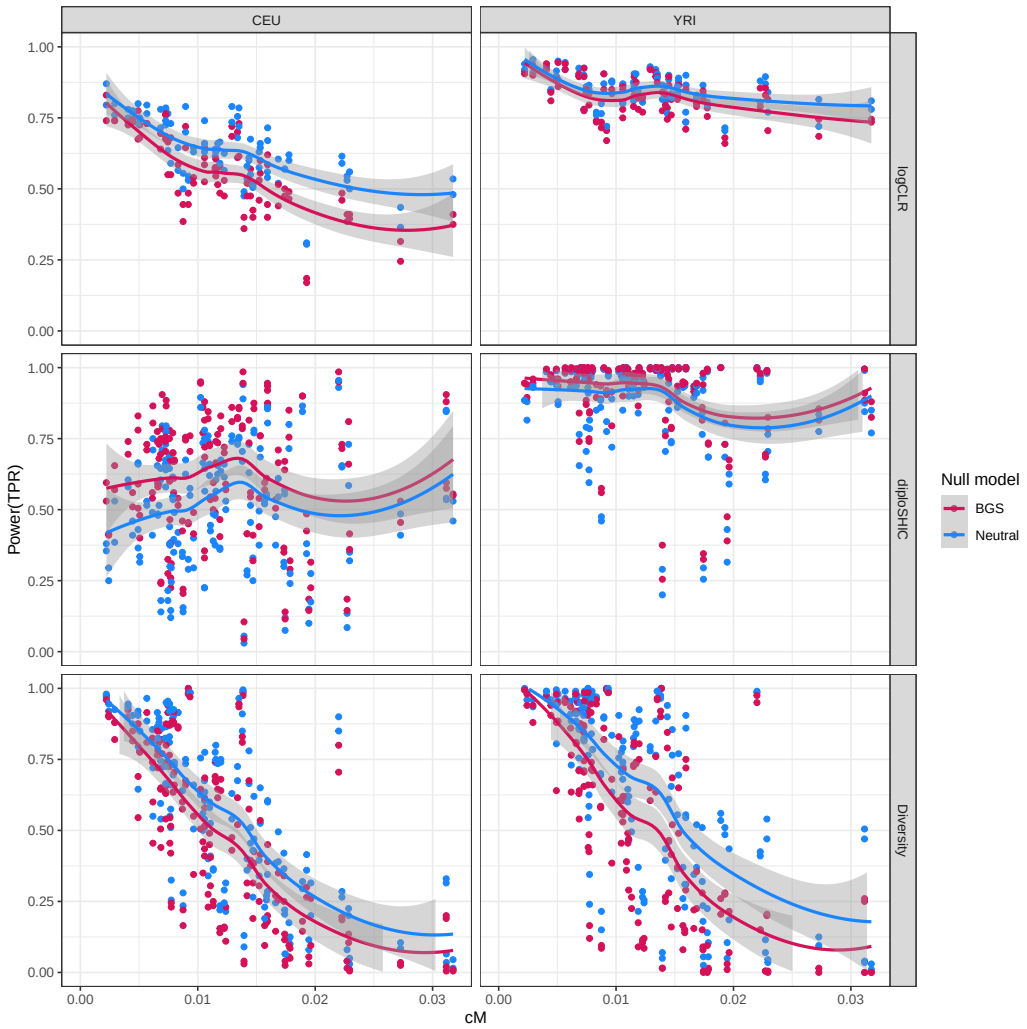


Figure 2.6. Recombination rates affect power to detect sweeps. The scatterplots depict the relationship between the genetic recombination rate and power for all three statistics (CLR, diploSHIC and Diversity). Note that power is computed by assuming either a neutral (blue) or background selection (red) null model.

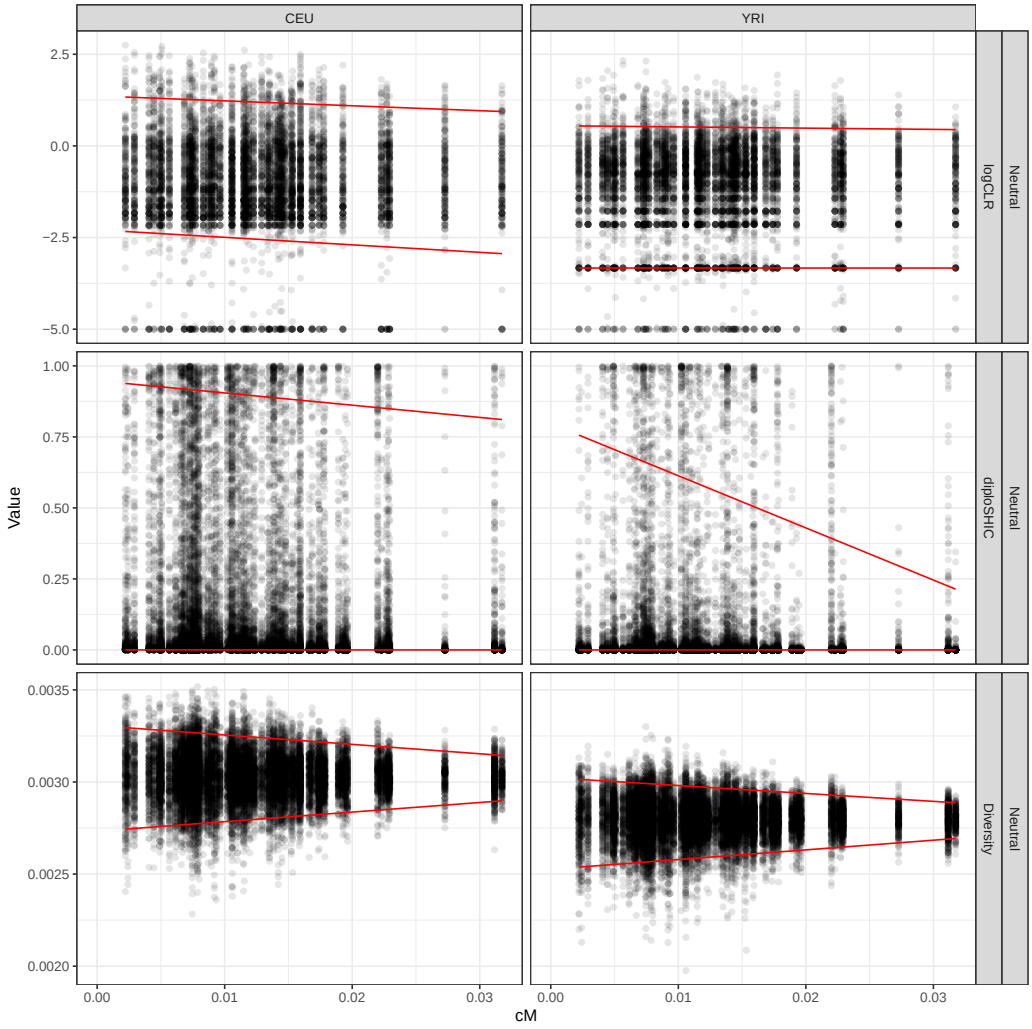


Figure 2.7. Recombination rates affect the variance of statistics under neutrality. Red lines delimit the central 90% of the distribution for the three statistics at each recombination rate.

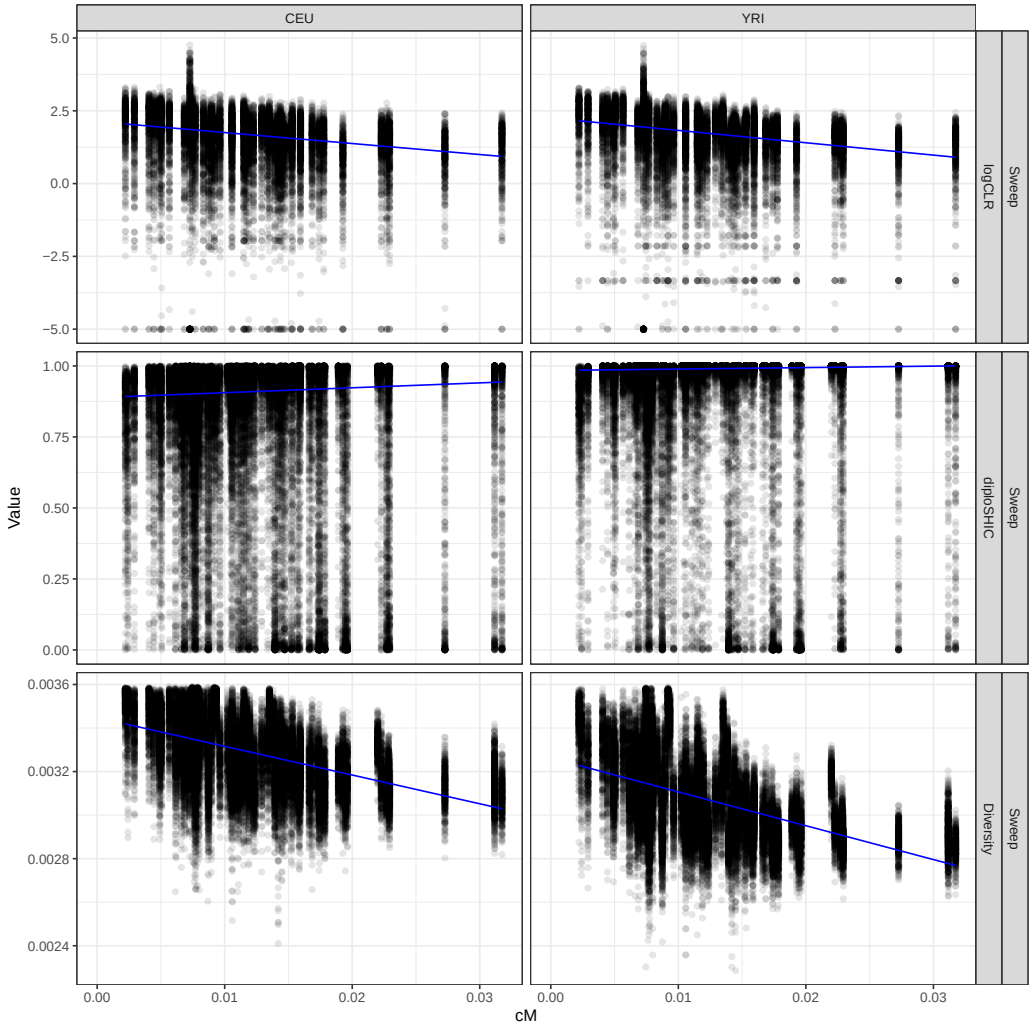


Figure 2.8. Recombination rates affect the median of statistics under the sweep model. The blue line is the median of the distribution for the three statistics at each recombination rate.

Taken together, our results demonstrate that power to detect sweeps varies considerably along realistic looking chromosomes. Indeed, power along chromosomes may vary almost as much as they do with different demographic models. It is necessary for future studies to consider these effects more carefully, and we provide the tools to do so using the `stdpopsim` library and software package. It is possible that previous scans have only found sweeps where they can be found. Thus, the effects of positive selection may have been understated in the literature, and important sweeps might have not been correctly identified. It seems important to better investigate whether ascertainment biases in reported sweeps correlate with power along chromosomes, and how this varies with different methods and across species.

2.4 Bridge

In Chapter II, I presented my contributions to make evolutionary tools more robust and efficient. To gain insights into the complex interactions between evolutionary processes that shape genetic variation, it is necessary to increase the biological realism of our models. This increase in realism comes at a cost: simulations that are closer to reality take longer to run and consume more computational resources. In Chapter III, we apply our ideas presented in this chapter on how to parallelize multi-population simulations and on simulating realistic chromosomes. The realistic simulations we produce in Chapter III were extremely costly, despite our best efforts to optimize, and each replicate took over 30 days to run consuming at the peak over 160Gb of RAM. Nevertheless, we were able to push the boundaries in what is possible with our tools and computational resources today to answer a long-standing question in evolutionary genetics: What are the relative contributions of different evolutionary processes in shaping genetic variation in the great apes?

CHAPTER III

SHARED EVOLUTIONARY PROCESSES SHAPE LANDSCAPES OF GENOMIC VARIATION IN THE GREAT APES

This chapter was published in the journal *Genetics* in 2024. Andrew D. Kern and Peter L. Ralph are co-authors on this paper. Co-authors and I conceptualized the study, I curated the dataset, I performed analyses and simulations with input from my co-authors, and I wrote the manuscript with editorial assistance from my co-authors.

The citation for this publication is as follows:

Rodrigues, M. F., Kern, A. D., & Ralph, P. L. (2024). Shared evolutionary processes shape landscapes of genomic variation in the great apes. *Genetics*, iyae006. https://doi.org/10.1093/genetics/iyae006

3.1 Introduction

Genetic variation is determined by the combined action of mutation, demographic processes, recombination and natural selection. However, there is still no consensus on the relative contributions of these processes and their interactions in shaping patterns of genetic variation. Two major open questions are: how does the influence of selection compare to other processes? And, to what degree is genetic variation influenced by beneficial versus deleterious mutations?

Genetic variation can be measured within a species or between species with two related metrics: within-species genetic diversity and between-species genetic divergence. Both can be estimated with genetic data by computing the per site average number of differences between pairs of samples within a species or between two species, and these are estimates of the mean time to coalescence.

(Note that we do not discuss *relative divergence*, which is often measured using F_{ST}.) Evolutionary processes impact diversity and divergence in different ways, so the relationship between these carries information regarding these processes.

Natural selection directly impacts genetic diversity because it can reduce the frequencies of alleles that are deleterious (negative selection) or increase those of beneficial alleles (positive selection). Selection can also directly affect between-species genetic divergence. Deleterious alleles are more likely to be lost from the population, thus reducing divergence at the affected sites. On the other hand, beneficial alleles have a higher probability of fixation. This leads to an increase in the rate of substitution at sites under positive selection that in turn increases divergence between species. Thus, contrasting patterns of diversity and divergence at the same time can help disentangle between modes of selection (Hudson et al., 1987). Indeed, perhaps the most widely used test for detecting adaptive evolution, the McDonald-Kreitman test, compares diversity and divergence contrasted between neutral (e.g., synonymous) and functional (e.g., non-synonymous) site classes (McDonald & Kreitman, 1991). This test and its extensions have been applied to a myriad of taxa, and it has become clear that a substantial proportion of amino acid substitutions are driven by positive selection in a number of taxa (Galtier, 2016; Ingvarsson, 2010; Slotte, 2014; N. Smith & Eyre-Walker, 2002).

Selection also disturbs genetic variation at nearby locations on the genome, and this indirect effect of selection on diversity is called "linked selection". Linked selection can be caused by at least two familiar mechanisms: genetic hitchhiking and background selection. Under genetic hitchhiking, as a beneficial mutation quickly increases in frequency in a population, its nearby genetic background is carried along, causing local reductions in levels of genetic diversity. The size of the

region affected by the sweep depends on the strength of selection, which determines how fast fixation happens, and the crossover rate, because recombination allows linked sites to escape from the haplotype carrying the beneficial mutation (Kaplan et al., 1989; Maynard Smith & Haigh, 1974). Under background selection, neutral variation linked to deleterious mutations is removed from the population unless, as before, focal lineages escape via recombination (Charlesworth et al., 1993). Both of these processes leave similar footprints on patterns of within-species genetic diversity, and so attempts to determine the contributions of positive and negative selection in shaping levels of genetic variation genome-wide have proven to be difficult (Andolfatto, 2001; Y. Kim & Stephan, 2000), although the processes seem separable more locally (Schrider, 2020; Schrider & Kern, 2017). Importantly, linked selection does not affect between-species genetic divergence as strongly, as a beneficial or deleterious mutation does not affect the substitution rate of linked, neutral mutations (Birky & Walsh, 1988) (although it does affect divergence through ancestral levels of polymorphism (Begun et al., 2007; Phung et al., 2016)).

The effects of linked selection in shaping genetic variation are pervasive across genomes (Begun & Aquadro, 1992; Cai et al., 2009; Corbett-Detig et al., 2015; Lohmueller et al., 2011; Murphy et al., 2022). For example, dips in nucleotide diversity surrounding functional substitutions have been uncovered in many taxa, such as fruit flies (Kern et al., 2002; Sattath et al., 2011), rodents (Halligan et al., 2013), *Capsella* (R. J. Williamson et al., 2014) and maize (Beissinger et al., 2016). In *Drosophila melanogaster*, levels of synonymous diversity (which is putatively neutral) and amino acid divergence are negatively correlated (Andolfatto, 2007; Macpherson et al., 2007); positive selection can cause such a pattern if beneficial amino acid mutations are fixing and as they do reducing levels of linked neutral

variation via selective sweeps. In contrast in humans, levels of synonymous diversity are roughly the same near amino acid substitutions and synonymous substitutions, suggesting recent, recent fixations at amino acids sites may not be the result of strongly beneficial alleles (Hernandez et al., 2011; Lohmueller et al., 2011). However, in the human genome, amino acid substitutions tend to be located in regions of lower constraint than synonymous substitutions, implying that the signal of positive selection may be confounded by the effects of background selection (Enard et al., 2014).

Two major challenges remain in the way of a fuller characterization of the effects of selection on genetic variation: (i) it is hard to model interactions between evolutionary processes (e.g., sweeps within highly constrained regions), and (ii) model identifiability is challenging for some summaries of the data (e.g., sweeps and background selection may impact diversity in similar ways). Recent computational advances have made it possible for us to move from simpler backwards-in-time coalescent models (Hudson, 1983) to more complex and computationally demanding forward-in-time simulations, and these have provided a route to studying these hard to model interactions between evolutionary processes across multiple sites (Haller & Messer, 2019; Haller et al., 2019; Kelleher et al., 2016a). Simulation-based inference can then allow us to better describe the roles of different modes of selection and other processes in shaping genomic variation. However, the problem of identifying features of the data that are informative of the strength and mode of selection still remains.

One promising approach might be to compare patterns of genetic variation in multiple species jointly as each species can be thought of as semi-independent realizations of the same evolutionary processes (c.f. Won & Hey, 2005). In

speciation genomics studies, it is common to visualize large scale patterns of genetic variation along chromosomes (so-called landscapes of diversity and divergence), which may contain substantial information to help us disentangle evolutionary processes. Earlier empirical surveys have focused on the identification of regions of accentuated relative divergence between populations (Cruickshank & Hahn, 2014; Harr, 2006; Turner et al., 2005), although patches of increased divergence can be the result of myriad forces besides reproductive isolation and adaptation. Recent comparative studies have found that landscapes of diversity are highly correlated between related groups of species, such as *Ficedula* flycatchers (Burri et al., 2015; Ellegren et al., 2012), warblers (Irwin et al., 2016), stonechats (van Doren et al., 2017), hummingbirds (Battey, 2020), monkeyflowers (Stankowski et al., 2019) and *Populus* (Wang et al., 2020). Burri (2017) proposed that we could capitalize on correlated genomic landscapes to study the interplay between different forms of selection and other evolutionary processes. Neutral processes, such as retained ancestral diversity (i.e., incomplete lineage sorting) or migration, could potentially produce significant correlations in levels of diversity across species, however strong correlations have been observed among taxa with long divergence times and without evidence of gene flow. For example, Stankowski et al. (2019) found that landscapes of diversity and divergence are highly correlated across a radiation of monkeyflowers which spans one million year (or about $10N_e$ generations, where N_e is the effective population size), far longer than the time scale on which we expect to see effects of ancestral variation and incomplete lineage sorting (since the coalescent timescale spans just a few multiples of N_e). However, a shared process that independently occurs in the branches of a group of species could maintain correlations over long timescales. For example, if two species' physical arrangement

of functional elements and local recombination rates are similar, the direct and indirect effects of selection could make it so that peaks and valleys on the landscape of diversity are similar, maintaining correlation between their landscapes over evolutionary time (Burri, 2017; Delmore et al., 2018). Further, if mutational processes are heterogeneous across the genome in a manner that is shared among species, then correlated landscapes of diversity could be created through mutational variation as well.

Here, we aim to (i) describe whether and in what ways landscapes of within-species diversity and between-species divergence are correlated, and (ii) to tease apart the relative roles of positive and negative selection and other processes (e.g., ancestral variation, mutation rate variation) in shaping patterns of genetic variation. To understand processes driving these correlations, we employ highly realistic, chromosome-scale, forward-in-time simulations, since analytical predictions are not available. We use the great apes as a system to investigate correlated patterns of genetic variation because there is high quality population genomic data for all species (Prado-Martinez et al., 2013), the clade is about 12 million years old or $60N_e$ generations (but there have not been many chromosomal arrangements Jauch et al., 1992), and lastly the landscapes of gene density, recombination rate and mutation rate are roughly conserved (Kronenberg et al., 2018; Stevison et al., 2016). Our study demonstrates that correlated landscapes can be useful in distinguishing between modes of selection and the balance of direct and linked selection shaping genomic variation.

3.2 Methods

3.2.1 Genomic data. We retrieved SNP calls for ten great ape populations made on high coverage ($\sim$25$\times$) short-read sequencing data from the

Great Ape Genome Project (Prado-Martinez et al., 2013), mapped onto the human reference genome (NCBI36/hg18). We analyzed 86 individuals divided into the following populations: human ($n = 9$ samples), bonobo ($n = 13$), Nigeria-Cameroon chimpanzee ($n = 10$), eastern chimpanzee ($n = 6$), central chimpanzee ($n = 4$), western chimpanzee ($n = 4$), eastern lowland gorilla ($n = 3$), western gorilla ($n = 27$), Sumatran orangutan ($n = 5$), Bornean orangutan ($n = 5$) (we excluded two samples from the original dataset: the Cross River gorilla and the chimpanzee hybrid). Prado-Martinez et al. (2013) applied several quality filters to the SNP calls (see Section 2.1 of their Supplementary Information) and, for each species, identified the genomic regions in which it would be unreliable to call SNPs (uncallable regions). For our downstream analyses, we only considered sites which were callable in all populations.

We calculated nucleotide diversity and divergence (d_{XY}) in non-overlapping 1Mb windows using `scikit-allel` (Miles et al., 2020). Windows in which there were less than 40% callable sites were not used in any of the analyses. For example, this yielded 129 (out of 132) 1Mb windows in chromosome 12 in which 75% of the sites were callable on average.

To tease apart the effects of GC-biased gene conversion (gBGC), we decomposed diversity and divergence by allelic states. gBGC is expected to affect weak bases (A or T) which are disfavored when in heterozygotes which also carry a strong base (G or C). Thus, one way understand the effects of gBGC is by comparing sites which were weak to those that were strong in the ancestor (ancestrally strong alleles are not affected by gBGC, but ancestrally weak alleles can be). We assumed that the state in the ancestor of the great apes to be the state seen in rhesus macaques (genome version RheMac2) — sites without enough

information in RheMac2 were excluded. Then, we computed divergence only considering sites which were ancestrally weak or ancestrally strong (Figure A.4). This approach has two major drawbacks: (i) many of the sites cannot be used because they are missing in RheMac2 and (ii) sites can be mispolarized. Thus, we came up with a second approach to tease apart the effects of gBGC on correlations between genomic landscapes. When comparing two landscapes of divergence (which encompass four species), we can classify each site by the change in state that happened without needing to polarize mutations by looking at the ancestor. For example, if we have allelic states for four species and we see A-A-T-T as the configuration of alleles at a particular site, we know that there must have been one mutation which changed the state from a weak base to another weak base (W-W). On the other hand, if we see A-G-A-A there must have been one mutation from weak to strong (W-S) (or vice-versa). Sites with multiple mutations (e.g., A-G-G-C) were removed from the analyses. Sites that did not change from W to S (or vice-versa) are not expected to be affected by gBGC, and we refer to these as W-W or S-S mutations (Figure 3.6A). Sites where there may have been a weak to strong change (W-S mutations) may be affected by gBGC (Figure 3.6B). We only considered windows with at least 5% of callable sites in these analyses.

3.2.2 Simulations. We implemented forward-in-time Wright-Fisher simulations of the entire evolutionary history of the great apes using `SLiM` (Haller & Messer, 2019; Haller et al., 2019). Each branch in the great apes' tree was simulated as a single population with constant size (Figure 3.1). Population splits occurred in a single generation, and there was no contact between populations post-split. Population sizes and split times were taken from the estimates in Prado-Martinez et al. (2013). Across all our simulations, we simulated crossover events

occurred with the sex-averaged rates from the deCODE genetic map (in assembly NCBI36/hg18 coordinates) (Kong et al., 2002). We then computed diversity and divergence in the same windows used for the real data using `tskit` (Kelleher et al., 2018; Ralph et al., 2020).

To improve run time, we simulated sister branches in parallel and recorded the final genealogies as tree sequences (Kelleher et al., 2016a). Further, neutral mutations were not simulated with `SLiM` and were added after the fact with `msprime`. The resulting tree sequences were later joined and recapitated (i.e., we simulated genetic variation in the ancestor of all great apes using the coalescent) using `msprime`, `tskit` and `pyslim` (Kelleher et al., 2016a, 2018; Rodrigues & Ralph, 2021). Despite our efforts to improve run time, our simulations of the entire history of the great apes were still incredibly costly (taking over a month to complete in many instances).

In our neutral simulations, we assumed that neutral mutations occurred at a rate of 2×10^{-8} new mutations per generation per site (Scally & Durbin, 2012), uniformly across the entire chromosome. To understand the effects of natural selection on landscapes, we simulated beneficial and deleterious mutations only within exons, assuming that the locations of exons were shared across all great apes (Kronenberg et al., 2018) and using exon annotations from the human reference genome NCBI36/hg18. We varied the proportions of neutral, beneficial and deleterious mutations within exons, but the distribution of fitness effects (DFE) for both deleterious and beneficial mutations were shared across all apes. The DFE for deleterious mutations was gamma-distributed with a fixed shape α and scale as estimated in Castellano et al. (2019), and the DFE for beneficial mutations followed an exponential distribution (Orr, 2003).

By default, we added neutral mutations to the simulated genealogies with `msprime` so that the total mutation rate (of neutral plus non-neutral mutations, if any) was constant along the genome. In addition, to simulate local variation in mutation rates along the chromosome, we selected three simulated genealogies (the fully neutral simulation, one with deleterious mutations and one with both beneficial and deleterious mutations) to add neutral mutations in a way that resulted in varying levels of neutral mutation rate variation along the chromosome. To do this, we built mutation rate maps by sampling mutation rates for each 1Mb window independently from a normal distribution with mean 2×10^{-8} and standard deviation chosen from $\sigma/2 \times 10^{-8} = \{0.005, 0.007, 0.011, 0.016, 0.023, 0.033, 0.048, 0.070.0.103, 0.150\}$. In simulations with non-neutral mutations, we subtracted the non-neutral mutation rate from the respective window mutation rate for the intervals that intersected with exons. In total, we explored 56 different parameter combinations with all the different simulations (see Table 3.1 and Table A.1 for the parameter space). The code used to produce the simulations can be found at https://github.com/kr-colab/greatapes_sims.

Regime	**Neutral**	**Deleterious only**	**Beneficial only**	**Both**
Proportion of deleterious mutations	0%	10% – 70%	0%	10% – 70%
Proportion of beneficial mutations	0%	0%	0.005% – 0.5%	0.005% – 0.5%
Deleterious DFE	—	Gamma distributed with $\bar{s} = \{-0.015, -0.03\}$ and $\alpha = 0.16$	—	Gamma distributed with $\bar{s} = \{-0.015, -0.03\}$ and $\alpha = 0.16$
Beneficial DFE	—	—	Exponentially distributed with $\bar{s} = \{0.01, 0.005\}$	Exponentially distributed with $\bar{s} = \{0.01, 0.005\}$

Table 3.1. Range of parameters explored in the simulations. Non-neutral mutations were only allowed within exons. "DFE" refers to the distribution of fitness effects. Gamma distribution was parameterized with shape α and mean $\bar{s}{=}\alpha/\beta$, where β is the rate parameter.

3.2.3 Visualizing correlated landscapes of diversity and divergence.

To compare landscapes of diversity and divergence along chromosomes, we computed the Spearman correlation between the landscapes

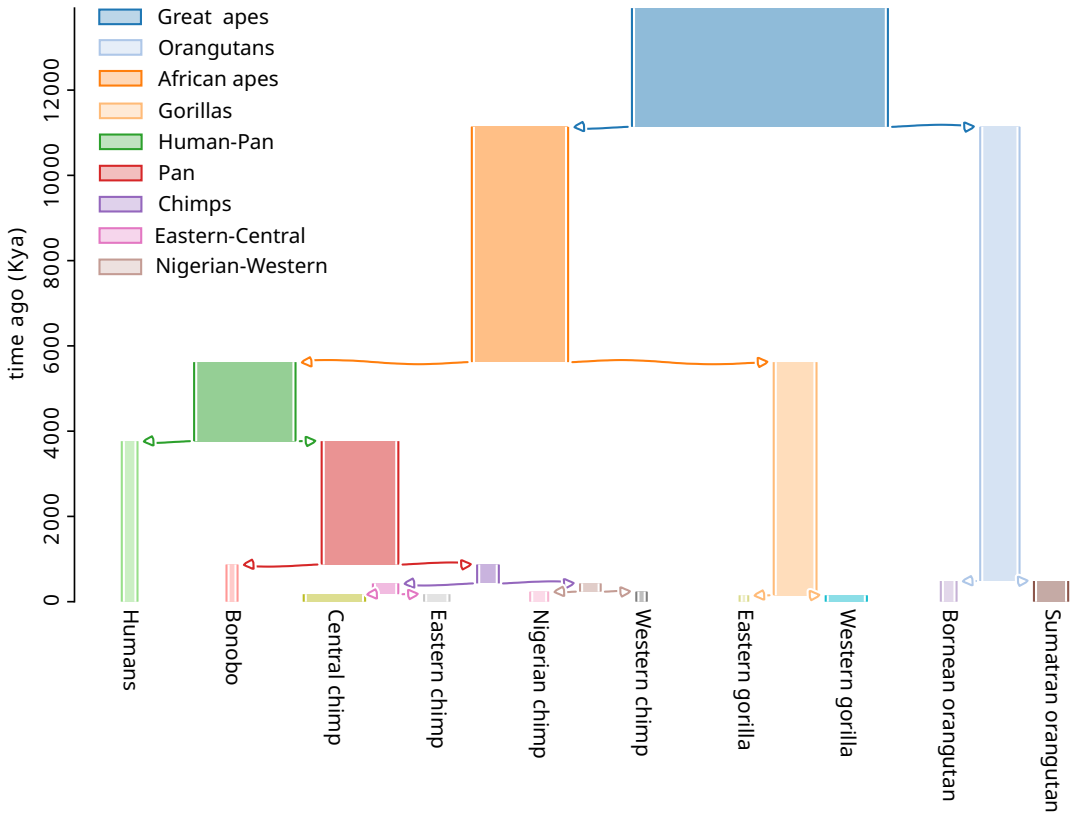


Figure 3.1. Range of parameters explored in the simulations. Arrows indicate population splits. Branch widths are proportional to population size. For example, the population size was $125,089$ for the great apes branch and $7,672$ for the humans branch. Figure was produced using *demesdraw* (Gower et al., 2022).

across windows within a chromosome. Because of computational constraints, we focus on chromosome 12. Chromosome 12 is one of the smallest chromosomes in the great apes, there are no major inversions, and it has good variation in exon density and recombination rate. The choice was made blindly before looking at the data, but we found it behaves similarly to other chromosomes (see Figure A.10 through Figure A.31).

We expected landscapes of two closely related species to be more correlated than the landscapes of two distantly related species. Thus, the correlation between any two landscapes of diversity and divergence is expected to depend on distances between them in the phylogenetic tree. We decided to plot our correlations against distance (in generations) between the most common recent ancestor (MRCA) of each landscape. These distances were computed using the demographic model estimated in Prado-Martinez et al. (2013). In comparing two landscapes of diversity, this amounts to the total distance between the two tips in the species tree. For instance, the phylogenetic distance dT between diversity in humans and diversity in bonobos is the sum of the lengths of the human, pan and bonobo branches in the species tree used for simulation (shown in Figure 3.1). In comparing a landscape of diversity to a landscape of divergence, this amounts to the distance between the species of the landscape of diversity and the MRCA of the two species involved in the divergence. For example, dT for the landscapes of diversity in humans and divergence between Sumatran orangutans and eastern gorillas would be the distance between the humans tip and the great apes internal node. dT for the landscapes of divergence between the orangutans and divergence between the gorillas would be the distance between the orangutan and gorilla

internal nodes. Some divergences may share branches in the tree, but these are excluded from our main figures; see subsection A.0.1 and Figure A.2.

3.3 Results

First, we will provide a qualitative view of the landscapes of diversity and divergence in the great apes. Then, we explore the correlations between landscapes in the real data and how they vary depending on phylogenetic distance. To understand the processes that can drive these correlations, we use forward-in-time simulations of the great apes history under different models (e.g., with and without natural selection). Lastly, we describe how genomic features are related to patterns of diversity and divergence in the real great apes data, and we speculate which processes can explain what we see in the data and simulations.

3.3.1 Landscapes of within-species diversity and between-species divergence.

There is considerable variation in levels of genetic diversity across the great apes (Figure 3.2). Species may differ in overall levels of diversity due to population size history: species with greater historical population sizes (e.g., central chimps and western gorillas) harbor the most amount of genetic variation (Prado-Martinez et al., 2013). Levels of diversity vary along the chromosome, but do not appear to be strongly structured. Instead, diversity seems to haphazardly fluctuate up and down along the chromosome, and this variation might be attributed to neutral genealogical and mutational processes alone. A notable feature is the large dip in diversity around the 50Mb mark, which is so extensive that it almost erases the differences between-species. This dip coincides with three of the windows with the highest exon density, possibly pointing to the role of selection in shaping genetic variation in those windows.

Levels of between-species genetic divergence also vary along the genome, by an even greater amount in absolute terms. Interestingly, diversity (π) varies (along the chromosome) by about 0.2%, whereas divergence (d_{XY}) varies by more than 0.5%. Because $d_{XY} = \pi^{\text{anc}} + rT$ (where π^{anc} is diversity in the ancestor, r is the substitution rate and T is the split time between the two species), this excess in variance may be due to the substitution process. Landscapes of divergence which share their most common recent ancestor (e.g., human-Bornean orangutan and bonobo-Bornean orangutan divergences — both colored in red in Figure 3.2A) overlap almost perfectly with each other. Curiously, divergence seems to accumulate faster in the ends of the chromosome, leading to a "smiley" pattern in the landscape of divergence — which is not apparent in the landscape of diversity. That is, with deeper split times, divergence in the ends of the chromosome seem to increase faster than in other regions of the genome (see how the divergences whose MRCA is the great apes look more like a convex parabola than a horizontal line in Figure 3.2A; see also Figure A.1).

In comparing landscapes across species side by side, a remarkable pattern emerges: levels of genetic diversity and divergence along chromosomes have similar peaks and troughs. To get a sense of how strong this observation is, we can compare it to one of the most well studied properties of genomic variation: the correlation between exon density and genetic diversity. We found that the correlation between human diversity and exon density is -0.2 (at the 1Mb scale), but the correlation between levels of diversity in humans and western gorillas is 0.48. Below, we dissect this observation of strong correlation between landscapes across the great apes and discuss the processes that may cause it.

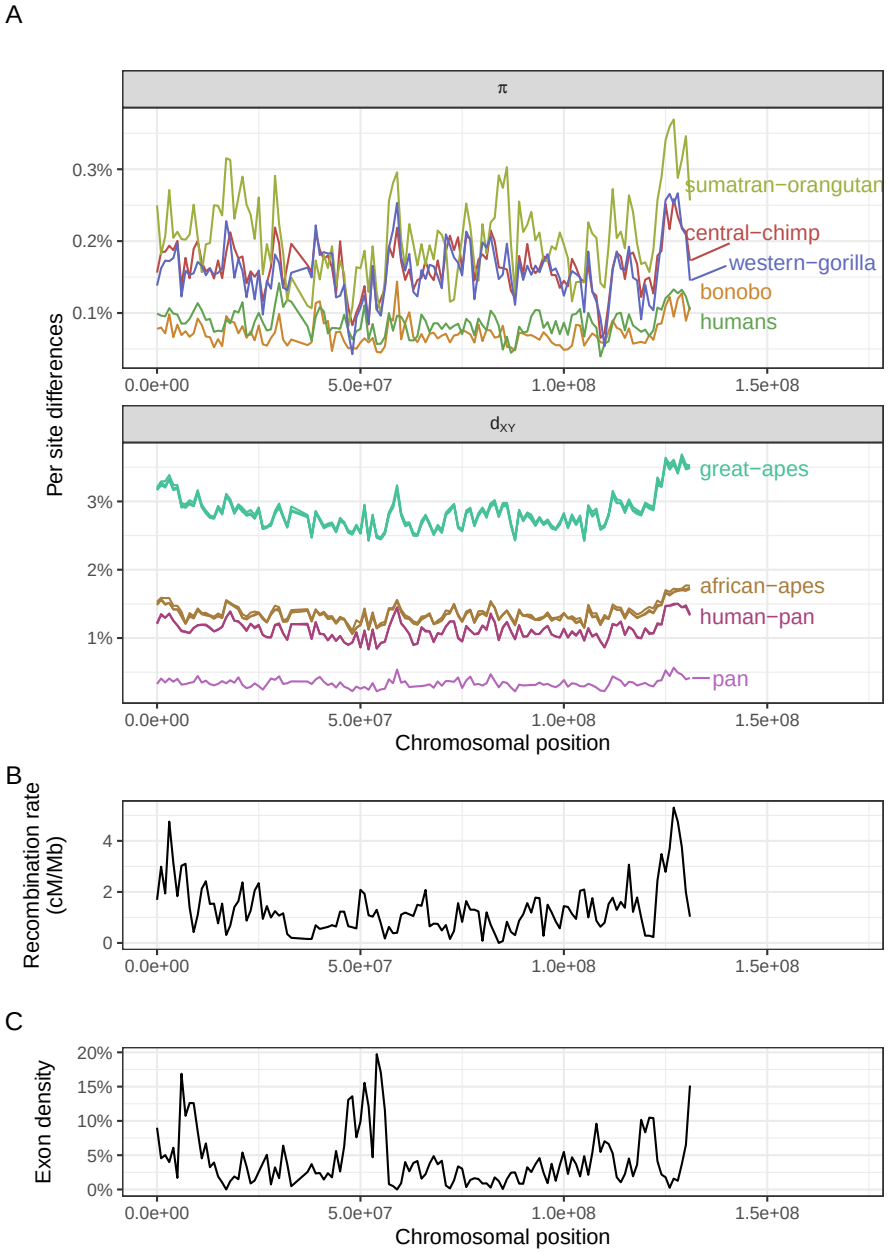


Figure 3.2. A) Landscapes of nucleotide diversity (π) and divergence (d_{XY}) in 1Mb windows along chromosome 12. Lines are colored by species on the top plot and by the most common recent ancestor (MRCA) on the bottom. Genomic windows with less than 40% of callable sites were masked. Only a subset of the species are displayed for clarity. B) Recombination rate estimates from humans (deCODE map; Kong et al., 2002). C) Exon density along chromosome 12, computed as the percentage of callable nucleotides in a window that fall within an exon.

3.3.2 Remarkable correlations between landscapes of diversity and divergence.

The landscapes of diversity and divergence are highly correlated across the great apes. To interpret this signal, we first need to understand what processes can cause such correlations, and so first we describe the toy example depicted in Figure 3.3. Both genetic diversity (π) and divergence (d_{XY}) are estimates of the mean time to the most recent common ancestor (multiplied by twice the effective mutation rate). Populations V and W split recently, and so ancestral variation contributes significantly to within-species diversity (i.e., the coalescences for samples within species happen before the species split). As a result, samples from one population may coalesce first with a sample from another population (e.g., samples v_2 and w_1), a pattern called incomplete lineage sorting (ILS) (see the branch marked with $*$ in the gene tree). This sharing of ancestral variation causes π_V and π_W to be correlated with each other. The probability two samples from V coalesce before the split with W is $1 - e^{\frac{-T}{2N_e}}$, where T is the split time and N_e is the effective population size. Therefore, split time (T) should be a good predictor of the correlation between two landscapes of diversity (and/or divergence). Thus, we decided to visualize correlations between landscapes of diversity and divergence by computing the phylogenetic distance dT, which is simply the distance in generation time between two statistics. For example, we define $dT(\pi_W, d_{XY}) = 2T_{VWXY} - T_{XY}$. Divergences may share branches by definition (irrespective of split times), as you can see with d_{VX} and d_{XY} (see subsection 3.2.3 for more details). In such cases, our chosen metric dT would not be a good proxy for expected correlations, so we omit such cases from our main figures. See subsection 3.2.3 and Figure A.2 for more on the correlations between landscapes that share branches.

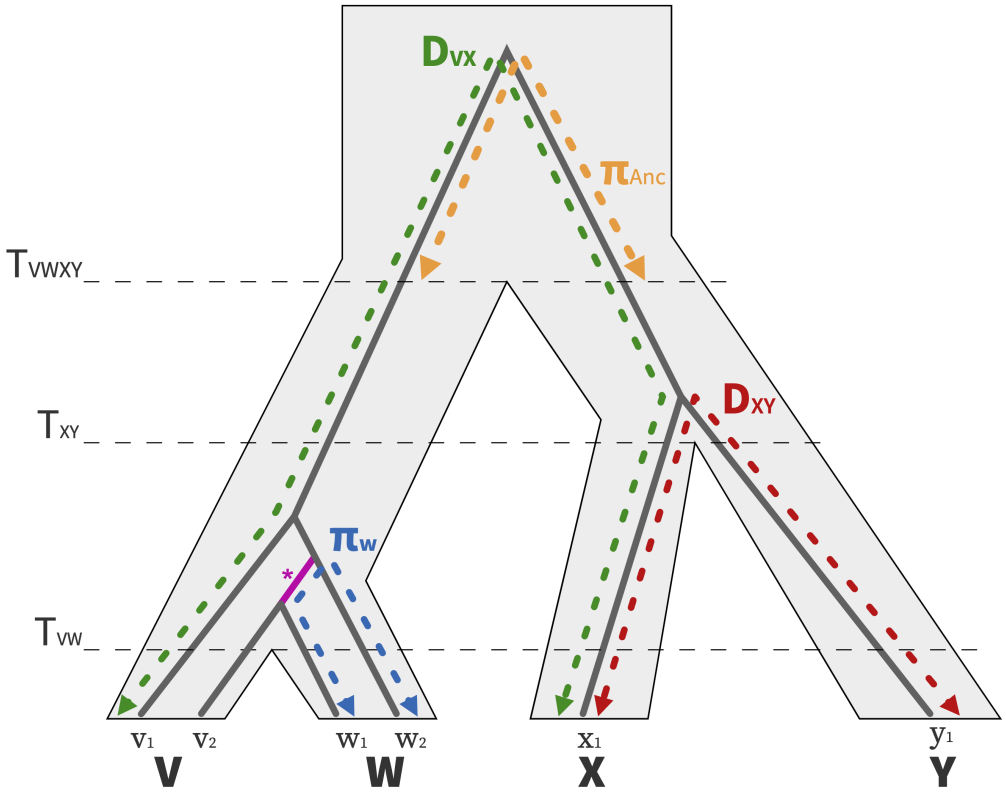


Figure 3.3. Visualizing the relationships between nucleotide diversity and divergence statistics between closely related taxa. A population and gene tree for four populations (V, W, X, Y) are depicted with the light gray polygon and gray solid line, respectively. The branch that is shared between π_V and π_W due to incomplete lineage sorting is highlighted in pink.

Figure 3.4 shows the pairwise correlations between great apes landscapes of diversity and divergence against phylogenetic distance (dT, which is computed from the split times of the model shown in Figure 3.1). We see ancestral variation seems to play a role in structuring correlations between landscapes: pairs of species that recently split have their landscapes of diversity highly correlated. Surprisingly, correlations still plateau at around 0.5. We expect ancestral variation to play a minor role when comparing orangutans and chimps, which separated around $60 \times N_e$ generations ago, but their landscapes are still highly correlated. Population size history seems to affect the correlation between landscapes since the weakest correlations involve the landscape of diversity of one of the species with small historical population sizes (i.e., bonobos, eastern gorillas and western chimps).

Correlations between landscapes of divergence and diversity and between landscapes of divergence are also quite high, often surpassing 0.5, and they also decay with phylogenetic distance (dT) (see middle and right most plots in Figure 3.4). In theory, these landscapes can also be correlated due to ancestral variation. To see how ancestral variation can create correlations even between landscapes with no overlap in the tree, consider Figure 3.3: divergence between X and Y and divergence between V and W can each contain contributions from ancestral diversity if lineages have not coalesced in both branches leading from the ancestor. If a particular portion of the genome happens to have higher diversity in the ancestor, it will also have higher divergence. Since this correlation is produced by sharing of ancestral variation, it is expected to have a very small effect except when branches are short. As discussed in subsection 3.2.3, two divergences can also be correlated by definition (because they share branches in the tree). For example, when comparing human-Bornean orangutan and gorilla-Bornean orangutan

divergence we expect some correlation because these divergences share the large African apes and orangutan branches in the tree (Figure 3.1). In Figure 3.4 we excluded these comparisons where branches are shared. Such comparisons can be seen in Figure A.2. We found that even these comparisons that share branches have an excess of correlation compared to a theoretical expectation (derived from a simplified neutral model), that is the correlations are above the $y = x$ line in Figure A.2 even for distantly related species.

There are many processes that could maintain landscapes correlated. Above, we discussed how we expect ancestral variation to explain these correlations. The alternative would be to have a process that structures variation along chromosomes which is shared across species. Using forward-in-time simulations, we set out to (i) confirm that ancestral variation alone is not causing landscapes to remain correlated, and (ii) test which process or processes that when shared among a group of species could maintain correlations in similar ways to what we observed in the great apes' data.

3.3.3 Neutral demographic processes. To assess the extent to which ancestral variation alone could explain our observations, we performed a forward-in-time simulation of the great apes' evolutionary history. As expected, the resulting landscapes of diversity and divergence are not well correlated (Figure 3.5). Ancestral variation seems to maintain correlations between some landscapes; for instance, the landscapes of diversity in central and eastern chimps have a 0.61 correlation, the highest across all pairs of comparisons (Figure 3.5A, point a). Nevertheless, correlations between landscapes of diversity and divergence decay quickly with phylogenetic distance to 0. Some distant comparisons are moderately correlated (e.g., the landscape of diversity in Bornean orangutans and divergence

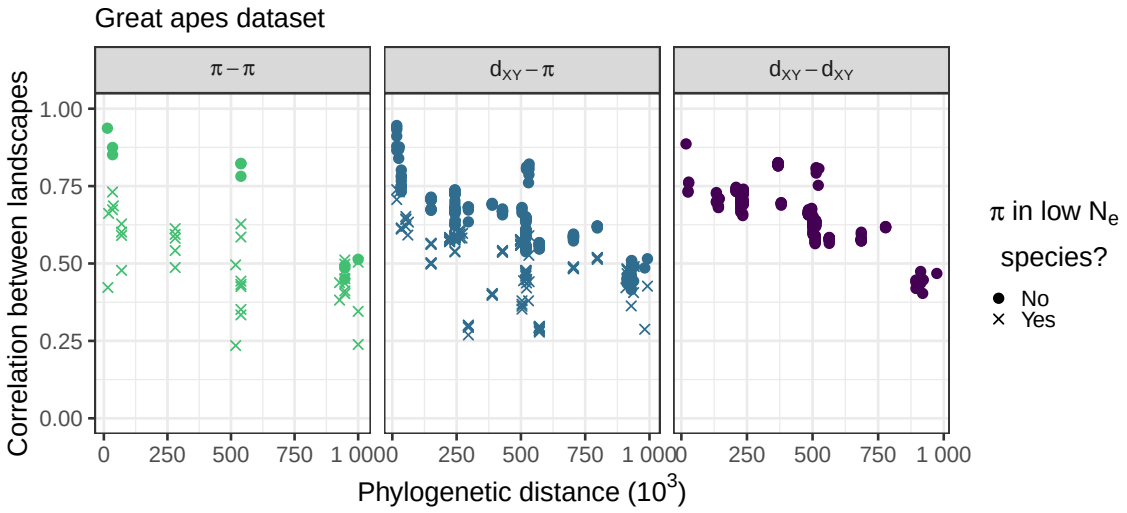


Figure 3.4. Correlations between landscapes of diversity and divergence across the great apes. Each point on the plots correspond to the (Spearman) correlation between two landscapes of diversity/divergence, computed on 1Mb windows across the entire chromosome 12. Correlations were split by type of landscapes compared ($\pi - \pi$, $\pi - d_{XY}$, $d_{XY} - d_{XY}$). dT is the phylogenetic distance (in number of generations) between the most common recent ancestor of the two landscapes compared (e.g., the dT for correlation between landscapes of diversity in humans and divergence between eastern gorillas and orangutans is distance between the humans and the great apes nodes in the phylogenetic tree, Figure 3.1). Note that species with low N_e — for which the estimated species N_e was less than 8×10^3: bonobos, eastern gorillas and western chimps — have a different point shape. Only comparisons for which the definition of the statistics do not overlap are shown, as explained in subsection 3.2.3.

between central and western chimps have a correlation coefficient of 0.23, see Figure 3.5A, point *b*), but that seems to be driven by the outlier window around 80Mb. This outlier window has a recombination rate close to 0 (Figure 3.2C), so the average nucleotide diversity over the window has a higher variance because of coalescent noise (see the extreme peaks and valleys in Figure 3.5). Recombination rate variation can create some moderate correlations, but when we look at multiple species at once it becomes clear that the mean correlation goes to 0.

3.3.4 GC-biased gene conversion. A prominent feature of the landscapes of divergence in the great apes is the faster accumulation of divergence in the ends of the chromosomes (Figure 3.2). This feature was not present in any of our simulations, so we sought to understand its possible causes. Double strand breaks are more common at the ends of chromosomes (Kong et al., 2002, 2010), and these can be repaired either by crossover or gene conversion events. GC-biased gene conversion (gBGC), the process whereby weak alleles (A and T) are replaced by strong alleles (G and C) in the repair of double-stranded breaks in heterozygotes, mimics positive selection – in that it increases the probability of fixation of G and C alleles (e.g., Galtier et al., 2009). We suspected gBGC could have caused the increased rate of accumulation divergence in the ends of chromosomes, as has been observed previously (Katzman et al., 2010), and contributes to the maintenance of correlations between landscapes over long time scales.

To tease apart the effects of gBGC on correlated landscapes, we partitioned divergence by mutation type (weak to weak, strong to strong and weak to strong). If correlations are being driven by gBGC, then we would expect the correlation between landscapes of divergence to be stronger for weak to strong mutations. We found that the overall correlations are very similar across mutation types,

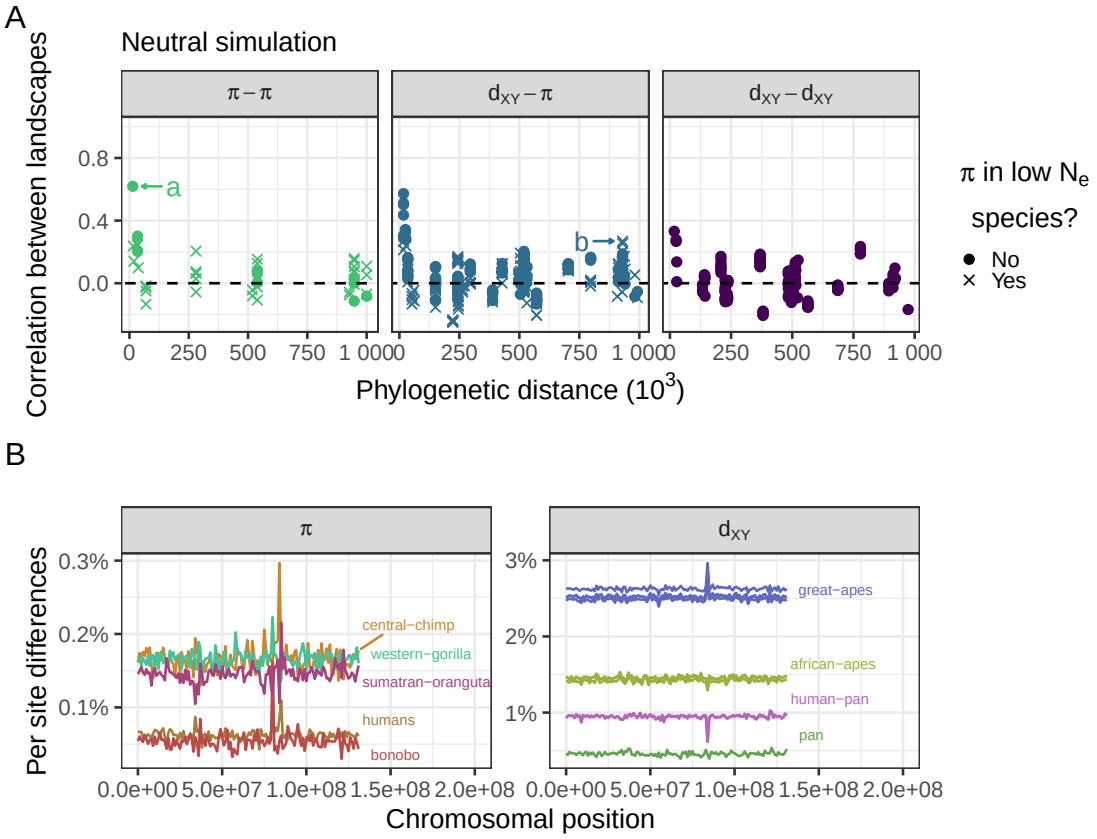


Figure 3.5. Landscapes are not well correlated in a neutral simulation. (A) Correlations between landscapes of diversity and divergence in a neutral simulation. See Figure 3.4 for more details. (B) Nucleotide diversity and divergence along the simulated neutral chromosome. See Figure 3.2A for details.

suggesting gBGC does not play a strong role in structuring the correlations between landscapes (Figure 3.6).

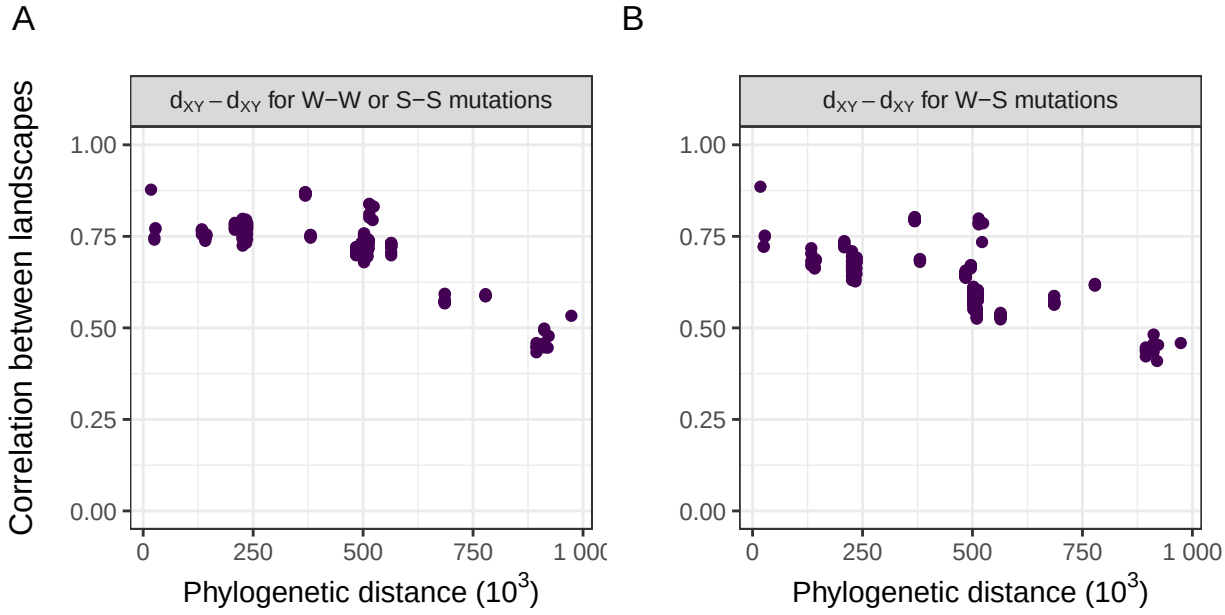


Figure 3.6. Correlations between landscapes of divergence partitioned by site type (W-W/S-S and W-S). W-W sites are sites in which the state did not change between-species (and remained weak which corresponds to A or T). Similar logic applies to S-S sites (S or strong states are G or C). W-S sites are sites in which a new mutation appeared either going from weak to strong or from strong to weak. Note these definitions do not rely on identifying the exact ancestral state, we simply compare the current states in the four species involved (two species per d_{XY} landscape). For example, if by looking at the four species we see the following states A,T,A,T the site would be classified as W-W. If we saw G,A,A,A the site would be classified as W-S. Other details are the same as in the rightmost panel in Figure 3.4.

3.3.5 Positive and negative natural selection. Another process whose intensity is likely correlated across all branches in the great apes tree is natural selection. If targets of selection and recombination maps are shared across species, then we would expect both the direct and indirect effects of selection to be shared across branches. It can be difficult to model natural selection in a

realistic manner because we do not know precisely which locations of the genome are subject to stronger selection. Nevertheless, exons are expected to have higher density of functional mutations than other places in the genome. Thus, we ran simulations in which beneficial and deleterious mutations can happen only within exons. Using human annotations, we simulated the great apes' history assuming a common recombination map and exon locations. See the landscapes resulting from these simulations in Figure A.3.

We found that negative selection can slightly increase correlations between landscapes (Figure 3.7A-C). If 30% of all mutations within exons were strongly deleterious (mean selection coefficient $\bar{s} = -0.03$), landscapes would be weakly correlated (Figure 3.7B). The correlations between landscapes rarely surpass 0.5, even with 70% of all mutations within exons being strongly deleterious (Figure 3.7C).

Positive selection, on the other hand, can quickly increase correlations between landscapes. A beneficial mutation rate within exons of $\bar{\mu_p} = 1 \times 10^{-12}$ produced moderate correlations between landscapes (Figure 3.7D). With too much positive selection, correlations can break down because of the contrasting effects of positive selection on diversity and divergence. That is, while positive selection increases fixation rates and hence divergence between-species, its linked effects decrease diversity within the species. This can create negative correlations between landscapes, as can be seen in Figure 3.7F. Note that some correlations between landscapes of diversity and divergence remain high when the divergence is computed between closely related species (e.g., central and eastern chimps). Divergence is $d_{XY} = \pi^{\text{anc}} + 2rT$, where π_{anc} is diversity in the ancestor, r is the substitution rate and T is the time since species split. Thus, for the divergences

in which the two species split recently are dominated by genetic diversity in the ancestor, correlations between $\pi - d_{XY}$ remain high because $d_{XY} \simeq \pi^{\text{anc}}$.

Positive and negative selection can work synergistically to produce correlated landscapes that look like the real data. For example, comparing figures Figure 3.7D,G,H which differ in rate of negatively selected mutations μ_n, it is possible to see that the correlations between landscapes start to resemble the real data with more deleterious mutations. Figure 3.7H seems to resemble the data fairly well, with $\pi - d_{XY}$ and $d_{XY} - d_{XY}$ correlations plateauing around 0.5. The $\pi - \pi$ correlations are a bit lower than the real data, however. Recent demographic events can affect genetic diversity and although our simulations are heavily parameterized with respect to the effects of selection, we are not capturing all the variation caused by more realistic demographic models. Figure 3.7D and H look very similar to each other. These have the same amount of positive selection, but the first did not have any negative selection. The major difference between them is that with negative selection there is a more clear separation between the correlations involving low N_e species, similar to what is seen in the data.

3.3.6 Mutation rate variation. Since mutation rate can vary along chromosomes, if this mutation rate map were shared across species, it would maintain correlations between landscapes over longer periods of time. To assess this, we used three of our previous simulated genealogies of the great apes and replaced all neutral mutations assuming a common neutral mutation rate map across the phylogeny: for each window, we drew a mutation rate from a normal distribution with mean 2×10^{-8} (the same as all other simulations) and standard deviation μ_{SD}. We found that, under neutrality, a mutation rate map with μ_{SD} close to $7\% \times 2 \times 10^{-8}$ would be needed to get correlations similar to the data

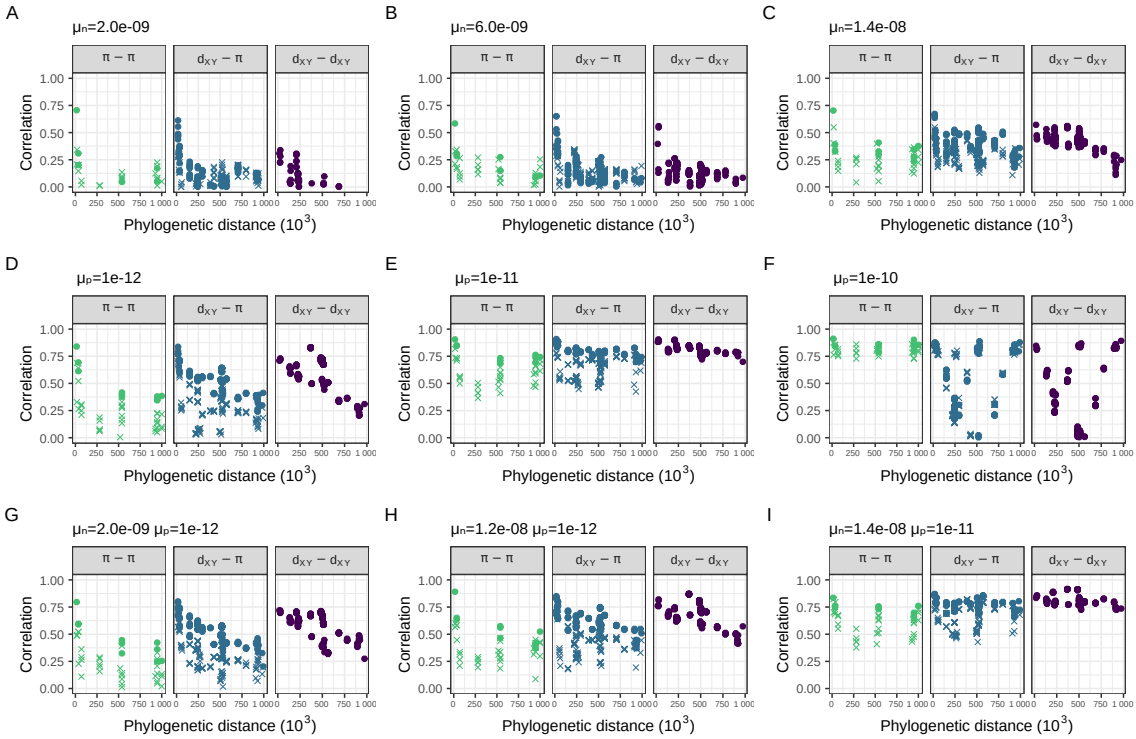


Figure 3.7. Correlations between landscapes of diversity and divergence in simulations with natural selection. (A-C) Simulations with negative selection. (D-F) Simulations with positive selection. (G-I) Simulations with both negative and positive selection. The selection parameters μ_n and μ_p are the rate of mutations in exons with negative and positive fitness effects, respectively. The mean fitness effect was $\bar{s} = -0.03$ for deleterious mutations and $\bar{s} = 0.01$ for beneficial mutations (see subsection 3.2.2 for more details). See how panel H looks the most like the data (Figure 3.4).

(Figure 3.8A-C). Although mean correlations look similar to the data, we see that correlations tend to increase slightly with time in the simulations with mutation rate variation. This is expected because windows with higher mutation rate accumulate divergence faster, creating a correlation with mutation rate that gets stronger with time. In the great apes' data, however, we see a slow but steady decrease in correlations with time.

When we added variation in the neutral mutation rate to simulations with selection, we found that a mutation rate map with a standard deviation of rates of slightly less than $\mu_{\mathrm{SD}} = 7\%$ could plausibly create the correlations observed in the real data (Figure 3.8D-I). The neutral and deleterious simulations with mutation rate variation fail to recover one aspect of the real data: the lower correlations between landscapes that include at least one low N_e species (seen in the $\pi - \pi$ and $\pi - d_{XY}$ comparisons of Figure 3.4). This feature, however, is seen in the simulations with both beneficial and deleterious mutations (Figure 3.8G-I).

3.3.7 Visualizing similarity between simulations and data.

To see how a particular simulation resembles the real data, we can use figures Figure 3.4 and Figure 3.7 to compare how the patterns of all 1260 pairwise correlations between landscapes match the real data. However, it is difficult to assess the fit of the simulated scenarios to real data from such a comparison. Instead, we use principal component analysis (PCA) and create a low dimensional visualization, shown in Figure 3.9, in which each point is a simulation or the real data (shown in yellow). We created this PCA from the 57×1260 matrix in which rows are the simulations and the data, and columns are the pairwise Spearman correlations between landscapes. Unlike in the plots above, here we include the correlations between overlapping landscapes (as detailed in subsection 3.2.3)

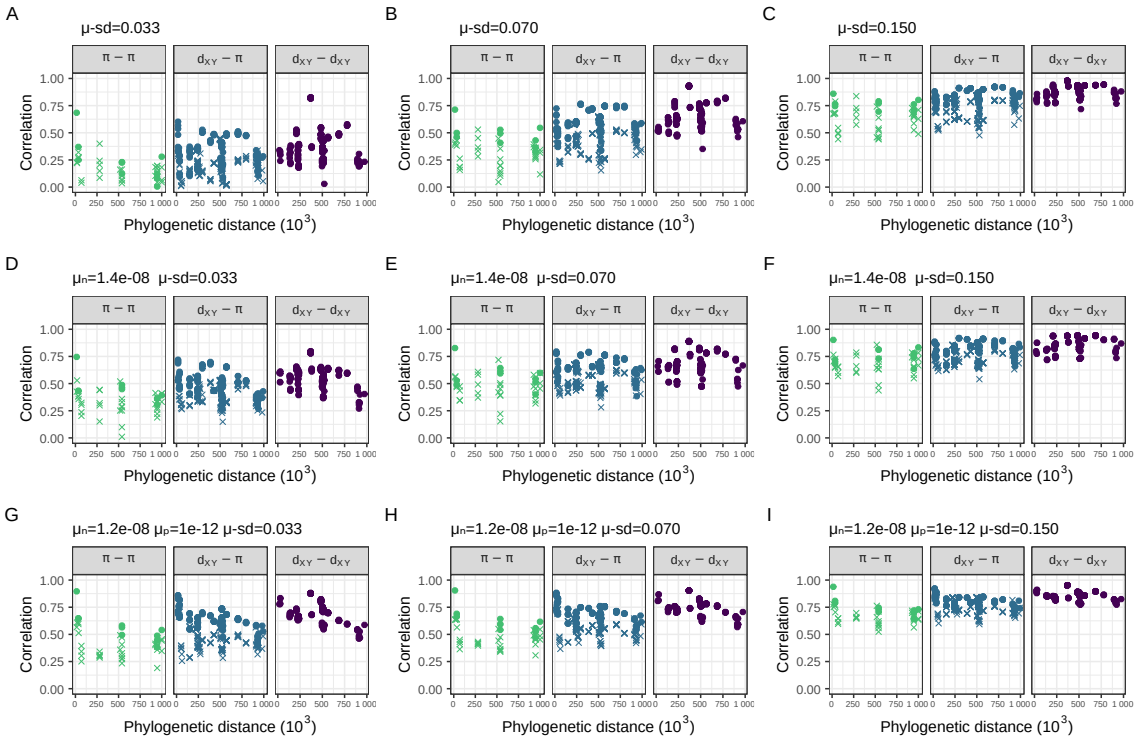


Figure 3.8. Correlations between landscapes of diversity and divergence across the great apes for simulations with variation in mutation rate along the chromosome. Panels A through I show different simulations in which we varied the standard deviation in neutral mutation rate between 1Mb windows, in each setting the standard deviation to the mean mutation rate (2×10^{-8}) multiplied by μ_{SD}. First row (A-C) use a neutral simulation, second row (D-F) a simulation with negative selection, and third row (G-I) a simulation with both positive and negative selection. The selection parameters μ_n and μ_p are the rate of mutations in exons with negative and positive fitness effects, respectively. The mean fitness effect was $\bar{s} = -0.03$ for deleterious mutations and $\bar{s} = 0.01$ for beneficial mutations (see subsection 3.2.2 for more details). See how without selection (A-C), the simulation with $\mu_{\text{SD}} = 7\%$ (panel B) looks close to the data (Figure 3.4). With selection, the simulation with both positive and negative selection and $\mu_{\text{SD}} = 3.3\%$ looks even more similar to the data (correlations between divergences decay over time, and there is a more pronounced differentiation between low and high N_e comparisons).

(Figure 3.9). In PC space, the data most closely resembles a subset of our simulations with both positive and negative selection (e.g., $\bar{\mu_p} = 1 \times 10^{-12}$ and $\bar{\mu_n} = 1.2 \times 10^{-8}$), including no or very little variation in mutation rates (less than 4%).

We also performed PCA on correlations computed at two different scales, 500Kb and 5Mb, in addition to the previously shown results for 1Mb (Figure A.5, Figure A.6). At 500Kb, the observed data are slightly more distant from simulations than at the higher scales, possibly because the recombination map used in simulations had a coarser resolution. Nevertheless, the observed data most closely match the simulations with both positive and negative selection in all scales.

3.3.8 Correlations between genomic features and diversity and divergence. Next, we describe how two important genomic features (i.e., exon density and recombination rate) are related to diversity and divergence in the real great apes data set. The correlations between recombination rate and genetic diversity are positive in all great apes (Figure 3.10A). The strongest correlation between genetic diversity and recombination rate is seen in humans, which is unsurprising given our recombination map was estimated for humans. Recent demographic events also seem to impact the strength of the correlation; for example, the correlation between recombination rate and diversity is higher in Nigerian chimps than in western chimps, which have a much lower recent effective population size. We found that diversity is negatively correlated with exon density across all species (Figure 3.10D). Contrary to what we observed with recombination rate, the correlation between exon density and diversity was even stronger in most other apes than in humans. Species with smaller N_e tend to show weaker correlation between diversity and exon density (see Nam et al., 2017 for related

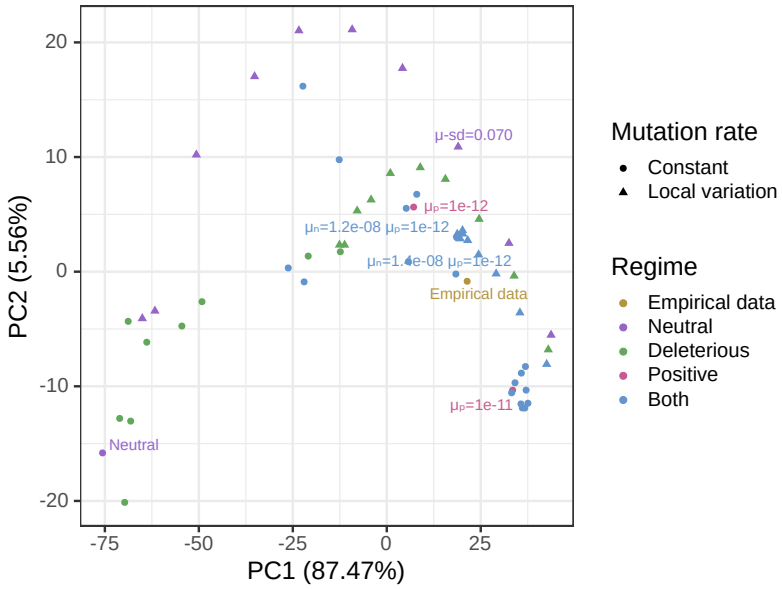


Figure 3.9. Principal component analysis (PCA) visualization of data and simulations. Colors differentiate the empirical data from simulations with different parameters: "Neutral" refers to the simulation without any selection, "Deleterious" refers to simulations with deleterious mutations, "Positive" refers to simulations with beneficial mutations, "Both" refers to simulations with both beneficial and deleterious mutations. The shape of the points differentiate simulations with constant mutation rate along the genome and variable local mutation rates. Local variation in mutation rates were added on top of three simulations: neutral, deleterious with $\mu_n = 1.4 \times 10^{-8}$, and both with $\mu_n = 1.2 \times 10^{-8}$ and $\mu_p = 1 \times 10^{-12}$. The PCA performed on a matrix containing all pairwise correlations between landscapes across the great apes (i.e., all $\pi - \pi$, $\pi - d_{XY}$ and $d_{XY} - d_{XY}$ comparisons) for the great apes dataset and simulations (with selection and with mutation rate variation). We excluded simulations with $\mu_p \geq 1 \times 10^{-10}$ from the PCA analysis because PC2 was capturing negative correlations caused by strong positive selection — as seen in Figure 3.7F.

findings). A striking feature of the correlations of between-species divergence and genomic features, shown in Figure 3.10, is that the correlations get stronger with the amount of phylogenetic time that goes into the comparison (i.e., the T_{MRCA}), in a way that is roughly linear with time.

To describe why this increase in correlation with time might occur, we turn to an analytic approach. Genetic divergence (D) in the i^{th} window between two species that split t generations ago can be decomposed as:

$$D_i(t) = \pi_i(t) + R_i t + \varepsilon_i,$$

where $\pi_i(t)$ is the genetic diversity in the ancestor at time t, R_i is the substitution rate in the window and ε_i is a contribution from genealogical and mutational noise (which has mean zero). This decomposition follows from the definition of genetic divergence as the number of mutations since the common ancestor, as depicted in Figure 3.3 (see how $D_{VX} = \pi^{\text{anc}} + 2RT_{VWXY}$).

The covariance between $D(t)$, the vector of divergences along windows, and a genomic feature X is, using bilinearity of covariance,

$$\text{Cov}(D(t), X) = \text{Cov}(\pi(t), X) + t\,\text{Cov}(R, X) + \text{Cov}(\varepsilon, X). \tag{3.1}$$

Happily, this equation predicts the linear change of the covariance with time that is seen in Figure 3.10C and perhaps Figure 3.10F. However, caution is needed because the correlation between diversity and the genomic feature ($\text{Cov}(\pi(t), X)$) may be different in different ancestors, and indeed the inferred effective population size is greater in older ancestors in the great apes (Figure 3.1).

Next consider covariances of diversity with recombination rate,
Figure 3.10C. Consulting the equation above, the fact that the covariance between divergence and recombination rate increases with time can be caused by two

factors (taking X to be the vector of mean recombination rates along the genome): (i) a positive covariance between substitution rates and recombination rates ($\mathrm{Cov}(R, X) > 0$), and/or (ii) greater genetic diversity in longer ago ancestors ($N_e(t)$ larger for larger t). It is unlikely that the increase in N_e in more ancient ancestors was sufficient to produce the dramatic increase in covariance seen in Figure 3.10C, since it would require $\mathrm{Cov}(\pi(t), X)$ to be far larger in the ancestral species than is seen in any modern species. On the other hand, there are various plausible mechanisms that would affect $\mathrm{Cov}(R, X)$. One factor that certainly contributes is the "smile": we found that divergence increases faster near the ends of the chromosomes where recombination rate is greater, probably in part because of GC-biased gene conversion. Interestingly, positive and negative selection are predicted to have opposite effects here: greater recombination rate increases the efficacy of both through reduced interference among selected alleles, so positive selection would increase substitution rate and hence increase $\mathrm{Cov}(R, X)$, while negative selection would decrease $\mathrm{Cov}(R, X)$. When considering only the middle half of the chromosome (i.e., excluding the effect of gBGC) (Figure A.7), the covariances between divergence and recombination rate flip to negative, and they continue to decrease over time. Thus, it seems that negative selection is the most important driver of divergence in the middle, whereas gBGC strongly affects the tails of the chromosome.

The covariance of diversity and exon density has a less clear pattern (Figure 3.10F), although it generally gets more strongly negative with time. This decrease could be a result of a negative covariance between substitution rates and exon density and/or an increase in the population sizes of the ancestors (if $\mathrm{Cov}(\nu, X) < 0$, as expected since ν is relative diversity and X is now exon density).

As before, positive selection in exons would be expected to produce a positive covariance between exon density and substitution rate, while negative selection would produce a negative covariance. It is hard to determine *a priori* which is likely to be stronger, because although negative selection is thought to be much more ubiquitous, a small amount of positive selection can have a strong effect on substitution rates. The fact that covariance generally goes down with time suggests that negative selection (i.e., constraint) is more strongly affecting substitution rates.

It is at first surprising that the correlations between exon density and divergence go up with time, but the covariances go down with time (Figure 3.10E,F). However, correlation is defined as $\mathrm{Cor}(D_t, X) = \mathrm{Cov}(D_t, X)/\,\mathrm{SD}(D_t)\,\mathrm{SD}(X)$. Thus, if the variance in divergences increases over time the correlations will decrease over time. Indeed, we see this happening as gBGC increases divergences on the ends of the chromosome faster than in the middle, leading to an increase in variance of divergence along the genome. This also explains why correlations of landscapes of very recent times are very noisy, but covariances are not. Indeed, the patterns are clearer when we exclude the tails of the chromosome (Figure A.7): there is only a modest increase in the correlation between exon density and divergence over time and the covariances go down with time more linearly.

3.4 Discussion

A central goal of population genetics is to understand the balance of evolutionary forces at work in shaping the origin and maintenance of variation within and between-species (Lewontin, 1974). While the field has been historically data-limited, with the current flood of genome sequencing data, we are poised to make progress on such old questions. Over the past decades, an important lever

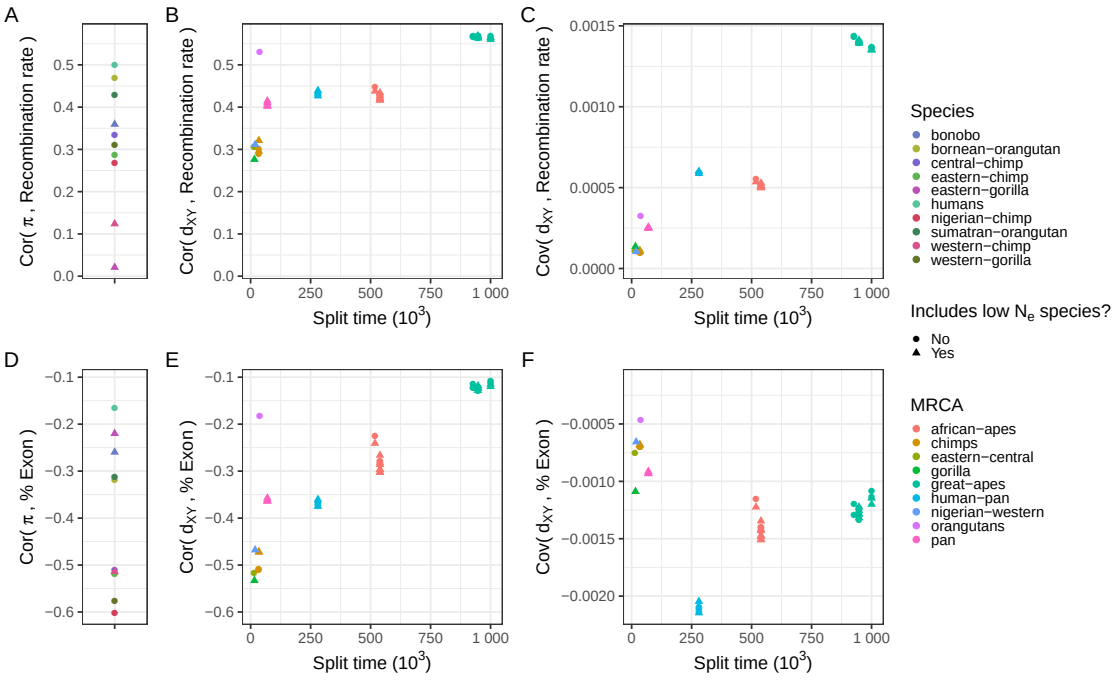


Figure 3.10. Correlations and covariances between landscapes of diversity and divergence and annotation features in the real great apes data. Exon density and recombination rates were obtained as detailed in Figure 3.2. Split time is the time distance between the two species involved in the divergence. Points are colored by the species of within-species diversity (π) in plots A and D. In plots B,C,E,F, the points are colored by the most common recent ancestor of the species for which between-species divergence was computed. Species with low N_e — for which the estimated species N_e was less than 8×10^3: bonobos, eastern gorillas and western chimps — have a different point shape.

in understanding the relative impact of genetic drift versus selection in shaping genomic patterns of variation has been to examine the relationship between *levels* of diversity and genomic features, such as recombination rate and exon density. The overarching observation has been that regions of reduced crossing over generally harbor less variation than regions of increased crossing over in many but not all species (e.g., Begun & Aquadro, 1992; Corbett-Detig et al., 2015). This observation is consistent with a role for linked selection shaping patterns of variation in recombining genomes, but the relative contributions of deleterious and beneficial mutations is still largely unknown. Indeed, it seems likely that some complex mixture of both processes shapes variation in natural populations (Kern & Hahn, 2018).

In this paper, we moved beyond genetic diversity within a single species to look at how divergence between closely related species changes with time and how this correlates with genomic features. Previous studies (e.g., Stankowski et al., 2019) looked at similar patterns (in monkeyflowers) and found strong correlations between landscapes of diversity and divergence between related species, despite deep split times. Landscapes of closely related species can remain correlated for two main reasons (i) shared ancestral variation or (ii) shared heterogeneous process. If two species recently split, their landscapes of diversity are expected to be correlated due to shared ancestral variation. If the process that structures genetic diversity along chromosomes is heterogeneous and somewhat shared between-species, then their landscapes are expected to remain correlated over longer periods of time. For example, if the effects of selection are concentrated in the same genomic regions in two species, then their landscapes of diversity will be correlated. By incorporating information from multiple species at once, we are able to pool

information across species and thus increase our power to disentangle the role of different evolutionary forces. Patterns across multiple species are more likely to be robust to the idiosyncrasies of any one species, such as demographic history. For instance within-species metrics can be confounded by demography: demographic events can create spurious troughs of diversity (Simonsen et al., 1995) or exacerbate the effects of background selection on diversity (Torres et al., 2018). However, correlations between landscapes can only be produced due to shared ancestral variation or a shared heterogeneous process.

In the great apes, we found that landscapes of within-species diversity and between-species divergence are highly correlated across the phylogeny. Those correlations are often stronger than those that have been historically used as evidence for the effects of selection on genetic variation. For example, the correlation between genetic diversity in humans and exon density is -0.2, yet the correlation between diversity in humans and diversity in western gorillas is 0.48. This stronger correlation may not be entirely due to shared landscape of selection — it may also be a result of shared ancestral variation (and incomplete lineage sorting), mutation rate variation, and/or GC-biased gene conversion. To understand how much of the correlation between landscapes can be attributed to ancestral variation, we performed extensive simulations of the great apes' evolutionary history, and found that ancestral variation explains very little of the correlations we observed. Thus, a shared heterogeneous process seems to be needed to explain the data.

Two neutral processes can be heterogeneous along the genome and shared across species: GC-biased gene conversion and mutation. GC-biased gene conversion (gBGC) is thought to be an important factor in shaping levels of

variation in humans (Chen et al., 2007; Glémin et al., 2015; Pouyet et al., 2018), and it has similar effects to those of natural selection. However, if gBGC were a major driver of correlations we would expect to see a difference in overall levels of correlation between different classes of substitution, and we do not (Figures 3.6 and A.4). As such gBGC seems to be a minor contributor to the correlations we observe, although it does seem to be leading to increased substitution rates near the telomeres (where divergences are increasing roughly 5% faster; see Figure 3.2 and Figure A.1). In birds, an excess of divergence near telomeres has been attributed to meiotic drives (Ellegren et al., 2012).

When the history of the great apes is simulated with a shared heterogeneous mutation map, correlations between landscapes do emerge. These were as strong as seen in the data when the rates were drawn from a normal distribution with a standard deviation of the mutation rate of at least a 7% of the mean mutation rate. However, our mutation map was perfectly shared among was species in our simulations, so it is possible that a mutation map which changes over time might move closely to match the data. T. C. A. Smith et al. (2018) estimated the standard deviation of de novo mutation rate in humans at the 1Mb scale to be around 25% of the mean mutation rate. However, the lack of congruency in de novo mutations identified in different data sets raises questions about the role of ascertainment biases that need to be addressed in future studies (Castellano et al., 2020). Our simulations showed a facet of shared mutational heterogeneity along the genome that we do not observe in real data: with variable mutation rate correlations increase over time, whereas in the real data they decrease. It is unknown how conserved mutation rate heterogeneity is across the great apes, so it remains to be seen how an evolving heterogeneous mutation rate map affects

landscapes of diversity and divergence. A major driver of mutation rate variation stems from CpG dinucleotides, which have much higher mutation rates than other sites (Agarwal & Przeworski, 2021; Hodgkinson & Eyre-Walker, 2011; Nachman & Crowell, 2000). Nevertheless, when we partitioned the landscapes of divergence by mutation types, we did not see an excess of correlation between landscapes with mutations that can be affected by CpG-induced mutation rate variation (Figures 3.6 and A.4).

Natural selection can also structure genetic variation heterogeneously along the genome. In simulations, both positive and negative selection are needed for the correlations between landscapes to resemble the data. By examining the correlations between landscapes (summarized in Figure 3.9), we found that the best fitting simulation is the one with a beneficial mutation rate within exons of 1×10^{-12} and deleterious rate within exons of 1.2×10^{-8}. Positive selection seems to be needed to explain one particular feature of the data: the separation between correlations involving a low N_e species (i.e., correlations are lower if diversity is computed in a species with a low N_e – as seen in humans, bonobos, western chimps and eastern gorillas; see Figure 3.4). Bottlenecks can erase sweep signatures (Jensen et al., 2005; Nielsen et al., 2005; Przeworski, 2002), but demography does not affect local variation in mutation rates, and it can exacerbate signatures of background selection (Torres et al., 2018). Thus, if sweeps are causing correlations between landscapes, we expect it to be more sensitive to the strong bottleneck in humans than the other processes. This conclusion largely agrees with previous studies which found that positive selection is necessary to explain reduction in genetic diversity surrounding genes in the great apes (Nam et al., 2017).

Another way we might characterize our simulations is through examination of substitution processes. In our best fitting simulation, we get a fixation rate of beneficial mutations of around 1×10^{-9} per generation per exon base pair, what amounts to approximately 9% of the fixations within exons (along the human lineage), i.e., about one new fixation of a beneficial mutation every 250 generations. Fixation rate within exons is decreased to around 60% of the rate in our neutral simulation due to the constant removal of deleterious mutations within these regions. Indeed, previous studies (Boyko et al., 2008; Laval et al., 2021; Zhen et al., 2021) have estimated that between 10% and 16% of amino acid differences between humans and chimpanzees were caused by positive selection, which is strikingly similar to our best fitting simulation. We would expect to see the fixation of around 16 beneficial mutations in the past 4000 generations, which is close to the number of hard sweeps genome scans for selections have found in humans over this same time period (Schrider & Kern, 2016, 2017). Our best fitting simulation with selection assumes that 60% of new mutations within exons are deleterious, similar to estimates from the site frequency spectrum (Boyko et al., 2008; Huber et al., 2017; B. Y. Kim et al., 2017). Thus while we have not done exhaustive model fitting due to computational constraints, our simulations reproduce estimates from studies which model a different facet of genetic variation (i.e., the site frequency spectrum).

Heterogeneous processes that correlate with a genomic feature will create differences in rates of substitution along the genome that correlate with the genomic feature. As shown in Equation (3.1), this implies that the covariance along the genome between a genomic feature and divergence is expected to increase with time, and the rate of increase is equal to the covariance between

that feature and the substitution rate. (It is important to note that varying covariances with ancestral diversity can be a confounding factor, and that the observation applies to covariance, not correlation.) Indeed, the covariance between divergence and recombination rate increases roughly linearly with time (see Figure 3.10C), as expected because the rate of gBGC-induced fixations are correlated with recombination rate. Once this effect is removed (see Figure A.7F), the covariance between exon density and divergence decreases linearly with time, as we would expect due to the effects of negative selection directly removing deleterious mutations in or near exons. The magnitude of this slope might produce a quantitative estimate of the strength of this effect, although more work is needed to disentangle confounders. It is important to contrast this observation, which applies mostly to the direct effects of selection, to other observations which also include linked effects (as discussed in Phung et al., 2016).

Although simulations allow for more biological realism, we made assumptions to constrain the parameter space explored. Our simulations used randomly mating populations of constant size, as inferred in Prado-Martinez et al. (2013). These population sizes were inferred using neutral model, and so are likely affected by the effects of selection (Jensen et al., 2005). Mean levels of diversity and divergence in our neutral simulation match the data, but simulations with natural selection differ, at times substantially (Figure A.9). On the other hand, our simulations with selection match the data more closely with respect to standard deviation in levels of diversity and divergence along genomes. However, inaccuracy of the demographic model should not affect any of our main observations, because the effects of demography on levels of variation along genomes are not shared across multiple species.

We chose exons to be the targets of selection in our simulations. Exons cover about 1% of the human genome, and in reality selection affects non-coding regions as well. However, a substantial portion of this selection affects *cis*-regulatory regions, whose density along the genome is well-predicted by coding sequence itself. Furthermore, highly conserved noncoding sequences have long been identified and characterized as functional (Bejerano et al., 2004; Katzman et al., 2007; Siepel et al., 2005). In the great apes, non-coding diversity is correlated with recombination rate, pointing to the role of selection (Castellano et al., 2020). However, it would be circular to include conserved noncoding elements in our simulations because such elements are identified based in part on levels of divergence, which themselves depend on ancestral levels of genetic diversity. Because conserved noncoding elements generally occur close to coding regions of the genome (at the 1Mb scale, the correlation between density of exons and PhastCons elements is around 0.6), we might expect a more realistic model to have the same amount of selection (in terms of total influx of selected mutations), but spread out over a somewhat wider region of the genome since we have omitted such sites. Even without considering all potential targets of selection, patterns of genetic diversity in our simulations match the data well: we see a correlation between simulated and observed diversity of 0.45 for chimps (Figure A.8), for example.

While it has long been recognized that genetic variation among species might be structured similarly due to shared targets of selection, our results demonstrate that these correlations contain important information about the processes at work that has yet to be utilized fully. Here we have used large-scale simulations to demonstrate the combination of forces required to pattern shared divergence and diversity as we observe it in nature. Indeed, our results show that

a combination of negative and positive selection, GC-biased gene conversion and mutation rate variation all contribute in shaping genetic variation in the great apes. Although some processes are not necessarily needed to recapitulate the real data (e.g., mutation rate variation), positive selection seems to be the only force that can explain most of our observations. There is clearly a need for future analytical work that might describe expected correlations across the genome given variation in local mutation, recombination, and selection. Further, statistical model fitting, based on theory or simulation is clearly desirable, although our experience suggests that the latter approach would prove computationally expensive.

3.5 Bridge

As presented in Chapter III, simulations can be useful to help us distinguish between evolutionary models. We showed that although many processes are consistent with great apes population genomic data, positive selection seems necessary to fully explain the observed data. Beyond qualitative assessments, in many cases we might want to infer parameters from a model. The evolutionary simulations of the entire great apes history were too costly to allow for proper parameter estimation. In Chapter IV, we present a new method for simulation-based parameter estimation using whole-genome genealogies. This method is expected to resolve a different bottleneck that plagues evolutionary inference: the scaling with number of samples and genome sizes. Whole-genome genealogies are much more efficient than genotype matrices (matrices with dimensions of number of samples by number of sites). Moreover, they are expected to organize genetic information in a way that can be more conducive for estimating evolutionary parameters. Our main contribution, a deep learning architecture that takes whole-genome genealogies as input, seems to perform well over different tasks. More importantly, it could enable evolutionary inference over biobank scale datasets (that contain millions of samples).

CHAPTER IV

A POWERFUL MACHINE LEARNING FRAMEWORK FOR EVOLUTIONARY INFERENCE USING WHOLE-GENOME GENEALOGIES

Nathaniel S. Pope, Peter L. Ralph and Andrew D. Kern are co-authors on this manuscript. Co-authors and I conceptualized the study, Nathaniel S. Pope and I developed the machine learning framework, I performed analyses with input from my co-authors, I wrote the manuscript with editorial assistance from my co-authors.

4.1 Introduction

Processes like mutation, recombination, assortative mating and selection leave footprints on genetic variation. A major goal of population genetics has been to invert this relationship and infer past, unobserved evolutionary events from genetic variation data (Schraiber & Akey, 2015). With the dramatic increase in our ability to generate whole-genome data, there is a growing need for more efficient inference methods capable of scaling well to handle over tens of thousands of individuals.

Although our interest is to infer evolutionary processes from population genetic data, it is usually simpler to first think about the effects of these processes on the shape of underlying genealogies (Wakely, 2016), and how this ultimately translates to observed genetic variation. In the absence of recombination, a genealogy is a tree that describes the shared ancestry of a sample of DNA molecules (e.g. haploid individuals), until they "coalesce" into a single common ancestor. The shared ancestry of individuals reflects historical processes that acted upon their ancestors. For example, recent contractions in population sizes will lead to genealogies where individuals coalesce relatively recently.

For recombinant DNA, the genealogy is not a tree but instead a graph that encodes the transmission of genetic material from ancestors to present-day individuals at every point in the genome, known as the "ancestral recombination graph" (ARG). The ARG records two types of historical events: coalescences (where separate lineages merge into a common ancestor), and recombinations (where a single lineage splits into multiple lineages, moving backward in time). Crucially, ancestors are "local" in the sense that only a subset of the ancestors in the entire ARG are associated with each location in the genome. Thus, one factorization of ARG is into a sequence of correlated marginal trees, where changes in tree structure are caused by recombination events (Kelleher et al., 2016a, 2019; Speidel et al., 2019). The branches of these trees have two dimensions: the time separating descendant from ancestor, and the length of sequence spanned by the tree containing the branch. A second factorization is into the minimal set of ancestral haplotypes and transmission paths between these ancestors, which is equivalent to collapsing coalescence nodes and branches across marginal trees.

If a genetic variant is segregating in a sample of present-day individuals, then the mutational events leading to that variant can be mapped to particular edges in the ARG. In particular, neutral mutations occur on a given edge at a rate proportional to its area. Thus, the distribution of variant frequency reflects the shape of the underlying ARG. More generally, summary statistics of frequencies of neutral variants are noisy measurements of equivalent summary statistics of branch areas across marginal trees in the underlying ARG (Ralph et al., 2020).

A typical strategy for evolutionary inference is to compare summary statistics (based on genotype matrices) to their expectations under a parametrized model. For example, the site frequency spectrum (SFS) describes the distribution

of allele frequencies across sites in the genome, and is frequently used to infer the demographic histories of natural populations (Gutenkunst et al., 2009; Schraiber & Akey, 2015). However, individual statistics that summarize genetic variation cannot capture all aspects of the data that are informative of the underlying processes. These limitations may be circumvented to some degree by using multiple summary statistics either in a composite likelihood framework or with likelihood-free methods (Caldas et al., 2022; DeGiorgio et al., 2016; Nielsen et al., 2005; Pavlidis et al., 2013; Sheehan & Song, 2016); and by stratifying data into windows across chromosomes (Flagel et al., 2019; Schrider & Kern, 2016; Sheehan & Song, 2016). However, there is inevitably a tradeoff between granularity and informativeness, both in the choice of summary (from raw genotypes to compact summaries like the SFS) and scope (from single bases to large windows).

Assume the underlying ARG for a dataset could be observed. Using the ARG as an input would solve many of the issues above (Rasmussen et al., 2014). This is because it is simultaneously: (A) sufficient, in the sense that it contains all information obtainable from observed genetic data; (B) compact, in the sense that shared genetic variation is encoded as relationships between ancestral haplotypes; (C) multiscale, in the sense that the span of ancestral haplotypes and edges naturally reflects the "spatial" and temporal scales at which shared ancestry impacts variation, without arbitrary discretization into windows.

This core idea of using ARGs for evolutionary inference has two challenges in practice. First, we cannot observe the ARG directly but instead we have to infer it. There are approaches for inferring ARGs in a more-or-less scalable way (Kelleher et al., 2019; Speidel et al., 2021; Zhang et al., 2023). But these will inevitably contain error (e.g., in genotyping, phasing, etc.), so any inference method

using inferred ARGs as input must be able to model error either explicitly or implicitly. Second, we need some way to "parameterize" the ARG in terms of quantities we are interested in. This would traditionally be done by specifying some generative process, like the coalescent with recombination, and calculating likelihood of observed genetic data conditional on this process (Fan et al., 2023). However, these likelihoods are only tractable under a very small class of models. Likelihood-free inference seems to be the most promising solution, specially now given recent advances in population genetic simulation tools that allow ARG recording (Haller et al., 2019; Kelleher et al., 2016a; Ralph et al., 2020).

Graph neural networks (GNNs), an emerging class of deep learning algorithms, provide a natural way to exploit ARGs for evolutionary inference. These networks fall into the broader message passing paradigm whereby information is aggregated in neighborhood of nodes, ultimately yielding an embedding for nodes in a graph (Battaglia et al., 2018; Bronstein et al., 2017; Hamilton et al., 2018). These node embeddings can then be used for downstream tasks such as node classification, graph classification, and edge prediction. GNNs can leverage information from genealogies (i.e., historical relationships between nodes and the timing of coalescent events) (Korfmann et al., 2023), but they would not take into account additional structure in ARGs induced by the spatial (along the chromosome) and temporal (from parent to child) ordering of nodes.

Here, we present a new method that uses whole-genome genealogies for evolutionary inference. Our main goal is to develop an architecture that can efficiently use genealogies for inference at different levels. As a proof-of-concept, we test our method on dating mutations in an ARG. Our approach, tsNN, outperforms an out-of-the-box graph neural network and current likelihood-based

approaches. Taken together, our results demonstrate that tsNN is a powerful and flexible framework for leveraging information from whole-genome genealogies for evolutionary inference.

4.2 Methods

4.2.1 tsNN. Taking inspiration from the graph neural network literature (Rossi et al., 2020), we developed an encoder that can take whole-genome genealogies as input and generates node embeddings (Figure 4.1). In summary, we obtain a lower dimension representation of these genealogies by passing messages between nodes. We start with an initialized embedding for nodes, and then we use an update scheme that takes an edge and updates the embedding of both child and parent nodes. Because the flow of information is structured from parent to child and from left-to-right, we hoped this neural network would learn a better representation of a tree sequence than a more naive approach, such as a simpler graph neural network. The node embeddings can then be decoded to obtain embeddings at different levels (e.g., at the edge, mutation or tree sequence level).

First, we define an encoder module for obtaining node embeddings. We randomly (or arbitrarily) initialize a vector with node features H of dimension number of nodes by number of node features. From a given tree sequence, we also compute a vector E of edge features of dimension number of edges by number of edge features. We define two edge features: the span (width in number of base pairs that an edge spans, scaled by the mutation rate) and edge length (the number of mutations on an edge). We traverse the tree sequence over edge differences (from letft-to-right, and optionally from right-to-left), and with each edge addition, we build a message that is used to update the node features H. The message for the edge that links the parent node p to the child node c is

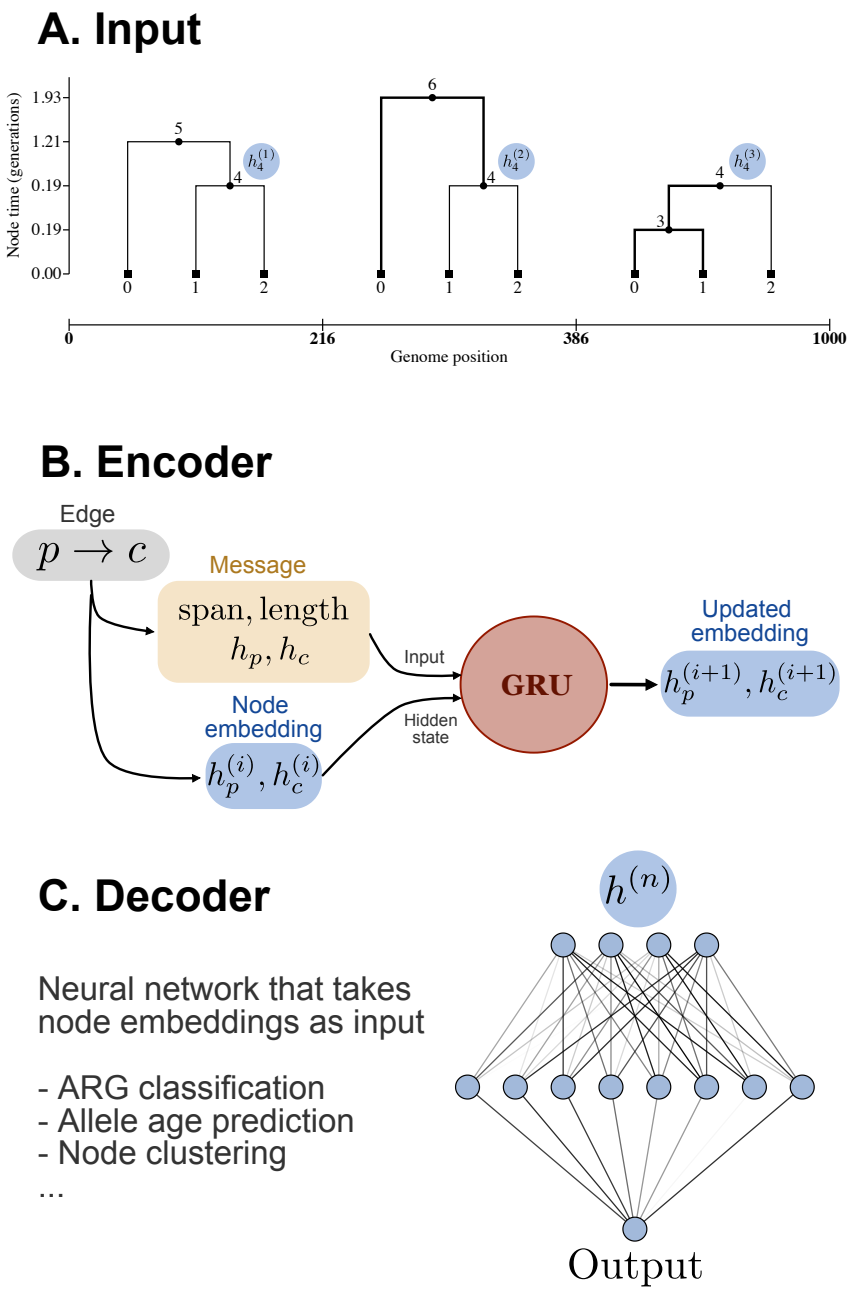


Figure 4.1. Schematic representation of the tsNN algorithm. A) An example tree sequence that can be used as input to tsNN. Edges that differ from the previous tree are highlighted with thicker lines. Note how the embedding of node 4 (h_4) is updated with edge differences along the tree sequence. B) The tsNN encoder updates node embeddings as it iterates over edge insertions along the tree sequence. The update is done with a Gated Recurrent Unit (GRU) that takes the span and length of the edge along with the node embedding for the parent and child nodes. C) The node embeddings can be decoded in different ways for a supervised task.

$$m_{pc} = e_{pc} \parallel (h_p \parallel h_c)$$

where e_{pc} is the row of E for the edge $p - c$, and h_p and h_c are the rows of H for nodes p and c. h_p and h_c are then updated with a Gated Recurrent Unit cell, such that

$$h_p^{\text{out}} \parallel h_c^{\text{out}} = \text{GRU}(m_{pc}, h_p \parallel h_c)$$

The node embedding module above can be combined with different neural network architectures to decode the node embeddings to infer parameters at different levels. For our mutation time task, we first compute edge embeddings, where for each edge we concatenate the corresponding parent and child node embeddings. For each mutation, we assign its embedding as the corresponding edge embedding (we ignore mutations that map to more than one edge). Then, we feed these embeddings into a Multi-layer Perceptron (MLP) of an arbitrary shape with an output size of 1 (one predicted mutation time for each mutation).

4.2.2 GNN. We also tested a simple graph neural network architecture (GNN), similar to Korfmann et al. (2023). The GNN leverages topological information, as well as edge features similarly to tsNN. The GNN performs graph convolutions (passing information between related nodes) to obtain a node embedding. However, most spatial (along chromosome) information is lost. These node embeddings can then be decoded in the same way as described above for tsNN.

4.2.3 Training and validation simulations. We used stdpopsim to simulate the ARG of a sample (100 diploid individuals) under the three

population Out-of-Africa demographic model (OutOfAfrica_3G09), with constant recombination rate of 1.3×10^{-8} and mutation rate of 2×10^{-8}. To avoid leaking of information from the true simulated ARGs, we reordered the nodes by mean number of descendants (as opposed to true node times) with the constraint that parent nodes are older the respective child nodes. The only features, beyond the topologies, that the neural networks saw were edge features (span and length). Span was scaled by the mutation rate and edge length was computed as the number of mutations on an edge.

We simulated 900 replicate ARGs under this model for training and 100 for validation (note no hyper-parameter tuning was performed). To compare tsNN and GNN, we simulated 10Kb genomes. However, we also trained tsNN on 1Mb genomes to understand the relationship between sequence length and accuracy. We trained both models with mean squared error loss on the $\log_{10}$ transformed mutation times.

4.3 Results

We devised a test to better understand whether a neural network can extract relevant evolutionary information from ARGs. The process of ARG inference is usually split into two steps: (i) local tree topology estimation, and (ii) inference of branch lengths and node times is performed using a "molecular clock" (after mapping mutations onto trees). The second step is straightforward, but strong coalescent priors are used and dating is heavily affected by issues in the data. Thus, we reasoned that estimating times from tree sequences is a simple, yet fundamental task in ARG inference. The choice of estimating mutation times, as opposed to node times, is due to the fact that true nodes are never known in practice, as the only actual data we can observe are mutations. Importantly,

we sought to estimate mutation times by presenting the neural networks with a modified true ARG, in which nodes are reordered based on the mean number of descendants. Further, the network only sees span (scaled by the mutation rate) and edge lengths (as the number of mutations in an edge).

4.3.1 Comparing tsNN and GNN. The fundamental difference between tsNN and GNN is the fact that tsNN induces a particular ordering to updating node embeddings, that we hypothesize is important for extracting evolutionary information. The node embeddings are updated with the edge insertion oder along a tree sequence (Figure 4.2). Because younger haplotypes persist far longer than old haplotypes, a young haplotypes will be involved in many more updates than an old one.

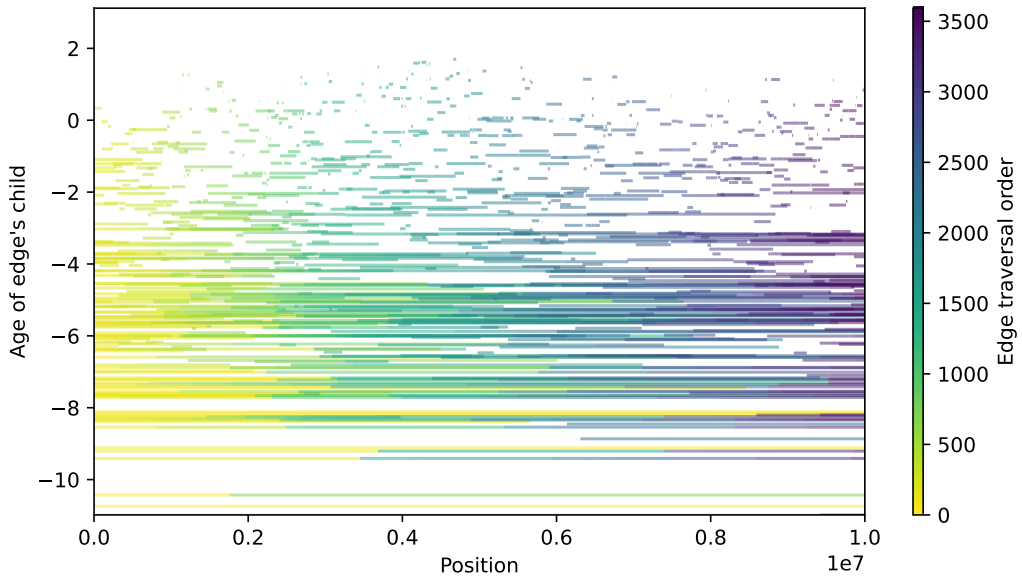
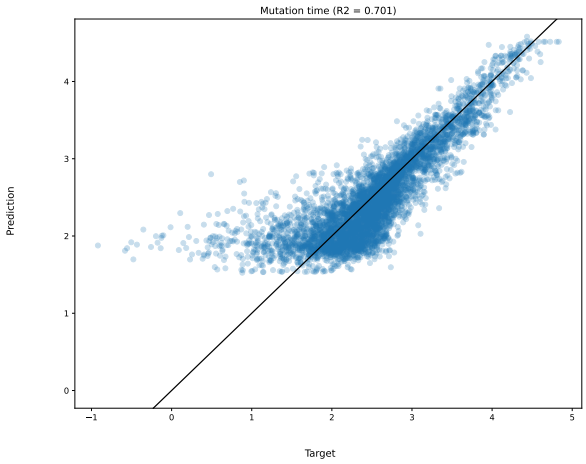
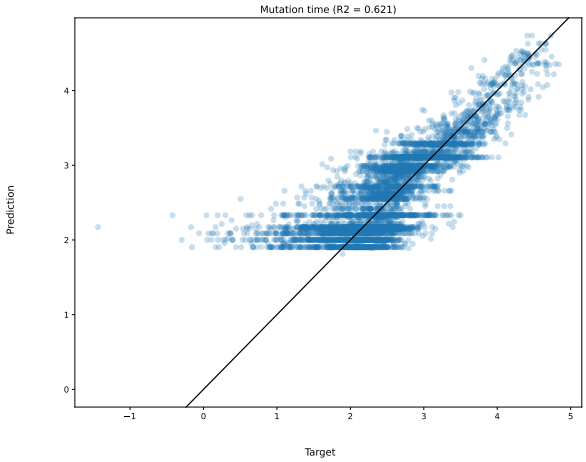


Figure 4.2. Edges in an arbitrary tree sequence, with the edge traversal order in tsNN represented by a color gradient. Edges are represented as horizontal lines, where the y-axis denotes the age of an edge's child and x-axis denotes the left and right positions along the chromosome.

We found that tsNN outperforms the GNN in inferring mutation times (Figure 4.4), even though the GNN architecture has about 6 times more parameters (560,000 trainable parameters in the GNN versus 90,000 parameters in the tsNN). Current dating algorithms can achieve an R^2 of around 0.9 (Wohns et al., 2022) using the true ARG. Our model trained on 10Kb sequences fail to properly estimate the time of mutations younger than $\log_{10}^2$ generations, greatly affecting overall accuracy.



(a) tsNN



(b) GNN

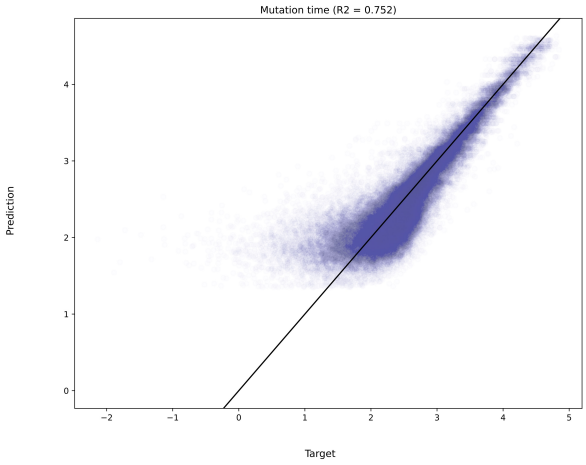
Figure 4.3. Accuracy of tsNN and GNN in predicting mutation times. Scatterplots show predicted against target mutation times for 100 tree sequences. Times are $\log_{10}$ transformed.

4.3.2 Improving inference of mutation times.

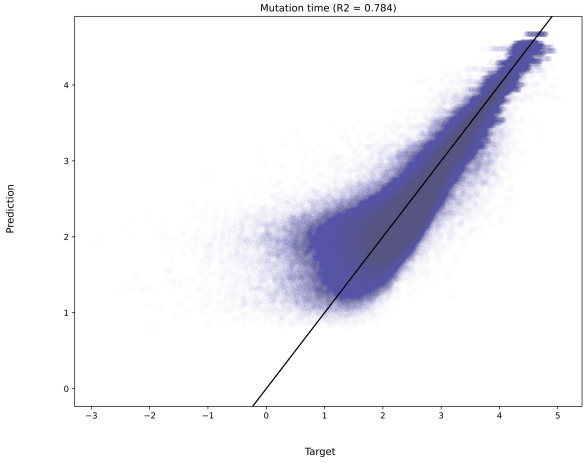
The failure to correctly infer the ages of young mutations could be caused by two factors: (i) the training dataset does not contain enough young mutations, and (ii) the model needs larger sequences (more than 10Kb) to better learn the recombination clock (i.e., the spans of young haplotypes will not be artificially truncated by 10Kb). To test what could be causing this issue, we changed the training data either by scaling the mutation rate or by increasing the simulated chromosome. We found that accuracy increases as mutation rates increases (Figures 4.4a and 4.4b), suggesting that the number of young mutations does limit learning. However, increasing the chromosome length increases accuracy even more (Figure 4.4c). Thus, it seems the network is learning the recombination clock and using it to improve prediction of mutation times.

4.4 Discussion

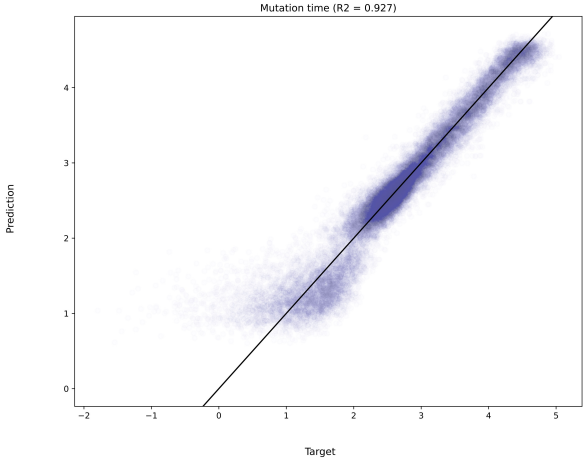
Ancestral Recombination Graphs (ARGs) can solve many issues currently plaguing evolutionary inference, because it fully encodes evolutionary outcomes in a compact manner. Treating the ARG as a latent parameter to be averaged out is not easy, because the state space is overwhelmingly large (Griffiths & Marjoram, 1996; Nielsen, 2000). More tractable approximations, such as the Sequentially Markovian coalescent (SMC), have been widely applied in evolutionary inference (McVean & Cardin, 2005; Schraiber & Akey, 2015), but the simplifying assumptions of these models limit their utility. An alternative lies in shifting towards inferring a single plausible ARG (Kelleher et al., 2019; Speidel et al., 2019; Wong et al., 2023). There is a clear need for methods that leverage an inferred ARG for population genetic inference and to quantify potential gains in accuracy. Nevertheless, it is not obvious how to use the all the information encoded in ARGs



(a) 10Kb with mutation rate of 2×10^{-7}



(b) 10Kb with mutation rate of 2×10^{-6}



(c) 1Mb with mutation rate of 2×10^{-8}

Figure 4.4. The effect of number of mutations and chromosome length on accuracy. Scatterplots show predicted against target mutation times for 100 tree sequences. Times are $\log_{10}$ transformed.

for inference, and current studies either compute summary statistics based on ARGs or assume independency between marginal trees (Fan et al., 2023; Hejase et al., 2022).

Our main goal has been to develop a framework that can leverage most of the information encoded in ARGs for evolutionary inference. We show that our neural network, tsNN, can learn how to date mutations in an ARG, outperforming likelihood-based models (tsNN $R^2 = 0.927$, tsdate $R^2 = 0.902$) (Wohns et al., 2022). By comparing performance on different training sets, we demonstrate that the neural network learns to leverage both the mutation and recombination clocks to produce accurate estimates of mutation times.

Although ARGs carry sufficient information for inferring evolutionary events, ARG inference methods could significantly impact downstream applications. Indeed, the two most widely adopted ARG inference programs, tsinfer and Relate, tend to overestimate small coalescence times and underestimate large ones (Y C Brandt et al., 2022). In our mutation time inference task, we used the true ARGs for training and validation, but some of these biases can be alleviated in a supervised machine learning framework by training on inferred ARGs, as opposed to true ARGs.

A trickier issue might arise when the true generative process, which yielded the real data, does not match that of the training data. In our tests, we assumed that the training and validation data came from the exact model. In the future, tests where we slightly alter the simulation model for training will aid in understanding the robustness of tsNN to model mis-specification. Previous studies have shown that machine learning methods can be as robust to model mis-specification as their likelihood-based counterparts (Hejase et al., 2022). A

new approach to deal with the mis-specification issue, called Domain-Adaptive Networks, can mitigate mis-speficication by leveraging target-domain data in an unsupervised manner (Mo & Siepel, 2023).

An exciting avenue for future research lies in expanding the tasks that tsNN can perform, what can be accomplished by developing different decoders. For example, the rich information contained in the ARG could be used to distinguish between complex demographic models that are unidentifiable with allele frequency data (e.g., the site frequency spectrum) (Fan et al., 2023; Schraiber & Akey, 2015). The inference of selective sweeps has much to gain from leveraging ARGs, mostly because it allows us to bypass the specification of a window size. The span of ancestral haplotypes and edges naturally encode both the spatial and temporal scales over which processes impact variation, and an ARG-based method can implicitly leverage this information.

With the recent advances in scalable ARG inference methods, ARG-based evolutionary inference methods are still in its infancy. A powerful framework that is able to efficiently leverage all the rich information contained in ARGs is needed. The field of graph neural networks is ripe with ideas that can be applied to ARGs. Indeed, we demonstrate that our method, tsNN, represents an important step in this direction, but there still much work to be done in applying deep learning to gain new insights into past evolutionary events.

CHAPTER V

CONCLUSION

Understanding the balance of evolutionary forces shaping the origin and maintenance of genetic variation has been the core driver of population genetics (Lewontin, 1974). Our ability to collect data has exponentially increased over the last few decades, moving from allozyme gels of a few samples to whole-genome data of thousands of individuals. With this flood of data, we are poised to make huge progress on long standing evolutionary questions. However, the traditional framework for evolutionary inference has somewhat stalled new discoveries.

Many of the issues we are facing can be mitigated with evolutionary simulations. A great deal of progress has been made on evolutionary simulation tools (Adrion et al., 2020; Haller & Messer, 2019; Haller et al., 2019; Kelleher et al., 2016a). These advancements, coupled with huge increases in computational power, now allow us to use simulations to effectively infer previously intractable likelihoods, and to explore features of the data that are not easy to model mathematically. Indeed, over the past decade simulation-based inference has gained immense popularity (Caldas et al., 2022; Chan et al., 2018; Korfmann et al., 2023; Schrider & Kern, 2016; Torres et al., 2018).

As the questions and models increase in complexity, there is a growing need for standards in simulation models and for increase reproducibility. `stdpopsim` is a community-maintained library for previously published simulation models, which includes species-specific population genetic parameters (e.g., mutation rates, recombination maps) as well as demographic and selection models. In Chapter II, I presented my contributions to evolutionary simulation tools, which will help facilitate simulation-based inference in population genetics. Much more is needed

still in two fronts: (i) benchmarking our current tools under a common variety of evolutionary scenarios, and (ii) verifying for consistency of published models, for example by ensuring that the models can yield simulated data that actually resembles the real data.

It is now easier than ever to use evolutionary simulations to better understand complex models and features of the data for which there is no theory yet. Indeed, simulations have opened up new opportunities to better understand interactions between evolutionary processes. For example, Schrider (2020) used more realistic simulations to show that the footprints selective sweeps are not as easily confounded by background selection as previously thought (Andolfatto, 2001). In Chapter III, I leveraged complex and realistic simulations of the entire great apes history to learn about which processes have shaped genetic variation in the group. Without simulations, it would not be possible to jointly study the effects of positive and negative selection, mutation rate variation and GC-biased gene conversion on large-scale patterns of genomic variation.

Another factor that is bound to improve evolutionary inference is the shift towards Ancestral Recombination Graphs (ARGs). This data structure is compact, and so it can enable inference over larger scales, both in number of samples as well as in number of sites. ARGs are now used as a backbone to many different evolutionary simulation tools, allowing for a better integration (Haller et al., 2019; Kelleher et al., 2016b). The field has moved away from treating the ARG as a latent parameter to be averaged out (Griffiths & Marjoram, 1996; Nielsen, 2000) to inferring a single plausible ARG (Kelleher et al., 2019; Speidel et al., 2019). The state space for possible ARGs is overwhelmingly large, so this shift allows us to leverage ARGs at scale. However, it is still unclear how to actually use all the

information encoded in ARGs for inference. Indeed, many recent studies instead either develop topological summary statistics or treat marginal trees in the ARG as independent (Fan et al., 2023; Hejase et al., 2022). In Chapter IV, I proposed a new neural network framework that can make better use of ARGs for inference, but there is much to be done in this space still.

Computer simulations have drastically altered the field of evolutionary genetics in the past decade. One of the major downside of simulations is the computational cost. The parameter space for any moderately complex model can quickly become prohibitively large. No major leaps in the efficiency of simulators or in hardware are in the horizon, so likelihood-based models will be useful in complementing simulation studies. For example, it is often useful to constrain the parameter space using estimates from likelihood-based models. Simulation-based inference can produce biased estimates when the generative process in simulations does not match that of the actual data, an issue also known as model mis-specification. Some of this can be alleviated by incorporating error, such as genotyping error and missingness, to the idealized simulations. However, there is still no consensus on how to include such errors to simulations in a systematic way. There are tools for dealing with mis-specification in the broader machine learning literature, and these seem promising in evolutionary contexts as well (Mo & Siepel, 2023). I expect that our ability to leverage simulations to gain evolutionary insights will only increase with improvement in these areas.

APPENDIX

SUPPLEMENTAL MATERIAL FOR CHAPTER III

Supplementary material

A.0.1 Correlation between divergences that share branches.

Landscapes of divergence can be correlated by their definition, as they can share part of their histories. In most of our analyses (except for Figure A.2), we do not show the correlations for such cases but below we describe how this sharing would affect correlations (using a simplified theory). For example, in Figure 3.3 d_{VX} and d_{XY} share the branch X; depending on how the length of the branch X compares to the total tree length, these two landscapes are bound to be correlated. Assuming that mutations follow a Poisson process and that coalescences happen instantaneously, we derive the following. There are three non-overlapping parts in the tree between these, the branch from the XY ancestor to X with length $\mathrm{E}[\tau_X] = T_{XY}$, the branch from the XY ancestor to Y with length $\mathrm{E}[\tau_Y] = T_{XY}$ and the branch from V to the XY ancestor with length $\mathrm{E}[\tau_V] = 2T_{VWXY} - T_{XY}$. If we just consider the genealogical definition of divergence and assume $d_{VX} = \tau_V + \tau_X$ and $d_{XY} = \tau_X + \tau_Y$ (i.e., ignoring the contributions of ancestral diversity to divergence), then

$$\begin{aligned}
\mathrm{Cov}[d_{VX}, d_{XY}] &= \mathrm{Cov}[\tau_X + \tau_V, \tau_X + \tau_Y] \\
&= \mathrm{Cov}[\tau_X, \tau_X] + \cancelto{0}{\mathrm{Cov}[\tau_X, \tau_Y]} + \cancelto{0}{\mathrm{Cov}[\tau_V, \tau_X]} + \cancelto{0}{\mathrm{Cov}[\tau_V, \tau_Y]} \\
&= \mathrm{Var}(\tau_X) = \mathrm{E}[\tau_X] = T_X
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\mathrm{Cor}[d_{VX}, d_{XY}] &= \frac{\mathrm{Cov}[\tau_X + \tau_V, \tau_X + \tau_Y]}{\sqrt{\mathrm{Var}[\tau_X + \tau_V]\,\mathrm{Var}[\tau_X + \tau_Y]}} \\
&= \sqrt{\frac{\mathrm{Var}[\tau_X]^2}{(\mathrm{Var}[\tau_X] + \mathrm{Var}[\tau_V])(\mathrm{Var}[\tau_X] + \mathrm{Var}[\tau_Y])}} \\
&= \sqrt{\frac{\mathrm{Var}[\tau_X]}{\mathrm{Var}[\tau_X] + \mathrm{Var}[\tau_V]} \frac{\mathrm{Var}[\tau_X]}{\mathrm{Var}[\tau_X] + \mathrm{Var}[\tau_Y]}} \\
&= \sqrt{\frac{T_X}{T_X + T_V} \frac{T_X}{T_X + T_Y}} \\
&= \sqrt{p_{d_{VX}} p_{d_{XY}}}
\end{aligned}
$$

where $p_{d_{VX}} = \frac{T_X}{T_X + T_V}$ is the proportion of d_{VX} that is shared with d_{XY}, and $p_{d_{XY}} = \frac{T_X}{T_X + T_Y}$ is the proportion of d_{XY} that is shared with d_{VX}.

μ_N	μ_P	$\bar{s}_N$	$\bar{s}_P$	Regime	μ_{SD}
0	0	0	0	Neutral	0
0	0	0	0	Variable μ	0.005
0	0	0	0	Variable μ	0.007
0	0	0	0	Variable μ	0.011
0	0	0	0	Variable μ	0.016
0	0	0	0	Variable μ	0.023
0	0	0	0	Variable μ	0.033
0	0	0	0	Variable μ	0.048
0	0	0	0	Variable μ	0.070
0	0	0	0	Variable μ	0.103
0	0	0	0	Variable μ	0.150
0	1×10^{-12}	0	1×10^{-2}	Beneficial	0
0	1×10^{-11}	0	1×10^{-2}	Beneficial	0
2×10^{-9}	0	-3×10^{-2}	0	Deleterious	0
2×10^{-9}	1×10^{-11}	-3×10^{-2}	1×10^{-2}	Both	0
2×10^{-9}	0	-1.5×10^{-2}	0	Deleterious	0
2×10^{-9}	1×10^{-11}	-1.5×10^{-2}	1×10^{-2}	Both	0
2×10^{-9}	0	-1×10^{-2}	0	Deleterious	0
2×10^{-9}	1×10^{-12}	-1×10^{-2}	5×10^{-3}	Both	0
2×10^{-9}	1×10^{-12}	-1×10^{-2}	1×10^{-2}	Both	0
2×10^{-9}	0	-3×10^{-3}	0	Deleterious	0
2×10^{-9}	1×10^{-12}	-3×10^{-3}	5×10^{-3}	Both	0
2×10^{-9}	1×10^{-12}	-3×10^{-3}	1×10^{-2}	Both	0
2×10^{-9}	0	-1×10^{-3}	0	Deleterious	0
2×10^{-9}	1×10^{-11}	-1×10^{-3}	1×10^{-2}	Both	0
6×10^{-9}	0	-3×10^{-2}	0	Deleterious	0
6×10^{-9}	1×10^{-11}	-3×10^{-2}	1×10^{-2}	Both	0
6×10^{-9}	0	-1.5×10^{-2}	0	Deleterious	0
6×10^{-9}	1×10^{-11}	-1.5×10^{-2}	1×10^{-2}	Both	0
1×10^{-8}	1×10^{-11}	-1×10^{-3}	1×10^{-2}	Both	0
1.2×10^{-8}	0	-3×10^{-2}	0	Deleterious	0
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.005
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.007
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.011
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.016
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.023
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.033
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.048
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.070
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.103
1.2×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0.150
1.2×10^{-8}	1×10^{-11}	-3×10^{-2}	1×10^{-2}	Both	0
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.005
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.007
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.011
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.016
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.023
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.033
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.048
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.070
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.103
1.4×10^{-8}	0	-3×10^{-2}	0	Deleterious	0.150
1.4×10^{-8}	1×10^{-12}	-3×10^{-2}	1×10^{-2}	Both	0
1.4×10^{-8}	1×10^{-11}	-3×10^{-2}	1×10^{-2}	Both	0

Table A.1. Parameter space explored with simulations. μ_N and μ_P are the rates of mutations under negative and positive selection, respectively. $\bar{s}_N$ and $\bar{s}_P$ and the mean fitness effects of negatively and positively selected mutations. μ_{SD} is the scaled standard deviation of the mutation rate map. See Table 3.1 and subsection 3.2.2 for more details.

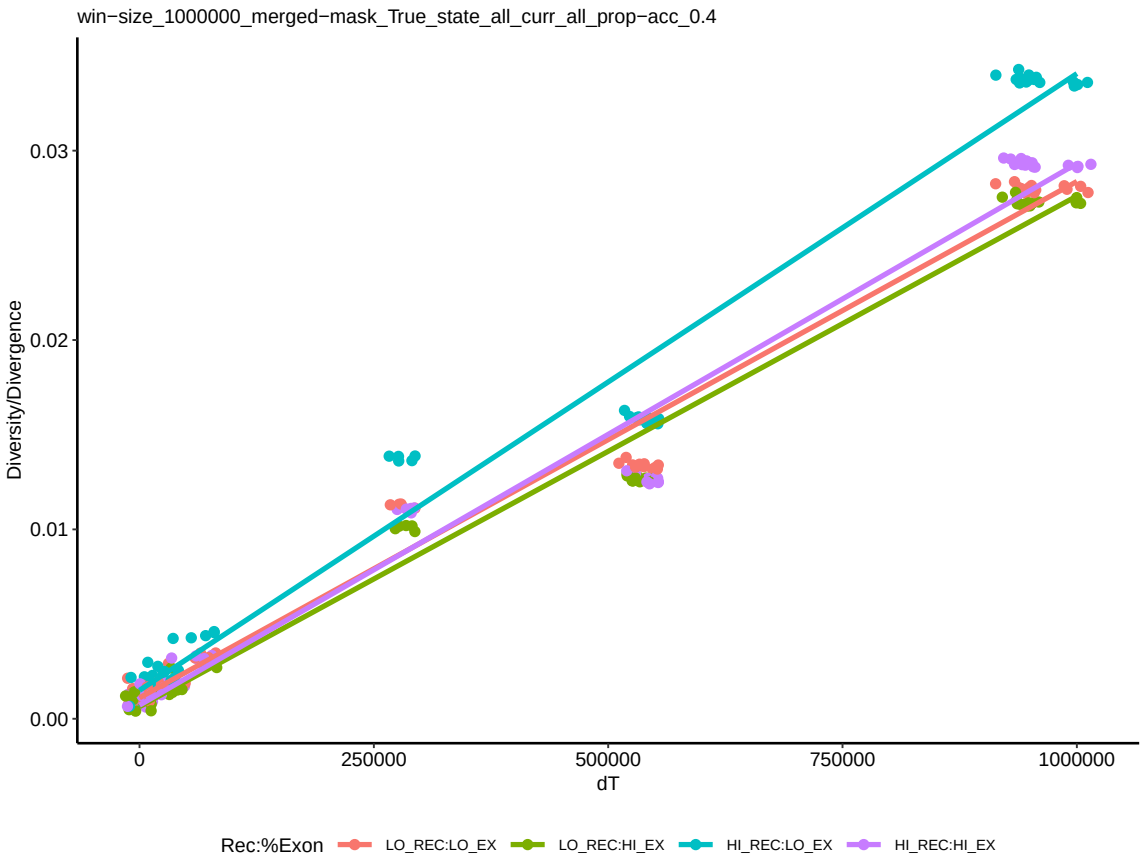


Figure A.1. Effect of exon density and recombination rate on the accumulation of genetic divergence in chromosome 12 with phylogenetic distance. Within-species genetic diversities are shown at $dT = 0$. Mean diversity and divergences were computed for four groups depending on whether they fell or not on the top 90% percentile of recombination rate and exon density.

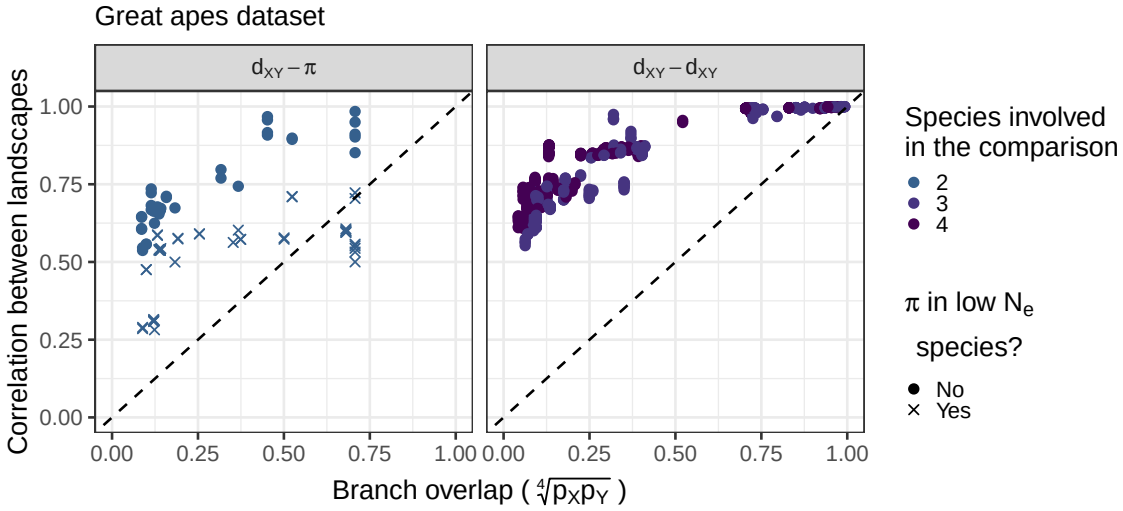


Figure A.2. Correlations between landscapes of diversity and divergence for comparisons with branch overlap. For example, diversity in humans and divergence between humans and bonobos share part of their history. Each point on the plots correspond to the (Spearman) correlation between two landscapes of diversity/divergence, computed on 1Mb windows across the entire genome. Correlations were split by type of landscapes compared ($\pi - d_{XY}$, $d_{XY} - d_{XY}$). The x-axis is a metric of expected branch overlap between the landscapes. See subsection A.0.1 for more information. Note that species with low N_e (bonobos, eastern gorillas and western chimps) have a different point shape. The colors reflect the number of species involved in the comparison. For example, the comparison between human-western gorilla and eastern chimp-Sumatran orangutan divergences includes four different species. On the other hand, the comparison between human-western gorilla and human-Sumatran orangutan divergences includes just three species.

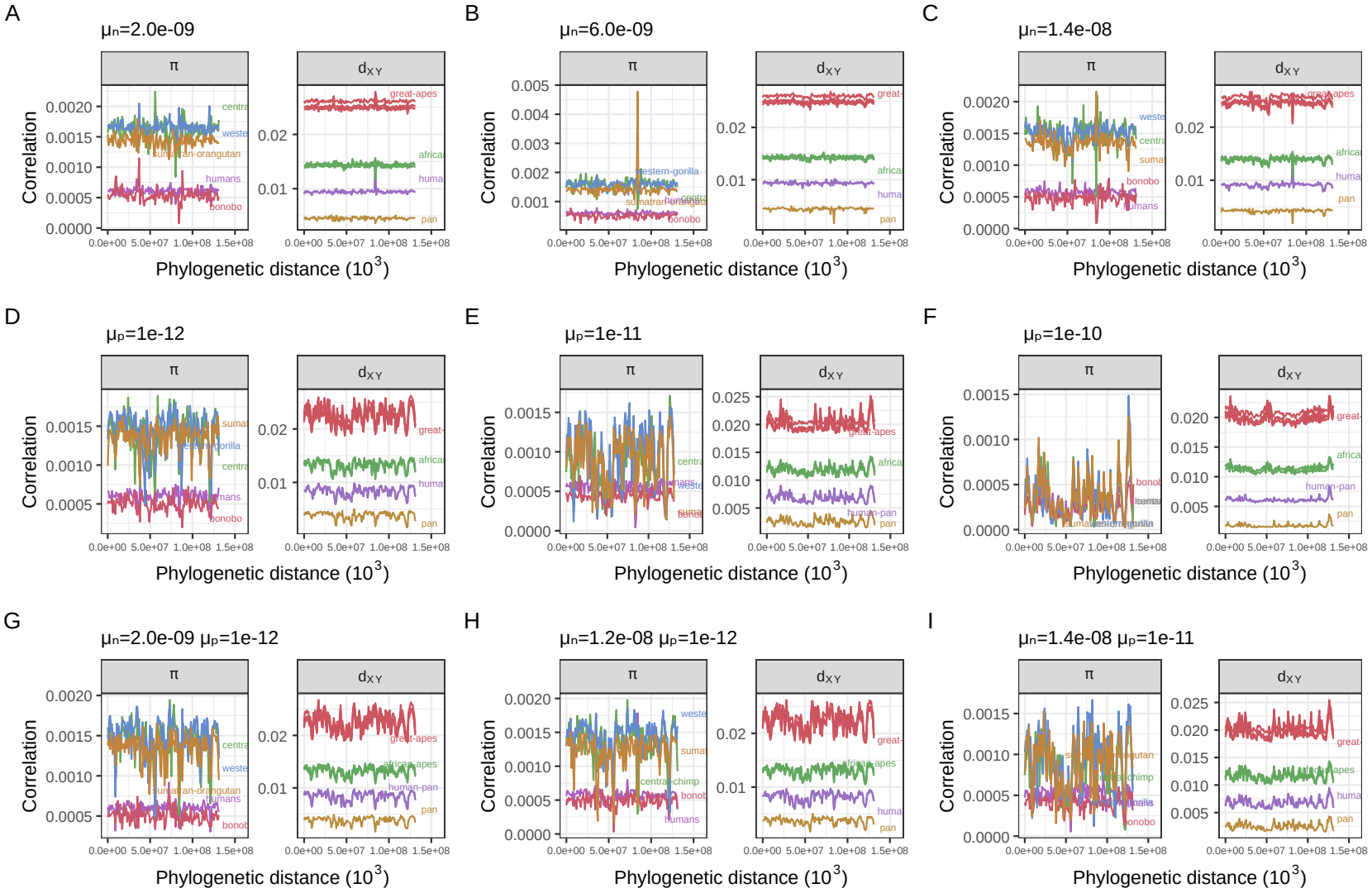


Figure A.3. Landscapes of diversity and divergence in selected simulations with natural selection. The selection parameters μ_n and μ_p are the rate of mutations in exons with negative and positive fitness effects, respectively. The mean fitness effect was $\bar{s} = -0.03$ for deleterious mutations and $\bar{s} = 0.01$ for beneficial mutations (see subsection 3.2.2 for more details). Other details are as in Figure 3.2.

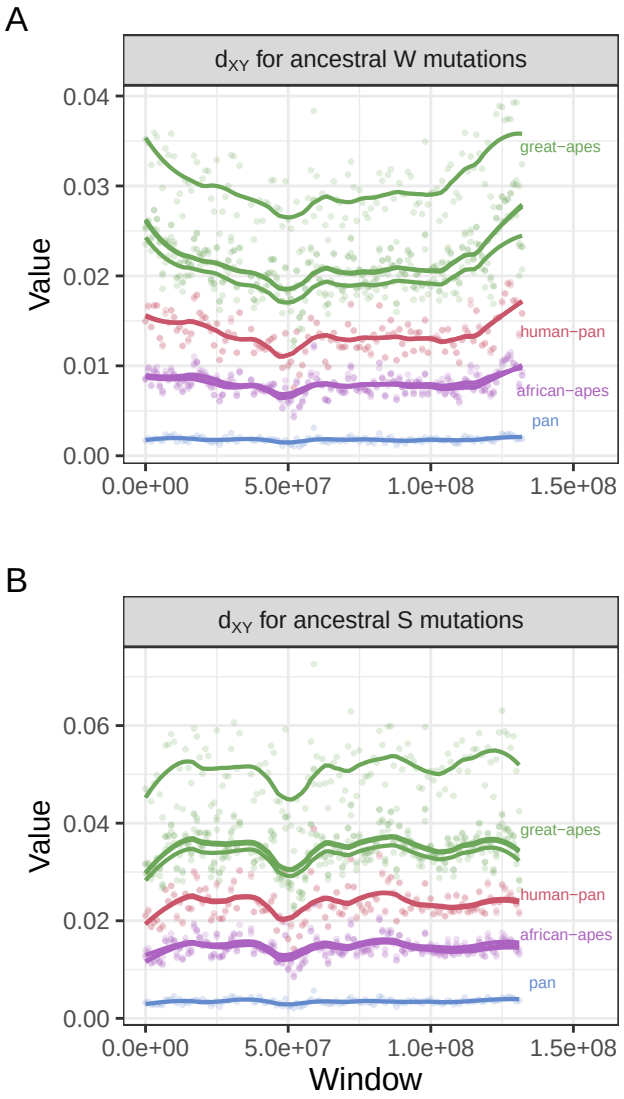


Figure A.4. Landscapes of divergence partitioned by allele state in the ancestor. Ancestral states were assumed to be the same as seen in rhesus macaques (RheMac2), and sites not called in macaques were not used. d_{XY} for W sites is simply the mean pairwise differences between samples in species X and Y per ancestral W sites (A/T). Similar reasoning applies for d_{XY} for S ancestral sites, but only considering (G/C) sites. Points were colored by the most common recent ancestor of the two species compared in each divergence. Lines were fitted using local linear regression. Note that for ancestrally weak mutations (A) there is an increase in divergence at the ends of the chromosomes, but that is not seen for ancestrally strong mutations (B).

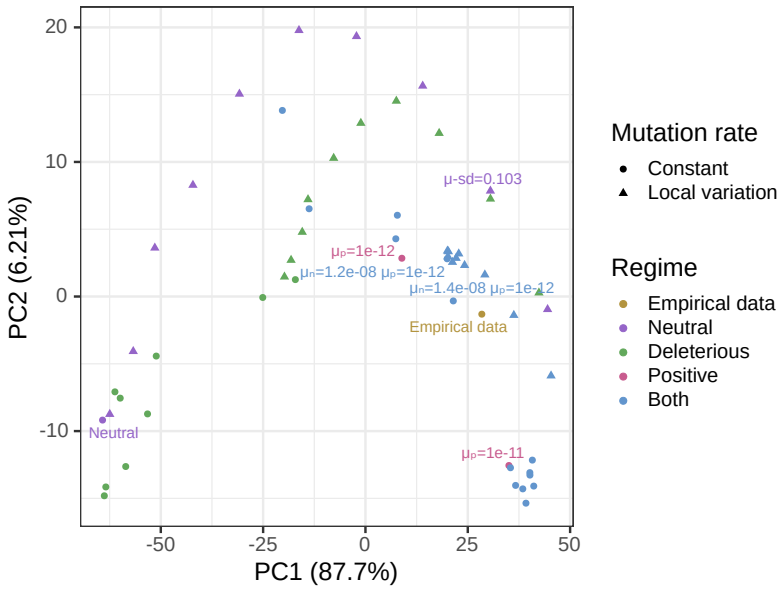


Figure A.5. PCA visualization of data and simulations at 500Kb. The colors differentiate the empirical data from simulations with different parameters: Neutral refers to the simulation without any selection, Deleterious refers to simulations with deleterious mutations, Positive refers to simulations with beneficial mutations, Both refers to simulations with both beneficial and deleterious mutations. The shape of the points differentiate simulations with constant mutation rate along the genome and variable local mutation rates. Principal component analysis (PCA) applied to a matrix with all pairwise correlations between landscapes across the great apes (including $\pi - \pi$, $\pi - d_{XY}$ and $d_{XY} - d_{XY}$ comparisons) for the great apes dataset and simulations (with selection and with mutation rate variation). We excluded simulations with $\mu_p \geq 1 \times 10^{-10}$ from the PCA analysis because PC2 was capturing negative correlations caused by strong positive selection — as seen in Figure 3.7F.

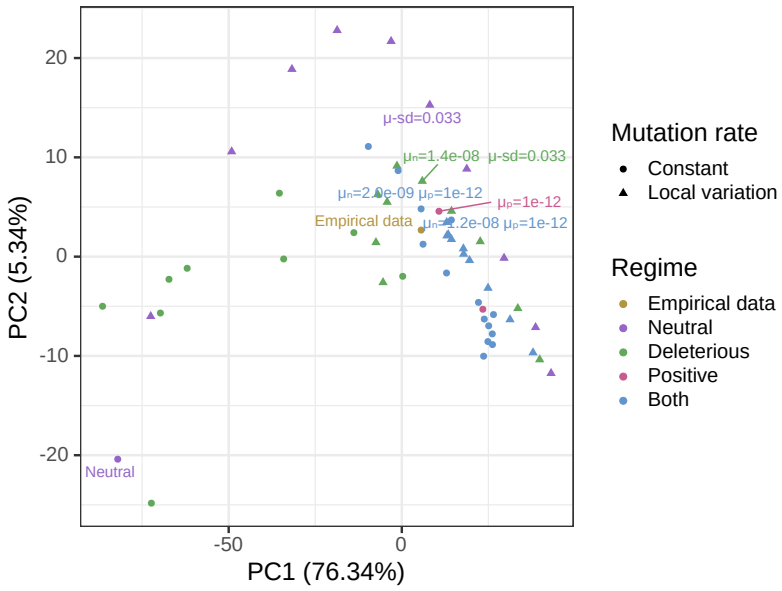


Figure A.6. PCA visualization of data and simulations at 5Mb. The colors differentiate the empirical data from simulations with different parameters: Neutral refers to the simulation without any selection, Deleterious refers to simulations with deleterious mutations, Positive refers to simulations with beneficial mutations, Both refers to simulations with both beneficial and deleterious mutations. The shape of the points differentiate simulations with constant mutation rate along the genome and variable local mutation rates. Principal component analysis (PCA) applied to a matrix with all pairwise correlations between landscapes across the great apes (including $\pi - \pi$, $\pi - d_{XY}$ and $d_{XY} - d_{XY}$ comparisons) for the great apes dataset and simulations (with selection and with mutation rate variation). We excluded simulations with $\mu_p \geq 1 \times 10^{-10}$ from the PCA analysis because PC2 was capturing negative correlations caused by strong positive selection — as seen in Figure 3.7F.

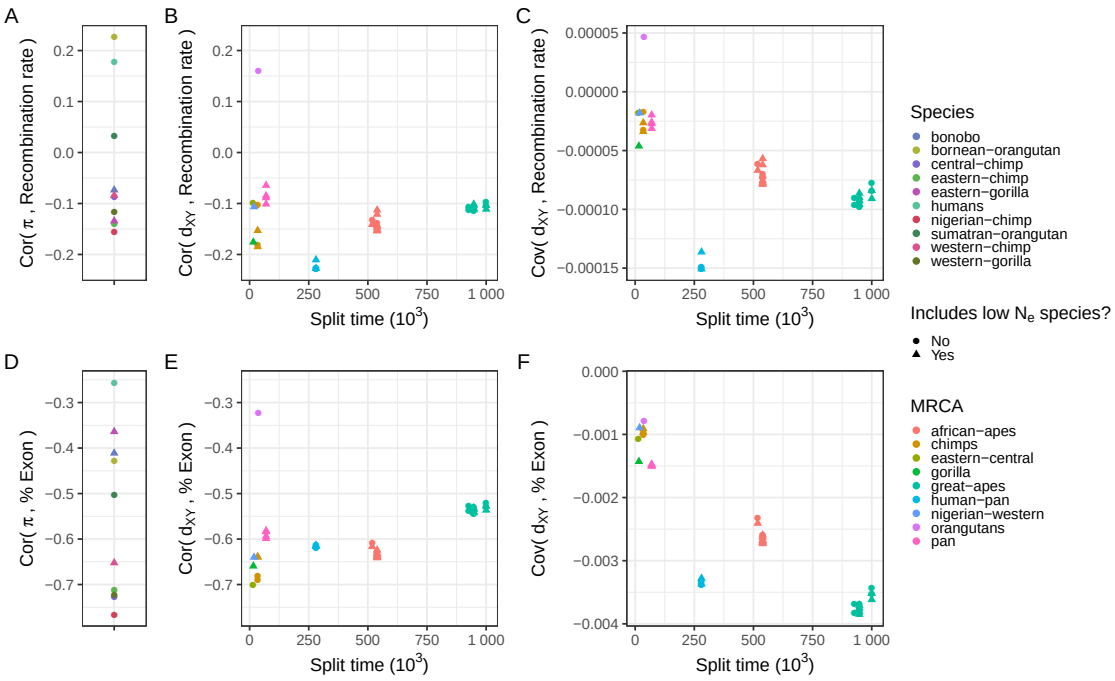


Figure A.7. Correlations and covariances between landscapes of diversity and divergence and annotation features in the real great apes data. Only windows in the middle half of chromosome 12 were included. Compare to Figure 3.10.

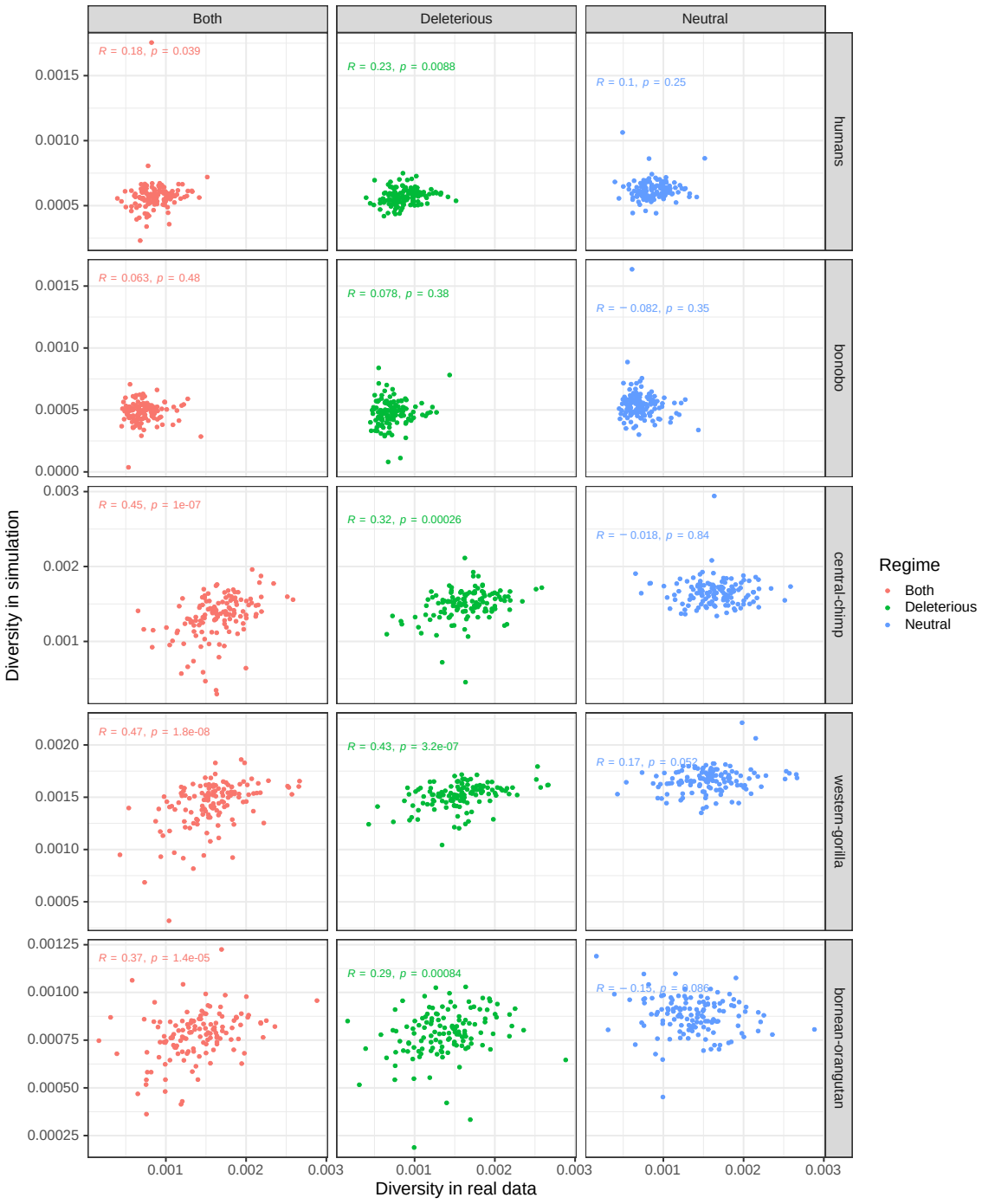


Figure A.8. Scatterplots of genetic diversity in the real great apes data against diversity seen in simulations. Simulation with deleterious mutations had $\mu_n = \exp 1.4{-}8$, and the simulation with both deleterious and beneficial mutations had $\mu_n = \exp 1.4{-}8$ and $\mu_p = \exp 1{-}12$.

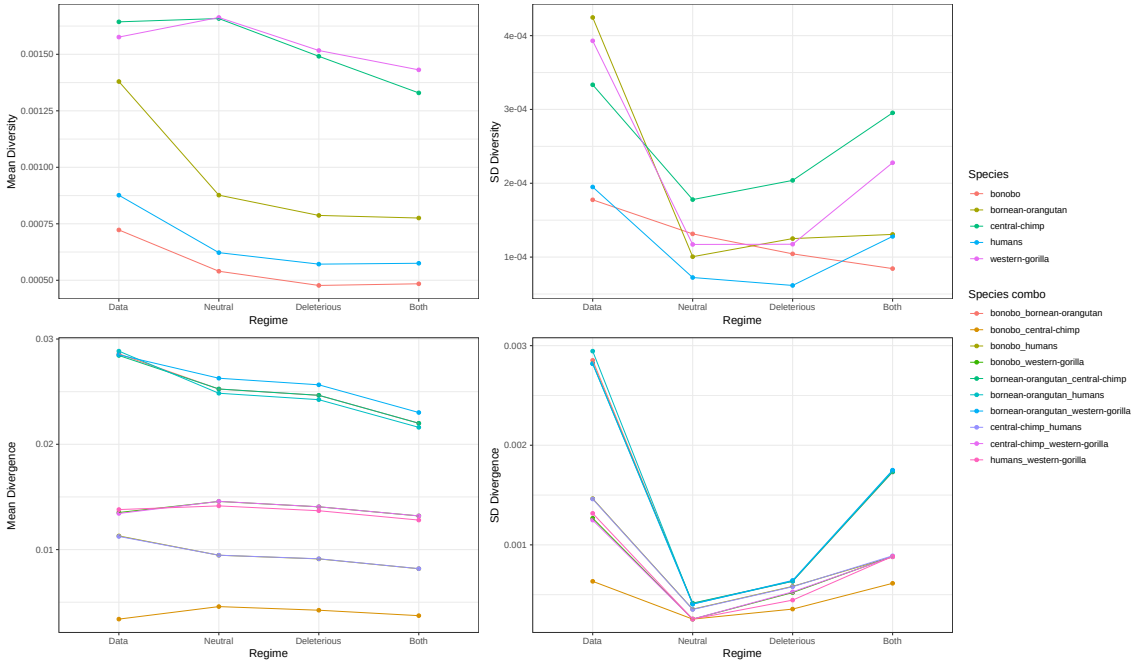


Figure A.9. Summaries of genetic diversity (top) and divergence (bottom) across all species for the data and simulations. Mean is shown on the left and standard deviation on the right. "Neutral" refers to the simulation without any selection, "Deleterious" refers to the simulation with deleterious mutations ocurring at a rate of 1.4×10^{-8}, "Both" refers to the simulation with both beneficial and deleterious mutations, with rates 1×10^{-12} and 1.4×10^{-8} respectively.

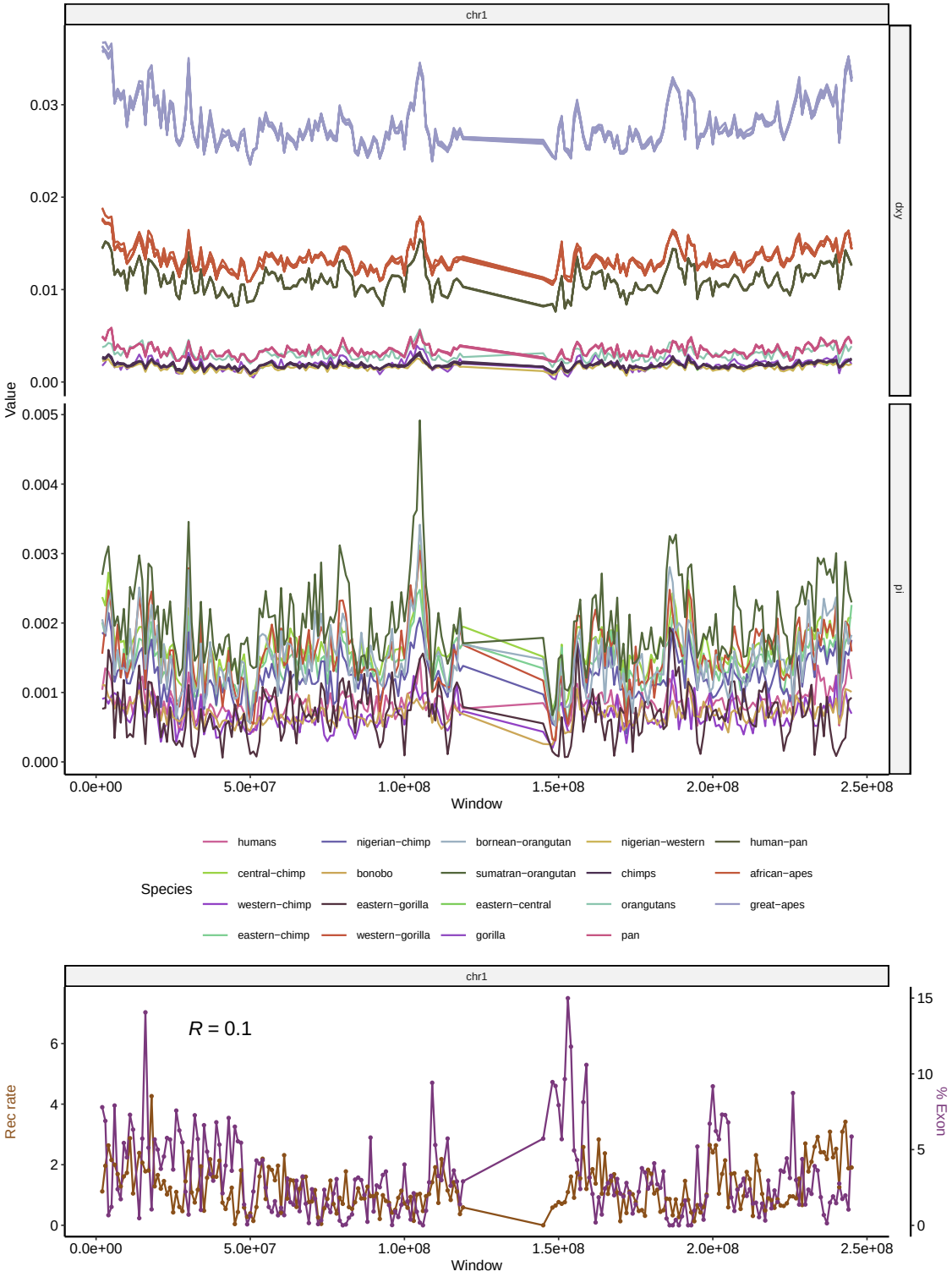


Figure A.10. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 1. See Figure 3.2 for more details.

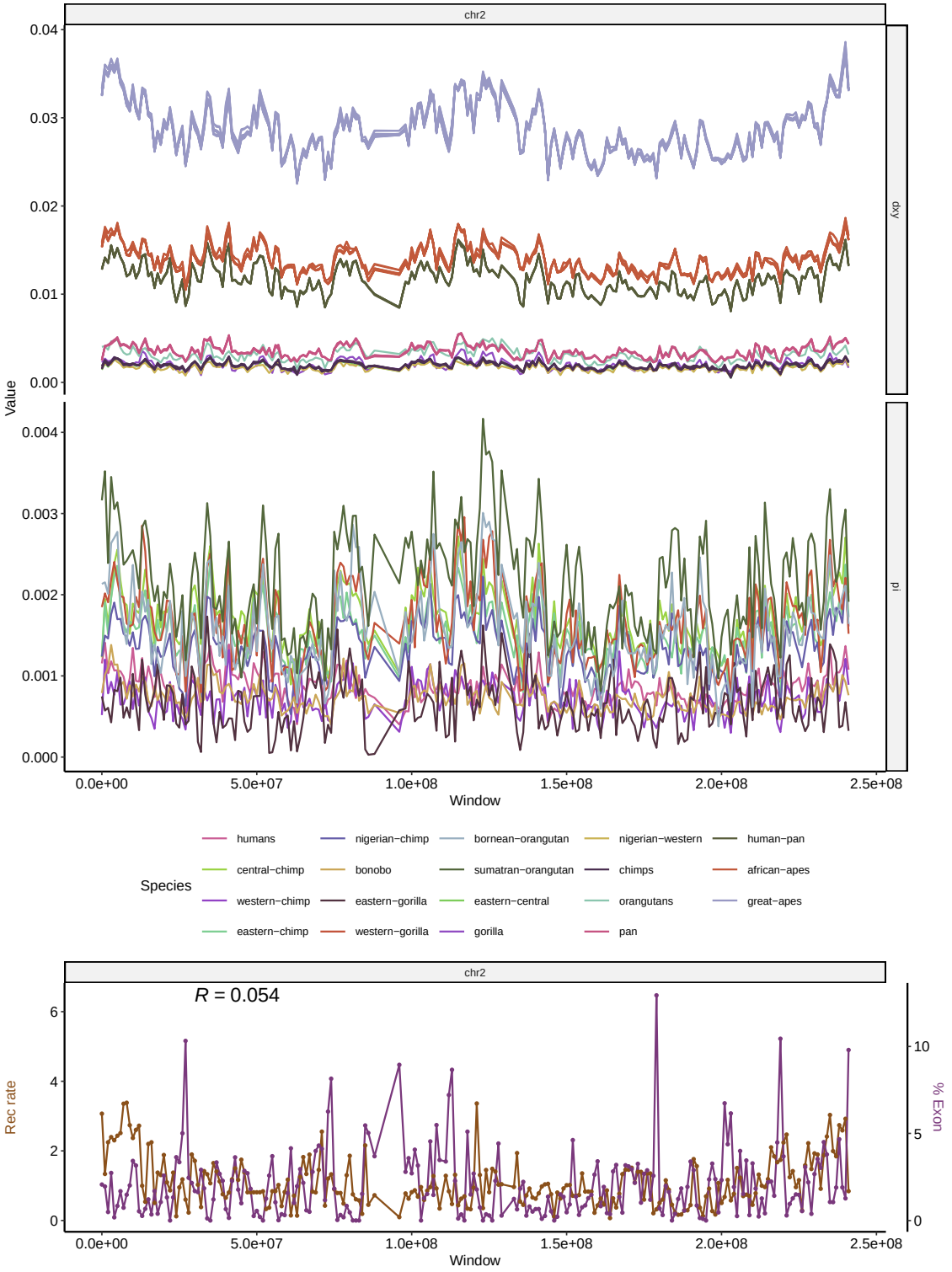


Figure A.11. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 2. See Figure 3.2 for more details.

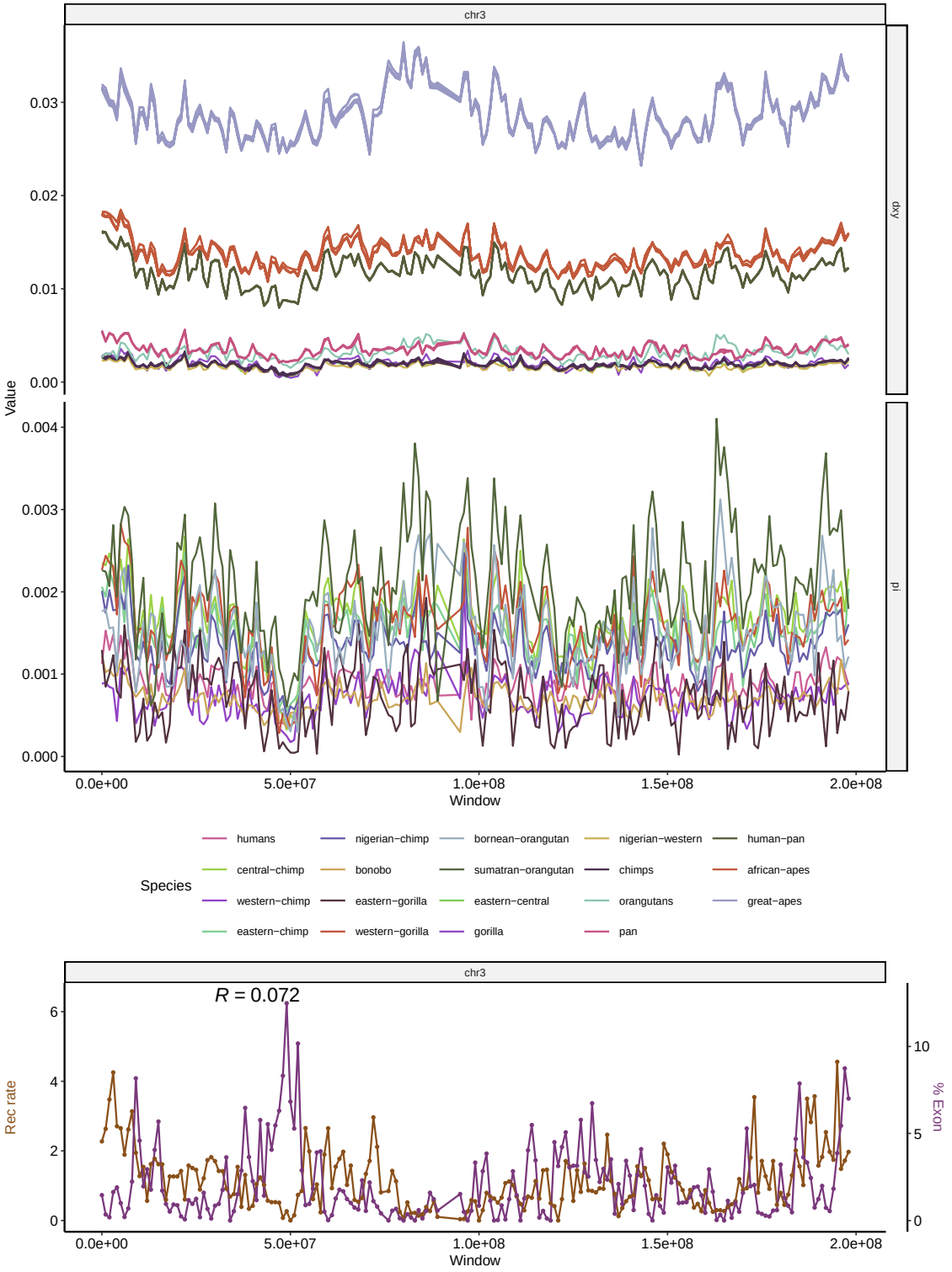


Figure A.12. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 3. See Figure 3.2 for more details.

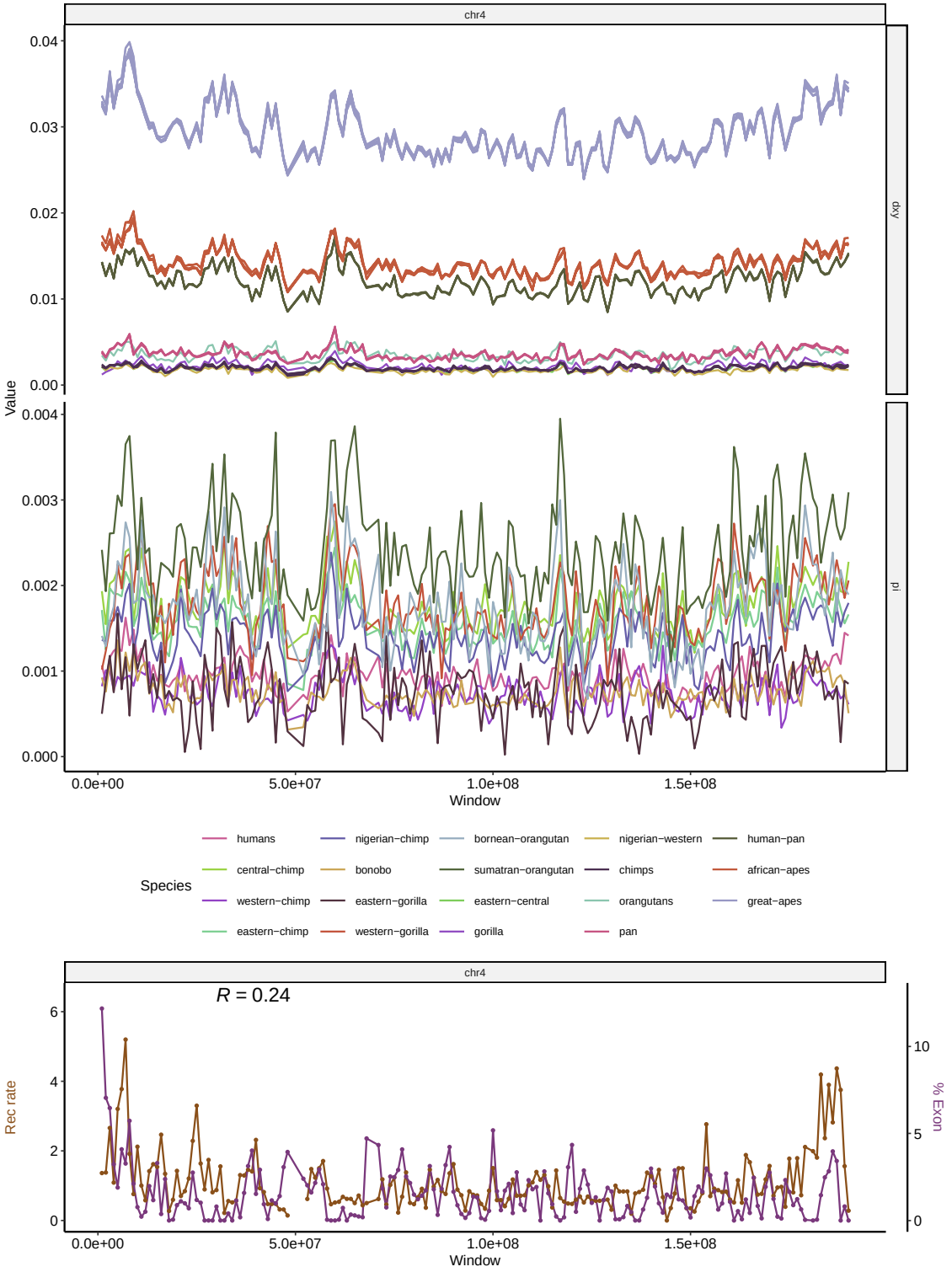


Figure A.13. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 4. See Figure 3.2 for more details.

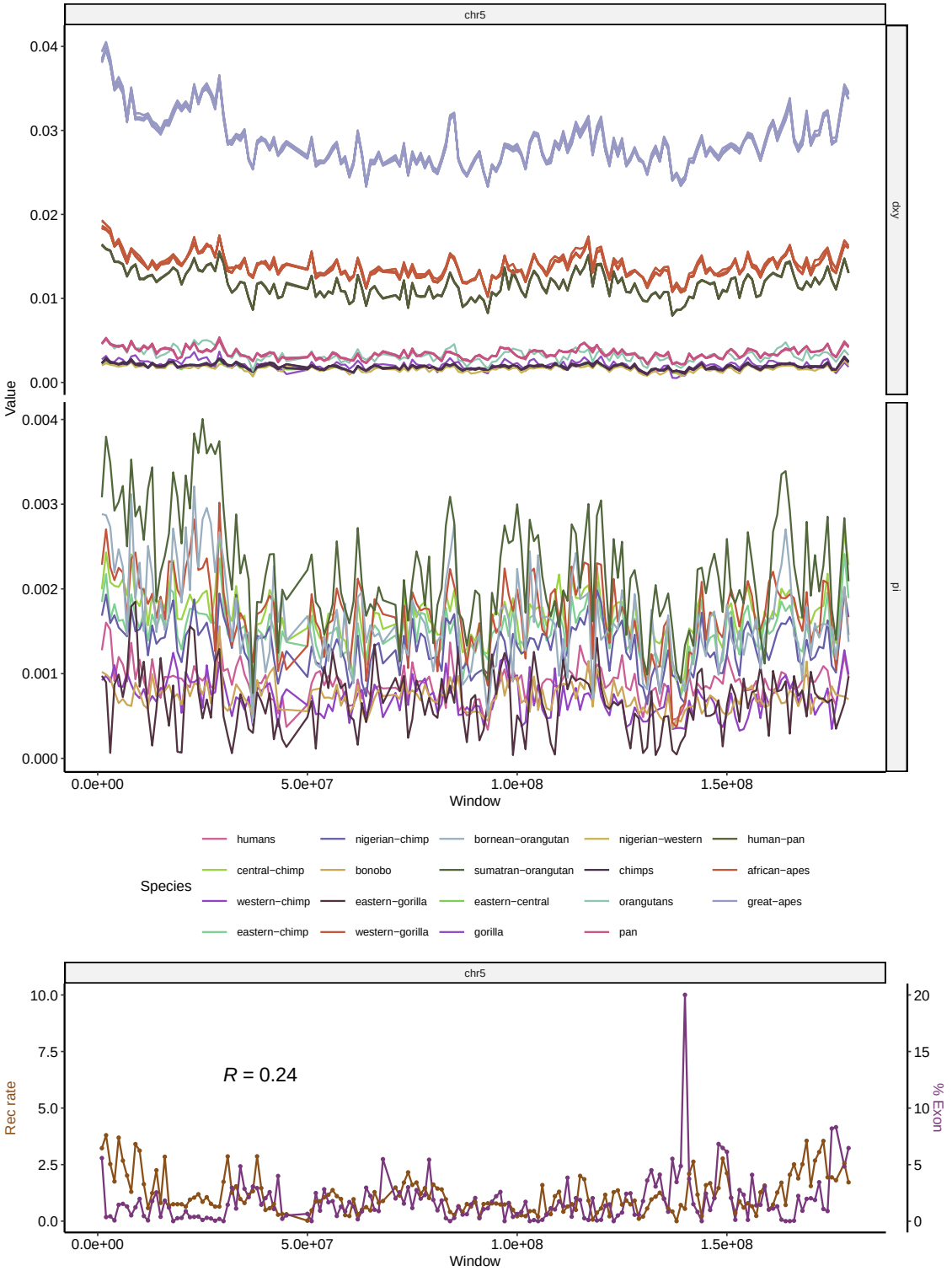


Figure A.14. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 5. See Figure 3.2 for more details.

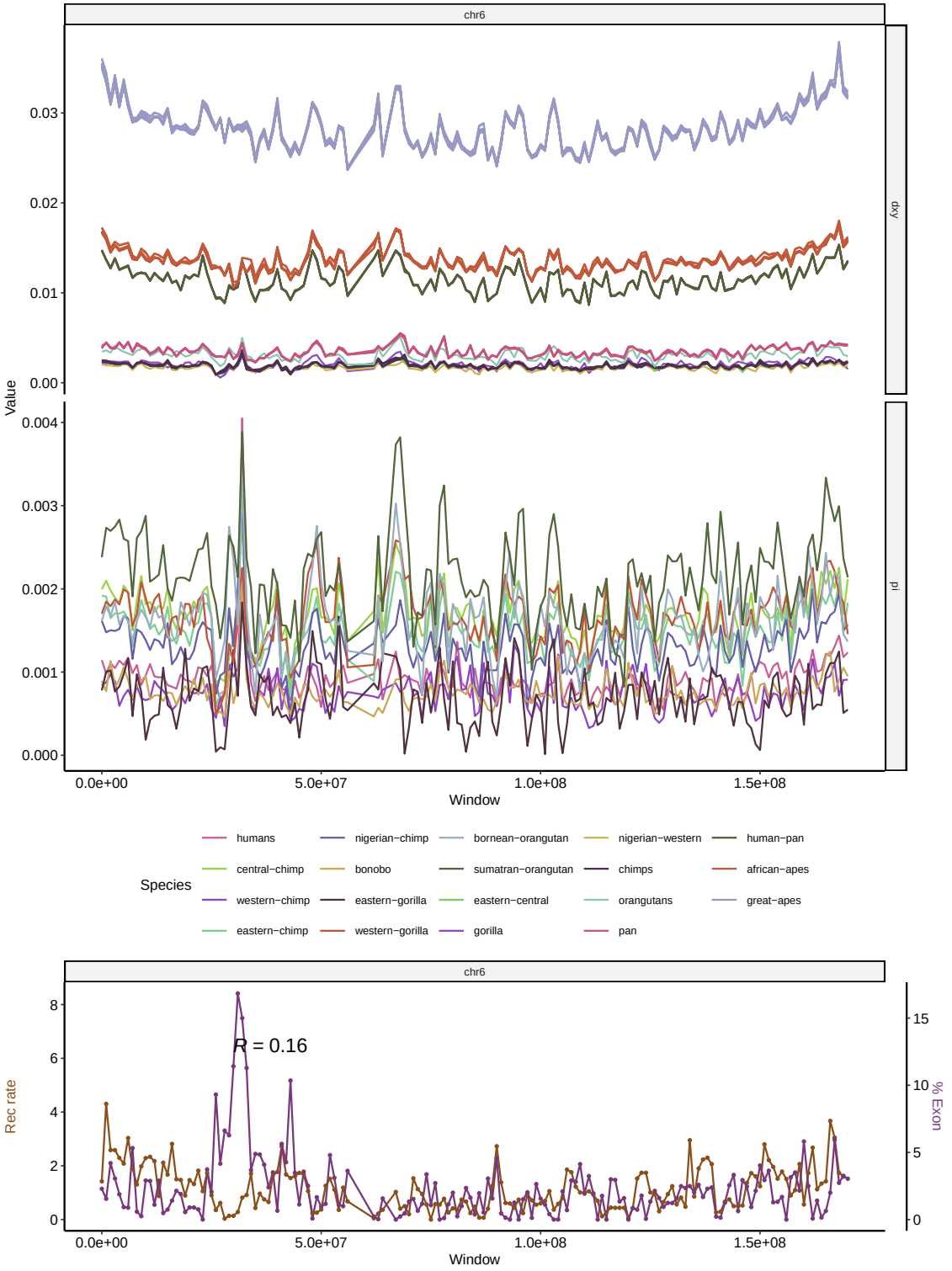


Figure A.15. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 6. See Figure 3.2 for more details.

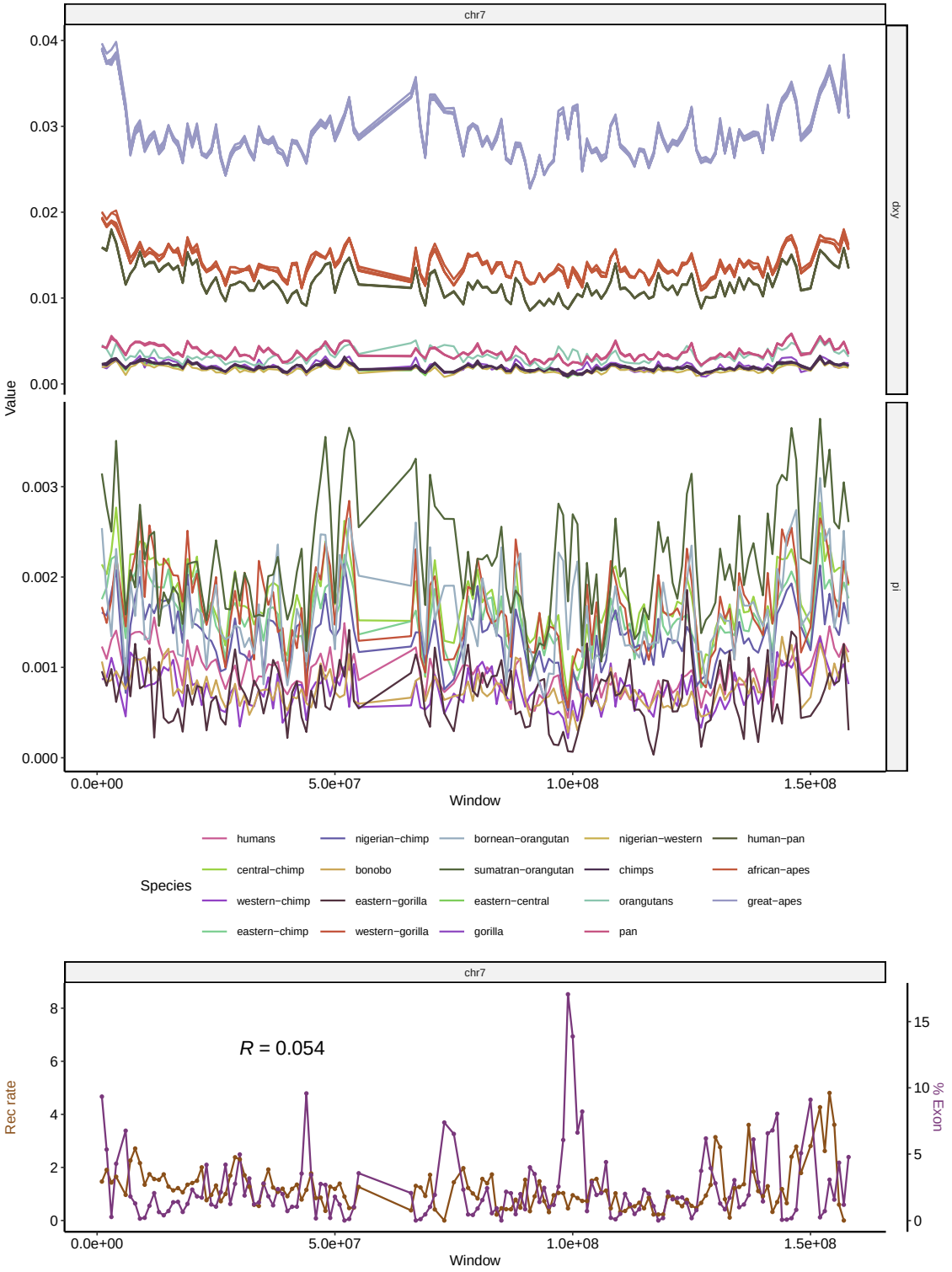


Figure A.16. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 7. See Figure 3.2 for more details.

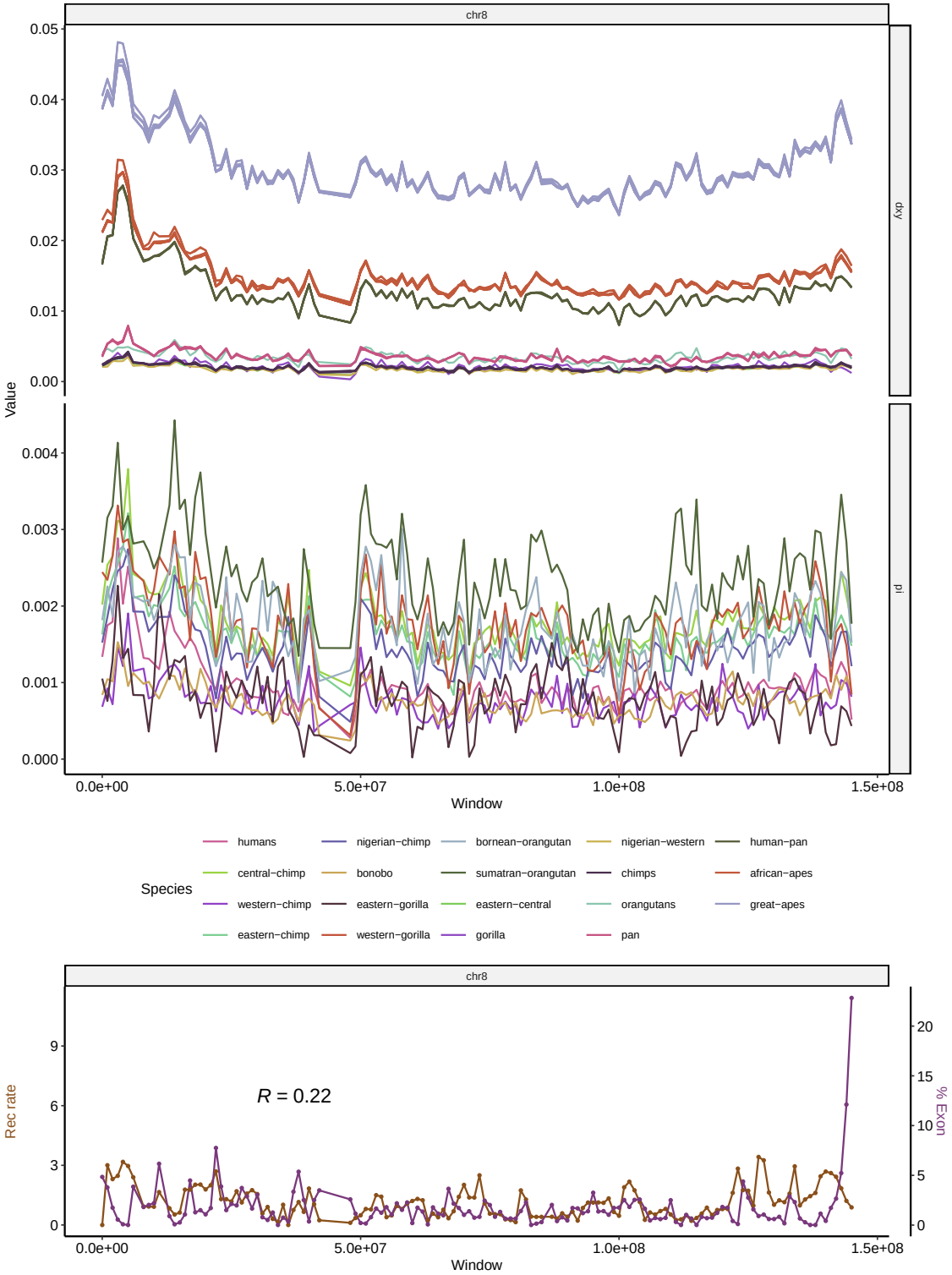


Figure A.17. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 8. See Figure 3.2 for more details.

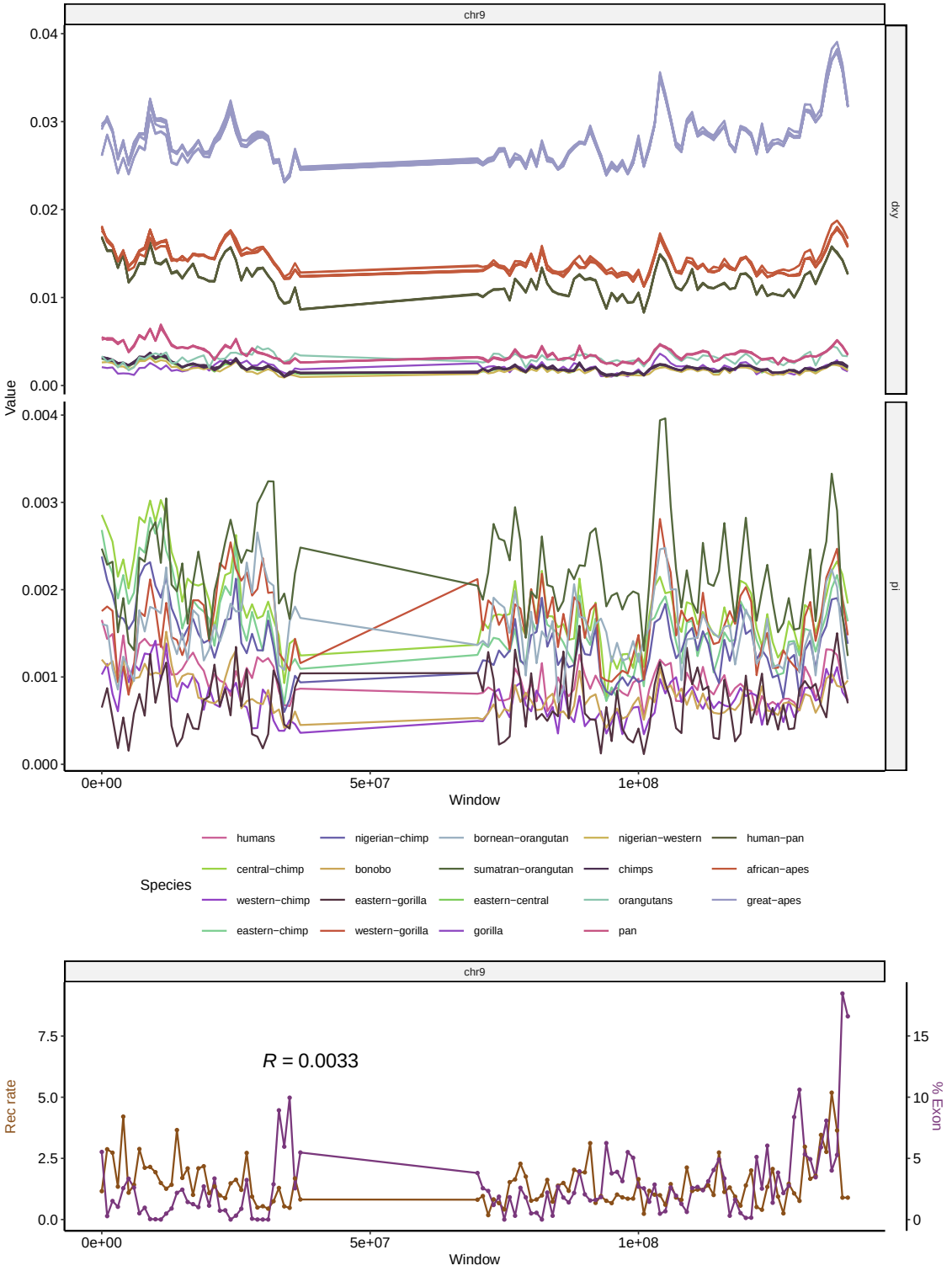


Figure A.18. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 9. See Figure 3.2 for more details.

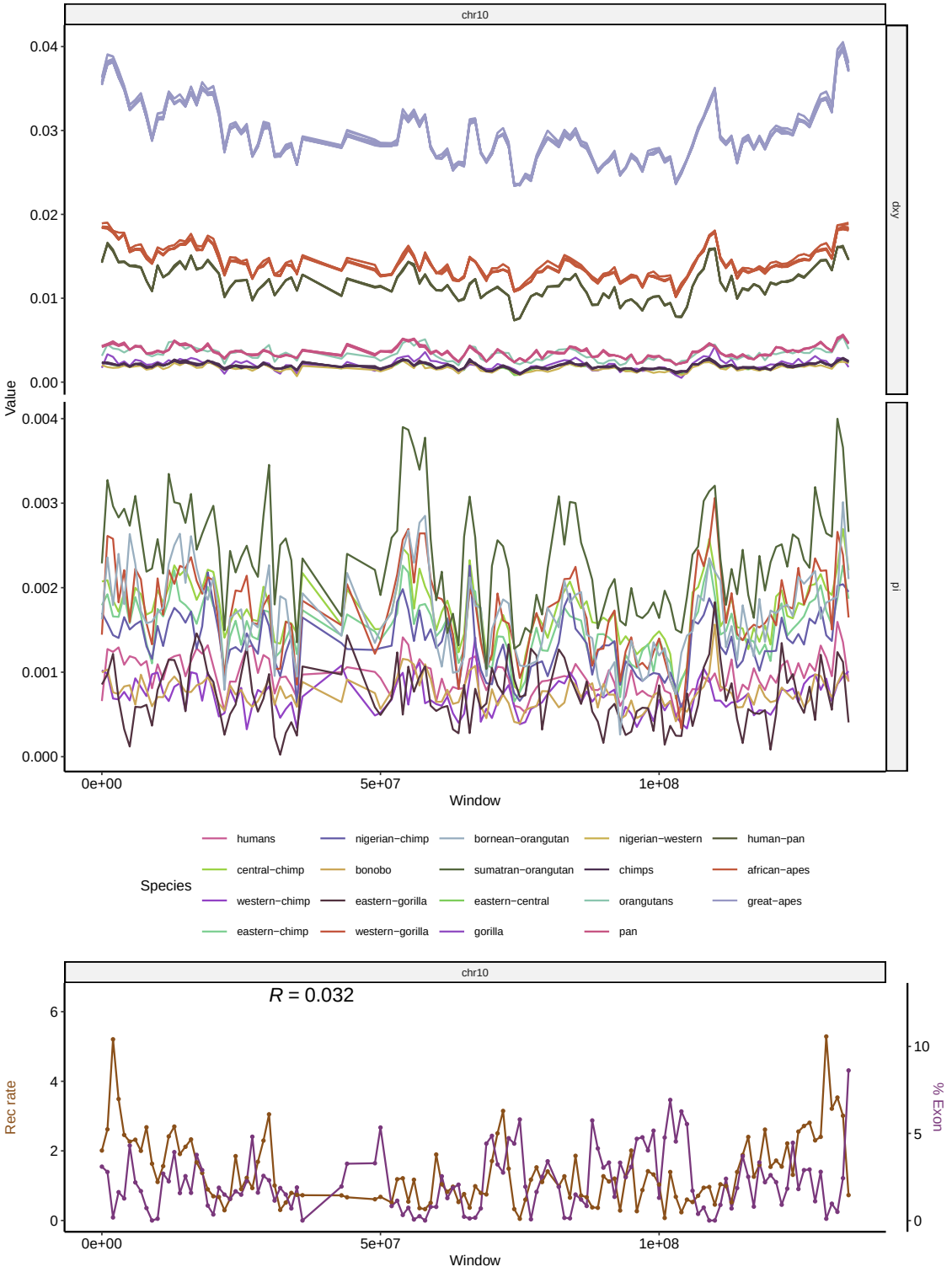


Figure A.19. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 10. See Figure 3.2 for more details.

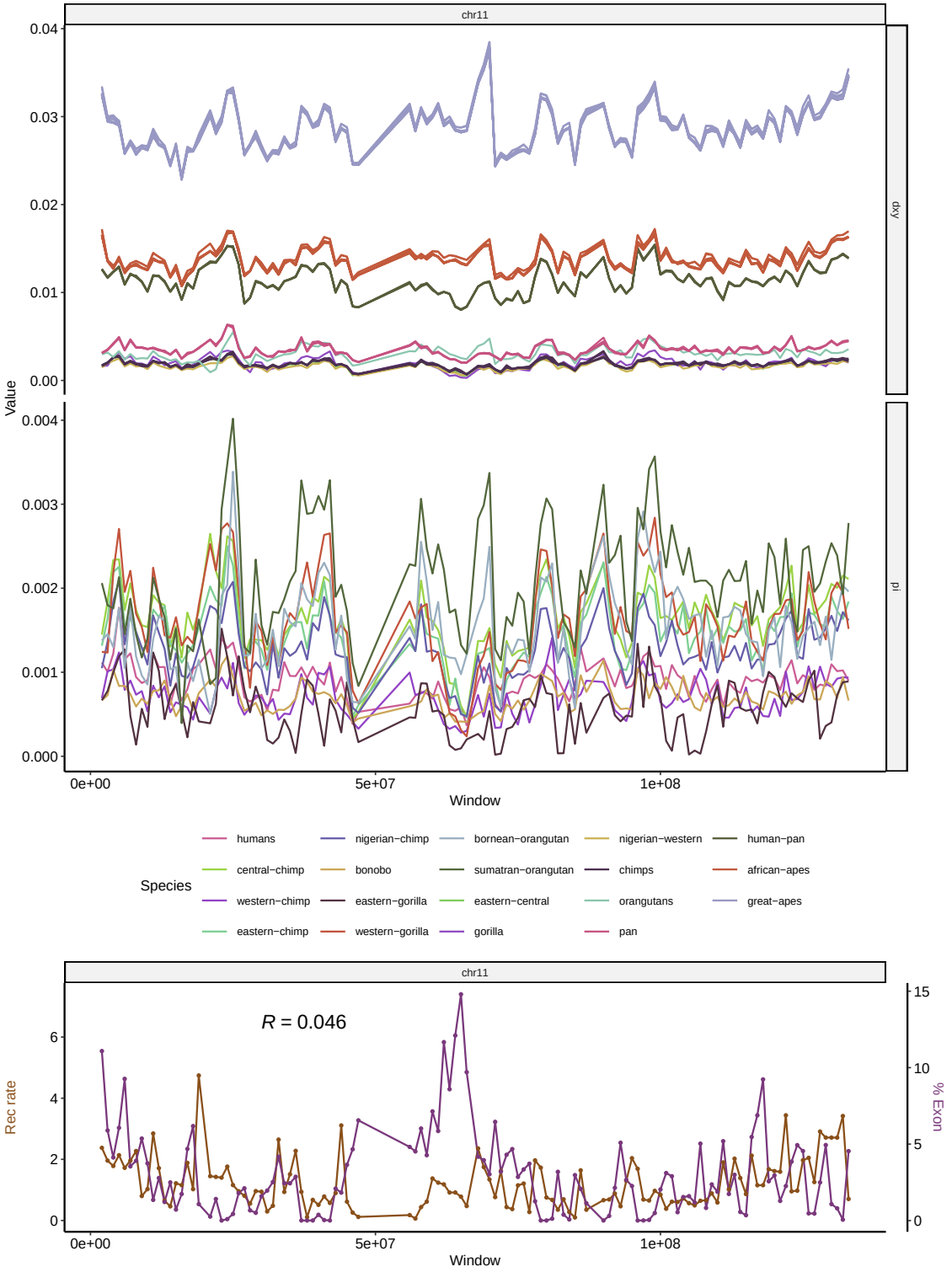


Figure A.20. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 11. See Figure 3.2 for more details.

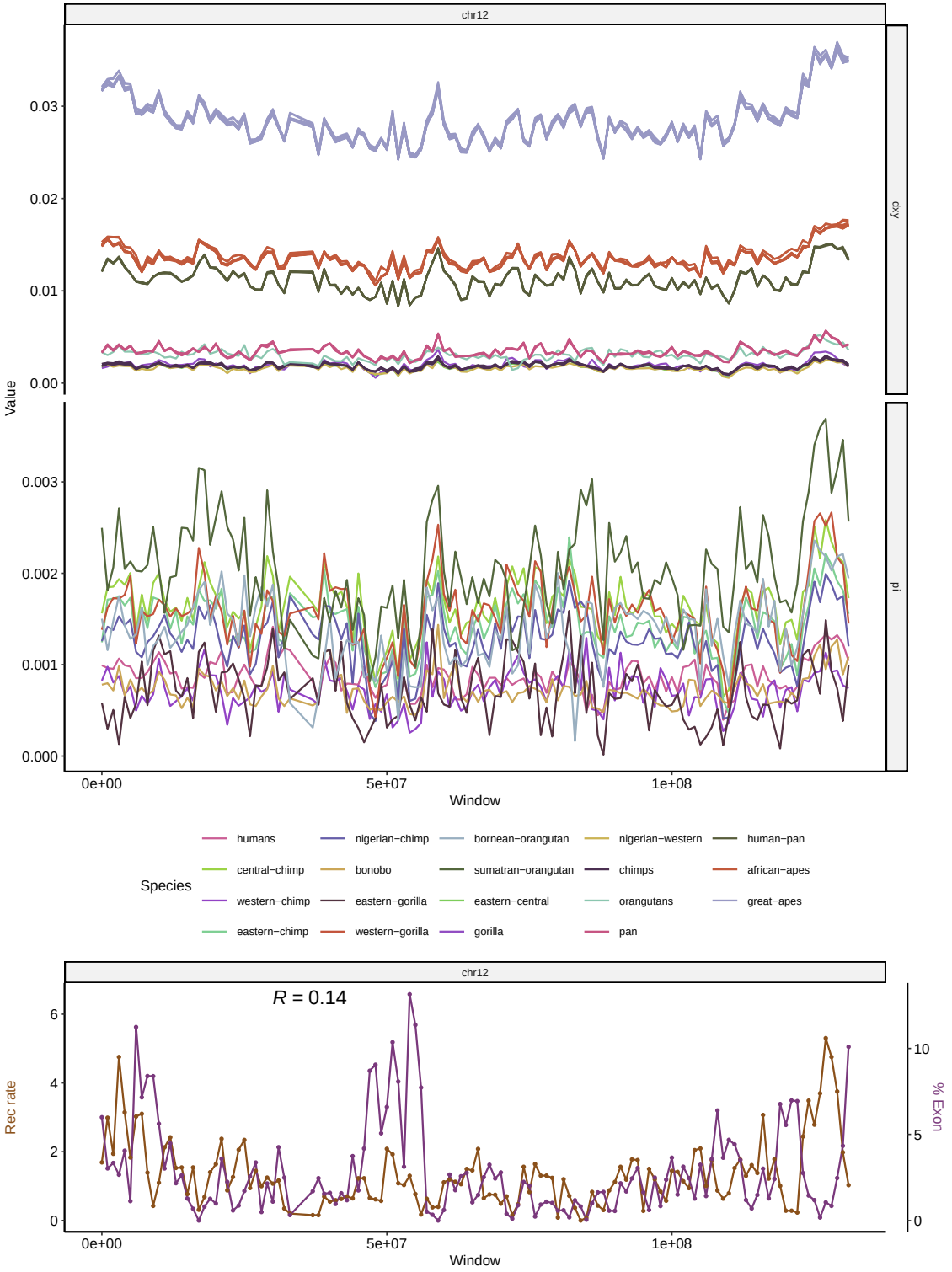


Figure A.21. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 12. See Figure 3.2 for more details.

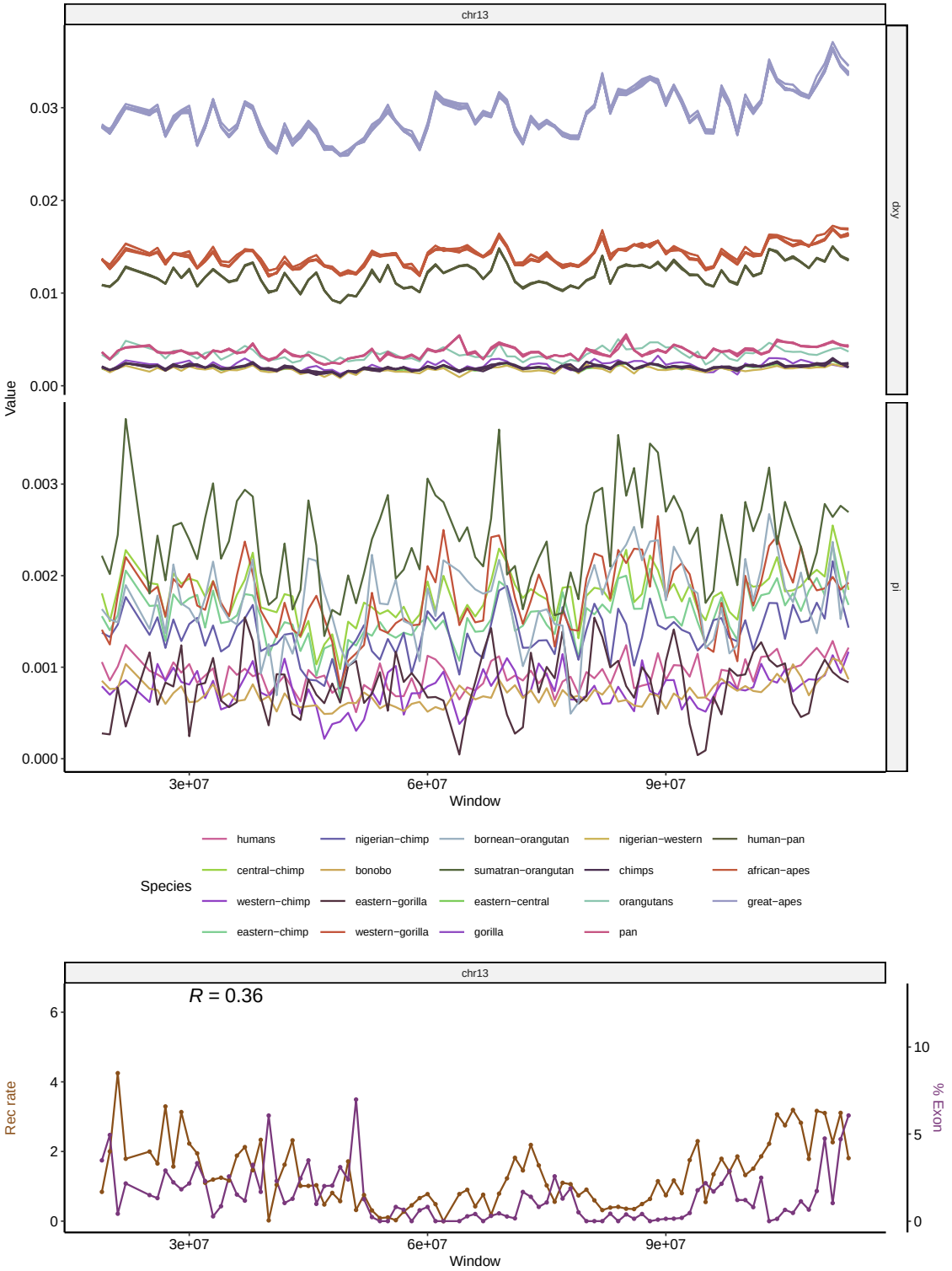


Figure A.22. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 13. See Figure 3.2 for more details.

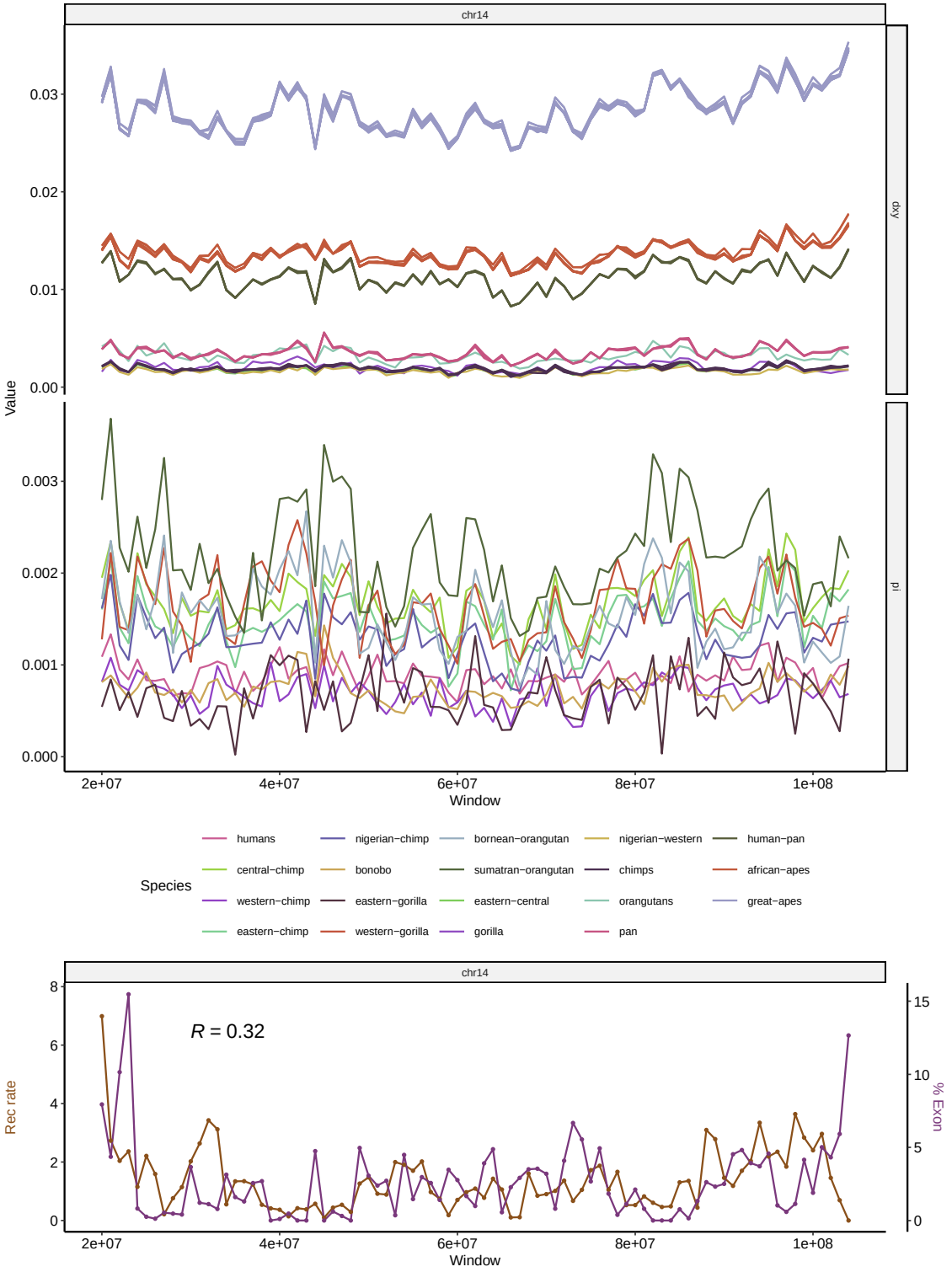


Figure A.23. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 14. See Figure 3.2 for more details.

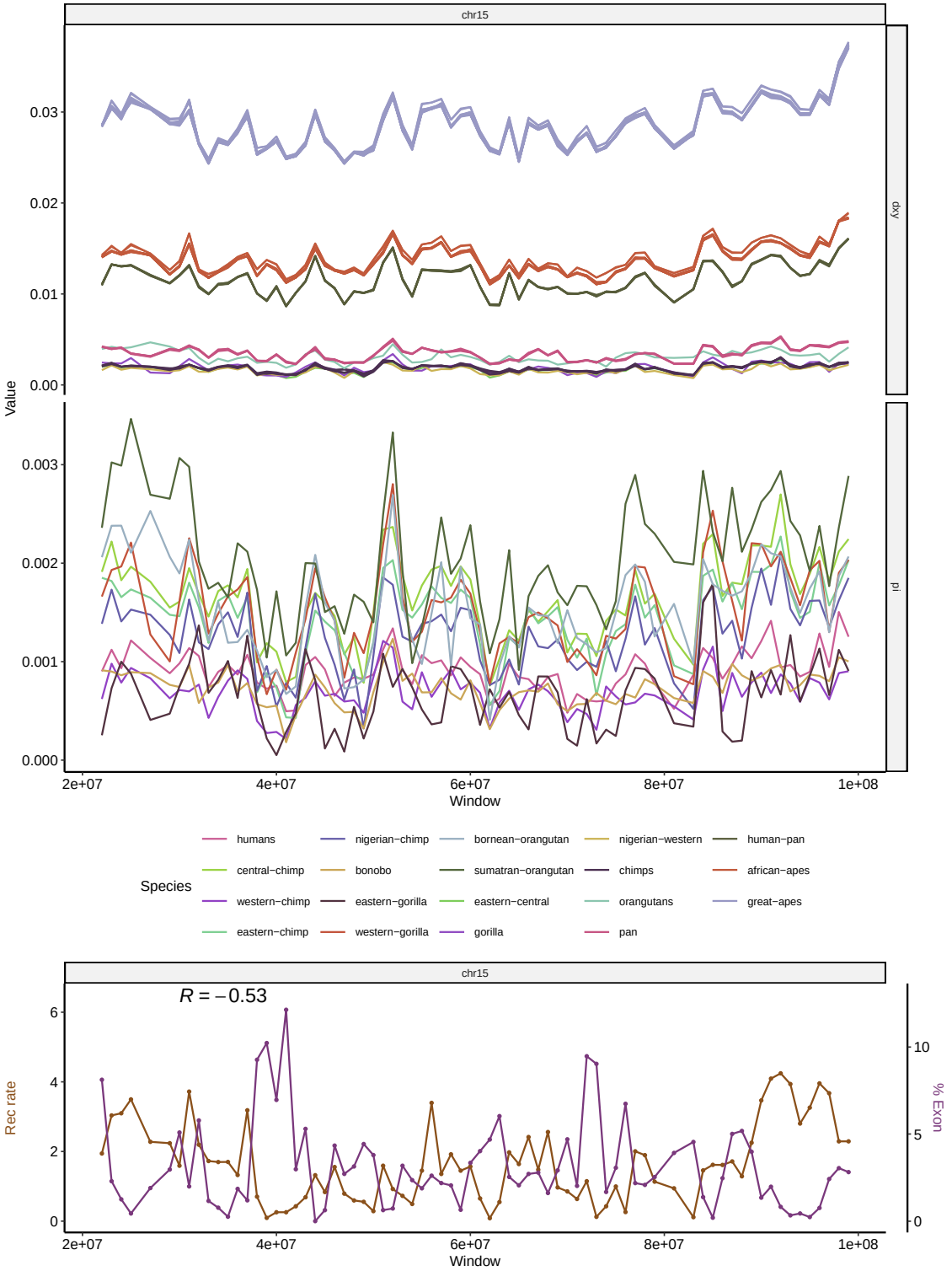


Figure A.24. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 15. See Figure 3.2 for more details.

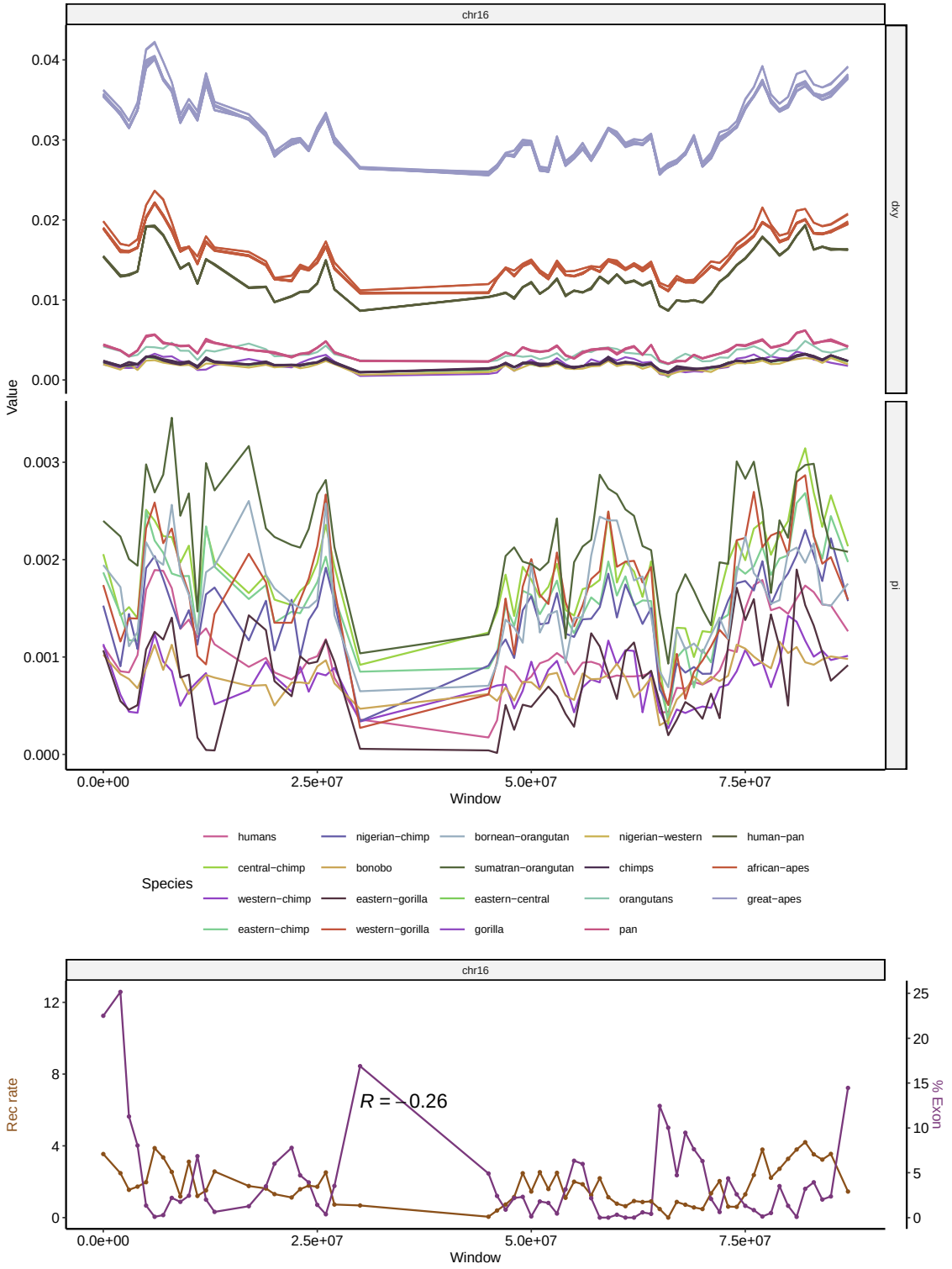


Figure A.25. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 16. See Figure 3.2 for more details.

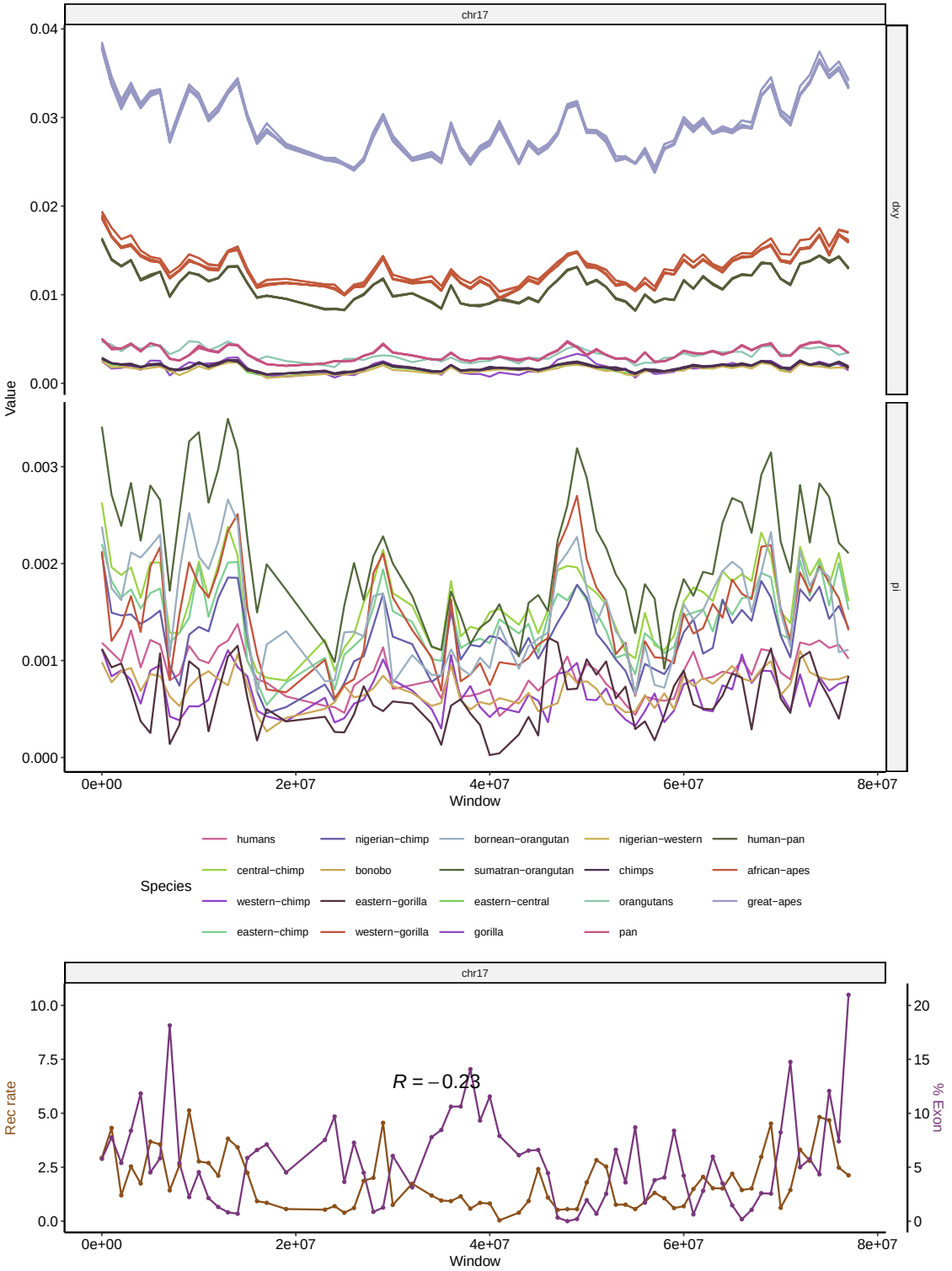


Figure A.26. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 17. See Figure 3.2 for more details.

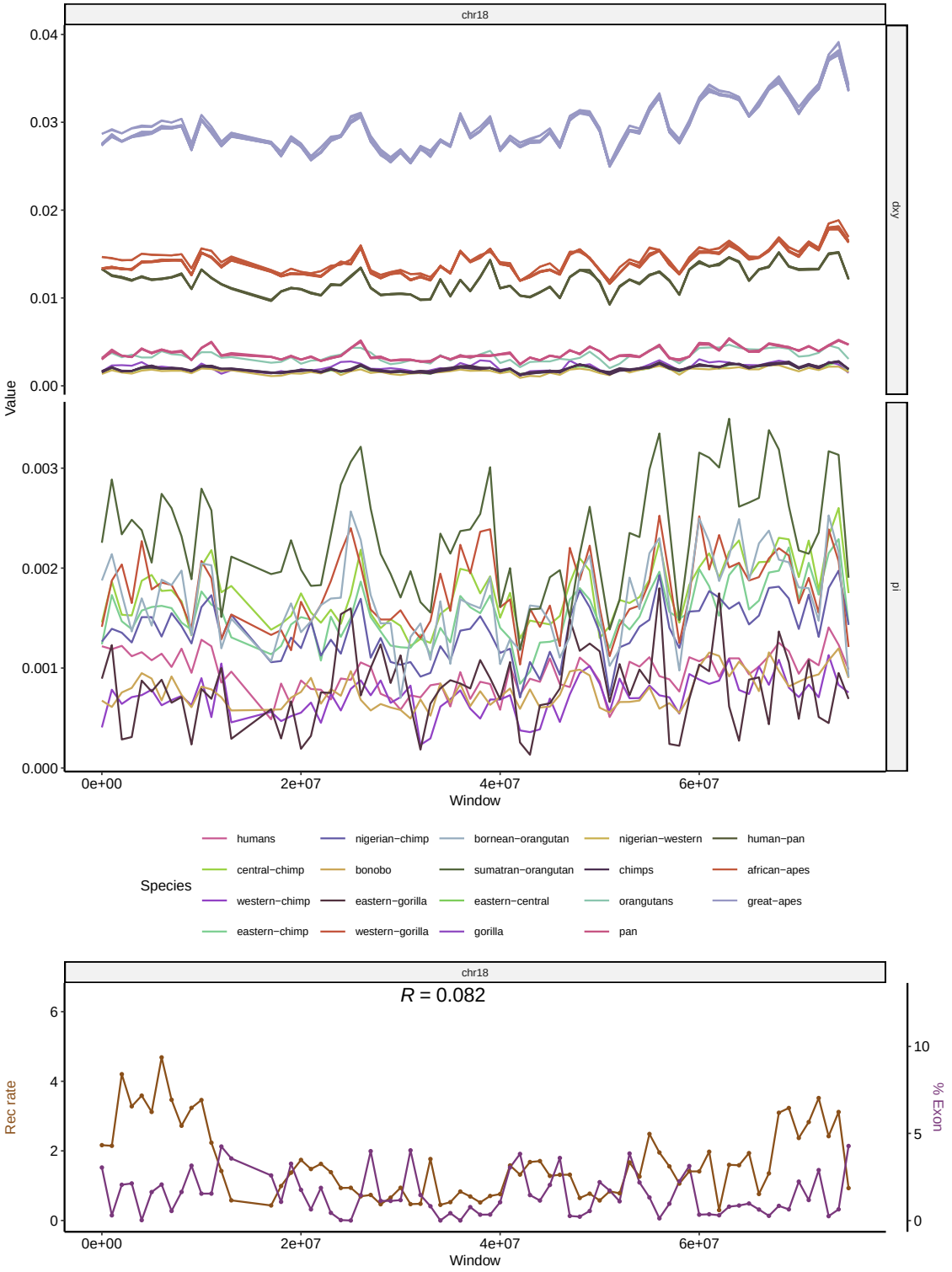


Figure A.27. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 18. See Figure 3.2 for more details.

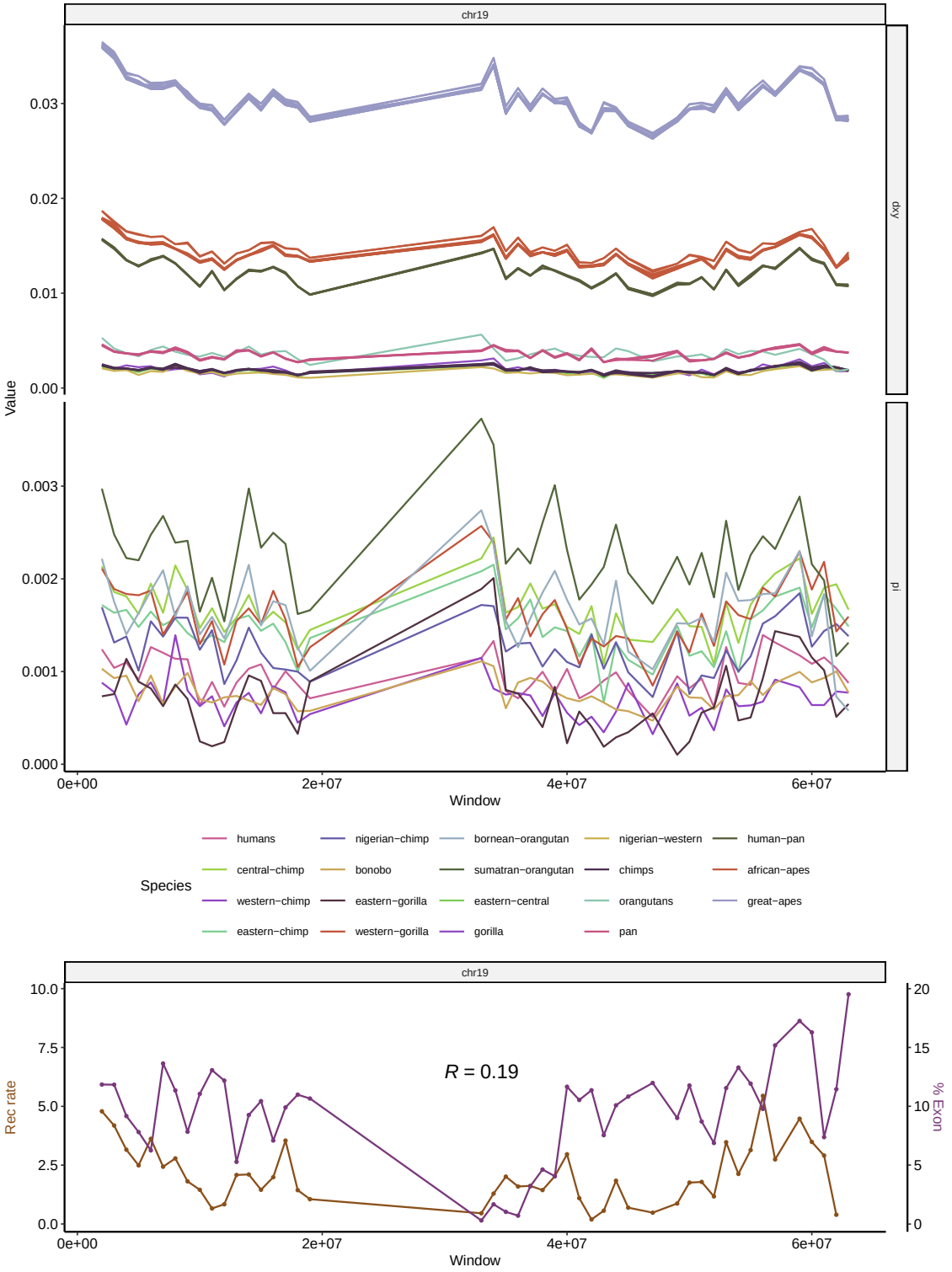


Figure A.28. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 19. See Figure 3.2 for more details.

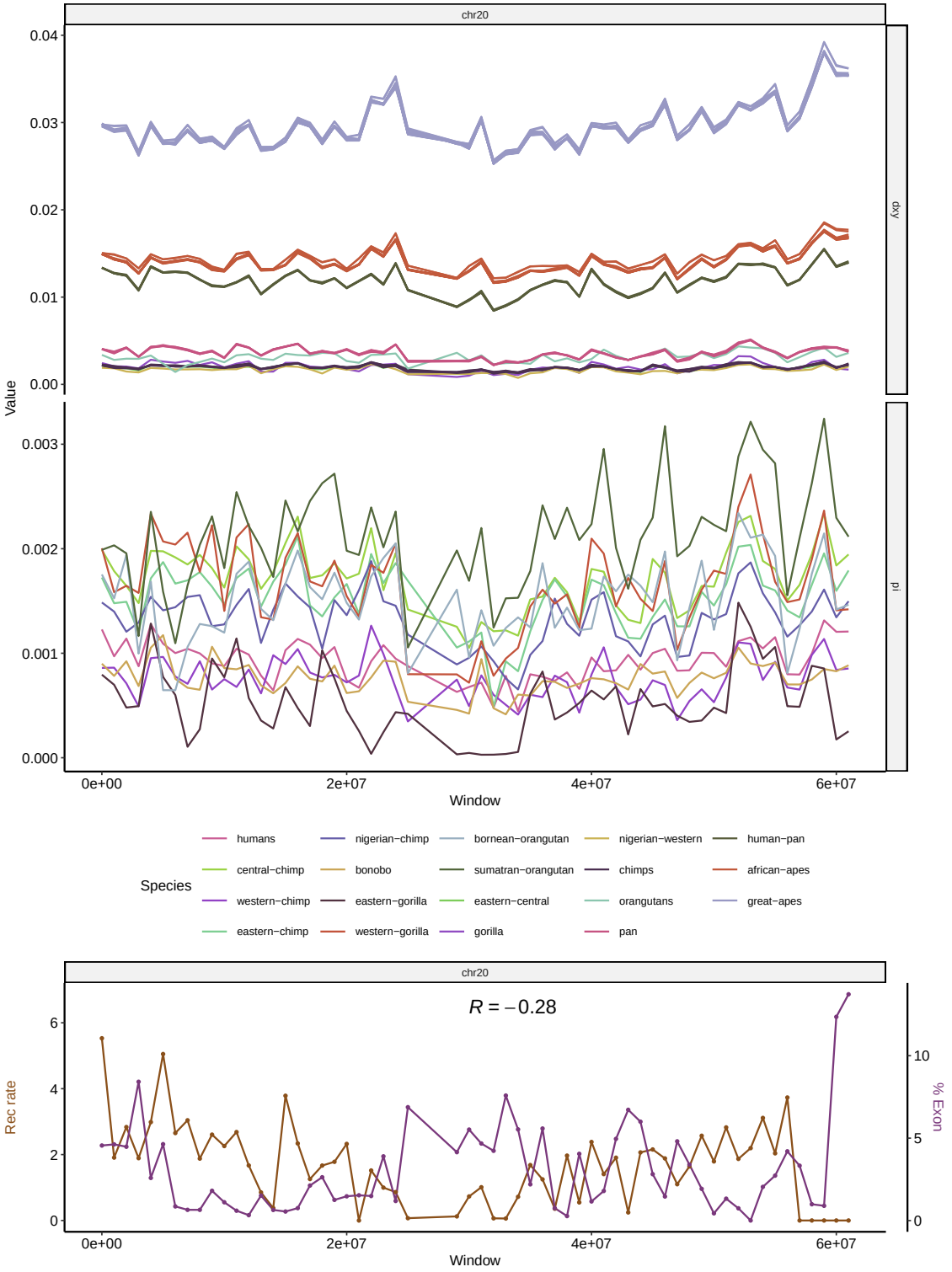


Figure A.29. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 20. See Figure 3.2 for more details.

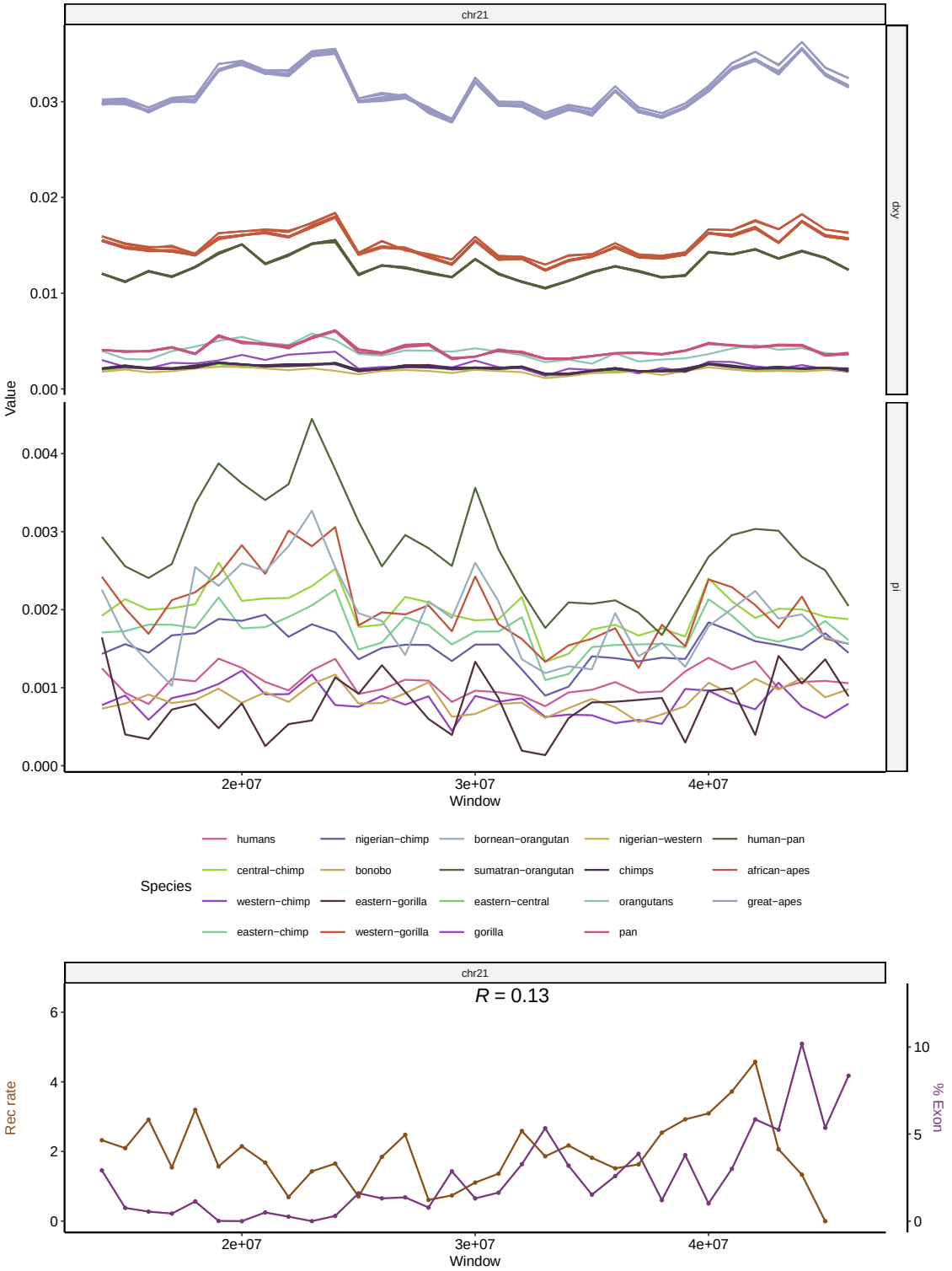


Figure A.30. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 21. See Figure 3.2 for more details.

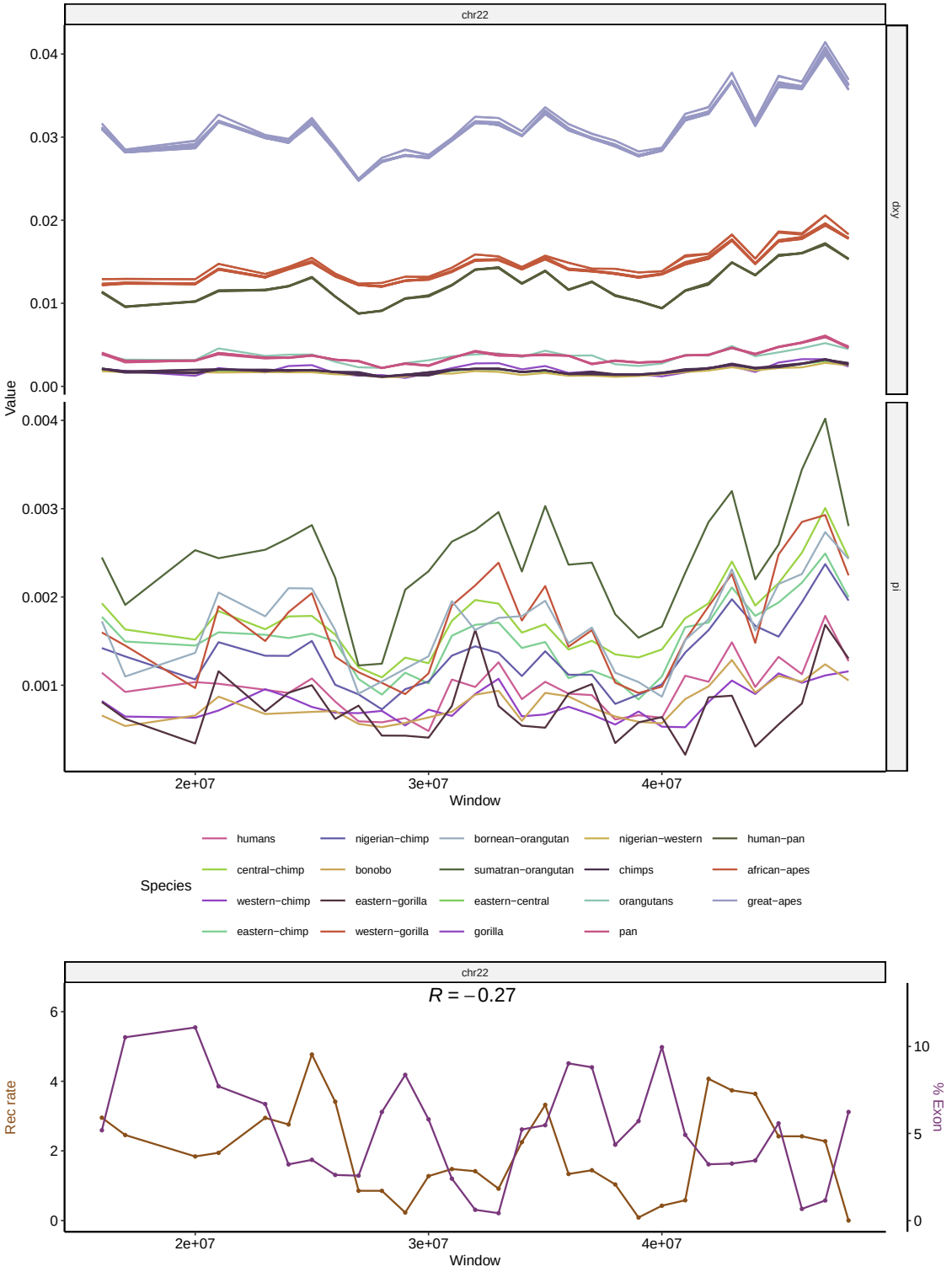


Figure A.31. Landscapes of diversity, divergence, exon density and recombination rate across chromosome 22. See Figure 3.2 for more details.

Bibliography

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., … Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Agarwal, I., & Przeworski, M. (2021). Mutation saturation for fitness effects at human CpG sites (J. Ross-Ibarra & P. J. Wittkopp, Eds.). *eLife*, *10*, e71513. https://doi.org/10.7554/eLife.71513

Andolfatto, P. (2001). Adaptive hitchhiking effects on genome variability. *Current Opinion in Genetics & Development*, *11*(6), 635–641. https://doi.org/10.1016/S0959-437X(00)00246-X

Andolfatto, P. (2007). Hitchhiking effects of recurrent beneficial amino acid substitutions in the drosophila melanogaster genome. *Genome Research*, *17*(12), 1755–1762. https://doi.org/10.1101/gr.6691007

Barton, N. H., & Etheridge, A. M. (2004). The effect of selection on genealogies. *Genetics*, *166*(2), 1115–1131. https://doi.org/10.1093/genetics/166.2.1115

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018, October 17). Relational inductive biases, deep learning, and graph networks. https://doi.org/10.48550/arXiv.1806.01261

Battey, C. J. (2020). Evidence of linked selection on the z chromosome of hybridizing hummingbirds*. *Evolution*, *74*(4), 725–739. https://doi.org/10.1111/evo.13888

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Begun, D. J., & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in d. melanogaster. *Nature*, *356*(6369), 519–520. https://doi.org/10.1038/356519a0

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in drosophila simulans [Publisher: Public Library of Science]. *PLOS Biology*, *5*(11), e310. https://doi.org/10.1371/journal.pbio.0050310

Beissinger, T. M., Wang, L., Crosby, K., Durvasula, A., Hufford, M. B., & Ross-Ibarra, J. (2016). Recent demography drives changes in linked selection across the maize genome. *Nature Plants*, *2*(7), 1–7. https://doi.org/10.1038/nplants.2016.84

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome.

Science (New York, N.Y.), *304*(5675), 1321–1325. https://doi.org/10.1126/science.1098119

Birky, C. W., & Walsh, J. B. (1988). Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Sciences*, *85*(17), 6414–6418. https://doi.org/10.1073/pnas.85.17.6414

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., & Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, *4*(5), e1000083. https://doi.org/10.1371/journal.pgen.1000083

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42. https://doi.org/10.1109/MSP.2017.2693418

Bulmer, M. G. (1976). The effect of selection on genetic variability: A simulation study [Publisher: Cambridge University Press]. *Genetics Research*, *28*(2), 101–117. https://doi.org/10.1017/S0016672300016797

Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, *1*(3), 118–131. https://doi.org/10.1002/evl3.14

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., Hogner, S., Moreno, J., Qvarnström, A., Ružić, M., Sæther, S.-A., Sætre, G.-P., Török, J., & Ellegren, H. (2015). Linked selection and recombination rate variation

drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome Research*, *25*(11), 1656–1665. https://doi.org/10.1101/gr.196485.115

Cai, J. J., Macpherson, J. M., Sella, G., & Petrov, D. A. (2009). Pervasive hitchhiking at coding and regulatory sites in humans. *PLOS Genetics*, *5*(1), e1000336. https://doi.org/10.1371/journal.pgen.1000336

Caldas, I. V., Clark, A. G., & Messer, P. W. (2022, July 20). Inference of selective sweep parameters through supervised learning [Pages: 2022.07.19.500702 Section: New Results]. https://doi.org/10.1101/2022.07.19.500702

Carvajal-Rodriguez, A. (2008). Simulation of genomes: A review. *Current Genomics*, *9*(3), 155–159. https://doi.org/10.2174/138920208784340759

Castellano, D., Eyre-Walker, A., & Munch, K. (2020). Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other homininae. *Genome Biology and Evolution*, *12*(1), 3550–3561. https://doi.org/10.1093/gbe/evz215

Castellano, D., Macià, M. C., Tataru, P., Bataillon, T., & Munch, K. (2019). Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. *Genetics*, *213*(3), 953–966. https://doi.org/10.1534/genetics.119.302494

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S., & Song, Y. S. (2018). A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in Neural Information Processing Systems*, *31*, 8594–8605.

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*(4), 1289–1303.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: Mechanisms, evolution and human disease. *Nature Reviews Genetics*, *8*(10), 762–775. https://doi.org/10.1038/nrg2193

Cleland, C. E. (2002). Methodological and epistemic differences between historical science and experimental science [Publisher: Cambridge University Press]. *Philosophy of Science*, *69*(3), 474–496. https://doi.org/10.1086/342455

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: Insights and challenges from drosophila studies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1736), 20160471. https://doi.org/10.1098/rstb.2016.0471

Coop, G., & Ralph, P. (2012). Patterns of neutral diversity under general models of selective sweeps. *Genetics*, *192*(1), 205–224. https://doi.org/10.1534/genetics.112.141861

Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLOS Biology*, *13*(4), e1002112. https://doi.org/10.1371/journal.pbio.1002112

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. https://doi.org/10.1111/mec.12796

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: Increased sensitivity, robustness and flexibility.

Bioinformatics, *32*(12), 1895–1897. https://doi.org/10.1093/bioinformatics/btw051

Delmore, K. E., Lugo Ramos, J. S., van Doren, B. M., Lundberg, M., Bensch, S., Irwin, D. E., & Liedvogel, M. (2018). Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evolution Letters*, *2*(2), 76–87. https://doi.org/10.1002/evl3.46

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., Künstner, A., Mäkinen, H., Nadachowska-Brzyska, K., Qvarnström, A., Uebbing, S., & Wolf, J. B. W. (2012). The genomic landscape of species divergence in ficedula flycatchers. *Nature*, *491*(7426), 756–760. https://doi.org/10.1038/nature11584

Enard, D., Messer, P. W., & Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. *Genome Research*, *24*(6), 885–895. https://doi.org/10.1101/gr.164822.113

Ewing, G. B., & Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13390]. *Molecular Ecology*, *25*(1), 135–141. https://doi.org/10.1111/mec.13390

Fan, C., Cahoon, J. L., Dinh, B. L., Ortega-Del Vecchyo, D., Huber, C., Edge, M. D., Mancuso, N., & Chiang, C. W. (2023). A likelihood-based framework for demographic inference from genealogical trees. *bioRxiv*, 2023.10.10.561787. https://doi.org/10.1101/2023.10.10.561787

Field, D., & Wills, C. (1997). Long, polymorphic microsatellites in simple organisms [Publisher: Royal Society]. *Proceedings of the Royal Society of*

London. Series B: Biological Sciences, *263*(1367), 209–215. https://doi.org/10.1098/rspb.1996.0033

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238. https://doi.org/10.1093/molbev/msy224

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., ... Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. [Place: England]. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Galtier, N. (2016). Adaptive protein evolution in animals and the effective population size hypothesis. *PLOS Genetics*, *12*(1), e1005774. https://doi.org/10.1371/journal.pgen.1005774

Galtier, N., Duret, L., Glémin, S., & Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in genetics: TIG*, *25*(1), 1–5. https://doi.org/10.1016/j.tig.2008.10.011

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps [Publisher: Public Library of Science]. *PLOS Genetics*, *11*(2), e1005004. https://doi.org/10.1371/journal.pgen.1005004

Glémin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., & Duret, L. (2015). Quantification of GC-biased gene conversion in the human genome. *Genome Research*, *25*(8), 1215–1228. https://doi.org/10.1101/gr.185488.114

Gower, G., Ragsdale, A. P., Bisschop, G., Gutenkunst, R. N., Hartfield, M., Noskova, E., Schiffels, S., Struck, T. J., Kelleher, J., & Thornton, K. R. (2022). Demes: A standard format for demographic models. *Genetics*, *222*(3), iyac131. https://doi.org/10.1093/genetics/iyac131

Griffiths, R., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination [Publisher: Mary Ann Liebert, Inc., publishers]. *Journal of Computational Biology*, *3*(4), 479–502. https://doi.org/10.1089/cmb.1996.3.479

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data [Publisher: Public Library of Science]. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, *19*(2), 552–566. https://doi.org/10.1111/1755-0998.12968

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the wright-fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eöry, L., Keane, T. M., Adams, D. J., & Keightley, P. D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genetics*, *9*(12). https://doi.org/10.1371/journal.pgen.1003995

Hamilton, W. L., Ying, R., & Leskovec, J. (2018, September 10). Inductive representation learning on large graphs. https://doi.org/10.48550/arXiv.1706.02216

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research*, *16*(6), 730–737. https://doi.org/10.1101/gr.5045006

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths [Publisher: Public Library of Science]. *PLOS Genetics*, *9*(6), e1003521. https://doi.org/10.1371/journal.pgen.1003521

Hejase, H. A., Mo, Z., Campagna, L., & Siepel, A. (2022). A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular Biology and Evolution*, *39*(1), msab332. https://doi.org/10.1093/molbev/msab332

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., & Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science (New York, N.Y.)*, *331*(6019), 920–924. https://doi.org/10.1126/science.1198878

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: Tools for population and evolutionary genetics [Number: 2 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *13*(2), 110–122. https://doi.org/10.1038/nrg3130

Hodgkinson, A., & Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, *12*(11), 756–766. https://doi.org/10.1038/nrg3098

Huber, C. D., Kim, B. Y., Marsden, C. D., & Lohmueller, K. E. (2017). Determining the factors driving selective effects of new nonsynonymous

mutations. *Proceedings of the National Academy of Sciences*, *114*(17), 4465–4470. https://doi.org/10.1073/pnas.1619508114

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data [Publisher: [Society for the Study of Evolution, Wiley]]. *Evolution*, *37*(1), 203–217. https://doi.org/10.2307/2408186

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153–159. https://doi.org/10.1093/genetics/116.1.153

Ingvarsson, P. K. (2010). Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in populus tremula. *Molecular Biology and Evolution*, *27*(3), 650–660. https://doi.org/10.1093/molbev/msp255

Irwin, D. E., Alcaide, M., Delmore, K. E., Irwin, J. H., & Owens, G. L. (2016). Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. *Molecular Ecology*, *25*(18), 4488–4507. https://doi.org/10.1111/mec.13792

Jauch, A., Wienberg, J., Stanyon, R., Arnold, N., Tofanelli, S., Ishida, T., & Cremer, T. (1992). Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(18), 8611–8615. Retrieved February 21, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC49970/

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., & Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA

polymorphism data. *Genetics*, *170*(3), 1401–1410. https://doi.org/10.1534/genetics.104.038224

Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., & Jensen, J. D. (2021). The impact of purifying and background selection on the inference of population history: Problems and prospects. *Molecular Biology and Evolution*, *38*(7), 2986–3003. https://doi.org/10.1093/molbev/msab050

Kaplan, N. L., Hudson, R. R., & Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, *123*(4), 887–899.

Katzman, S., Kern, A. D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R. K., Salama, S. R., & Haussler, D. (2007). Human genome ultraconserved elements are ultraselected. *Science*, *317*(5840), 915–915. https://doi.org/10.1126/science.1142430

Katzman, S., Kern, A. D., Pollard, K. S., Salama, S. R., & Haussler, D. (2010). GC-biased evolution near human accelerated regions (J. Zhang, Ed.). *PLoS Genetics*, *6*(5), e1000960. https://doi.org/10.1371/journal.pgen.1000960

Kelleher, J., Etheridge, A. M., & McVean, G. (2016a). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., Etheridge, A. M., & McVean, G. (2016b). Efficient coalescent simulation and genealogical analysis for large sample sizes [Publisher: Public Library of Science]. *PLOS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kelleher, J., Thornton, K. R., Ashander, J., & Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLoS*

computational biology, *14*(11), e1006581. https://doi.org/10.1371/journal.pcbi.1006581

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets [Number: 9 Publisher: Nature Publishing Group]. *Nature Genetics*, *51*(9), 1330–1338. https://doi.org/10.1038/s41588-019-0483-y

Kern, A. D., & Hahn, M. W. (2018). The neutral theory in light of natural selection. *Molecular Biology and Evolution*, *35*(6), 1366–1371. https://doi.org/10.1093/molbev/msy092

Kern, A. D., Jones, C. D., & Begun, D. J. (2002). Genomic effects of nucleotide substitutions in drosophila simulans. *Genetics*, *162*(4), 1753–1761. Retrieved February 9, 2020, from https://www.genetics.org/content/162/4/1753

Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: An updated approach to classifying selective sweeps. *G3 Genes—Genomes—Genetics*, *8*(6), 1959–1970. https://doi.org/10.1534/g3.118.200262

Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2017). Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, *206*(1), 345–361.

Kim, Y., & Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, *155*(3), 1415–1427. Retrieved February 8, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461159/

Kim, Y. (2006). Allele frequency distribution under recurrent selective sweeps. *Genetics*, *172*(3), 1967–1978. https://doi.org/10.1534/genetics.105.048447

Kimura, M. (1980). Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *77*(1), 522–526. https://doi.org/10.1073/pnas.77.1.522

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., & Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, *31*(3), 241–247. https://doi.org/10.1038/ng917

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. A., Frigge, M. L., Helgason, A., Thorsteinsdottir, U., & Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*(7319), 1099–1103. https://doi.org/10.1038/nature09525

Korfmann, K., Sellinger, T., Freund, F., Fumagalli, M., & Tellier, A. (2023, May 11). Simultaneous inference of past demography and selection from the ancestral recombination graph under the beta coalescent [Pages: 2022.09.28.508873 Section: New Results]. https://doi.org/10.1101/2022.09.28.508873

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A.,

... Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, *360*(6393). https://doi.org/10.1126/science.aar6343

Lauterbur, M. E., Cavassim, M. I. A., Gladstein, A. L., Gower, G., Pope, N. S., Tsambos, G., Adrion, J., Belsare, S., Biddanda, A., Caudill, V., Cury, J., Echevarria, I., Haller, B. C., Hasan, A. R., Huang, X., Iasi, L. N. M., Noskova, E., Obšteter, J., Pavinato, V. A. C., ... Gronau, I. (2023). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations [Publisher: eLife Sciences Publications Limited]. *eLife*, *12*. https://doi.org/10.7554/eLife.84874

Laval, G., Patin, E., Boutillier, P., & Quintana-Murci, L. (2021). Sporadic occurrence of recent selective sweeps from standing variation in humans as revealed by an approximate bayesian computation approach. *Genetics*, *219*(4), iyab161. https://doi.org/10.1093/genetics/iyab161

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto, P., & Przeworski, M. (2012). Revisiting an old riddle: What determines genetic diversity levels within species? *PLOS Biology*, *10*(9), e1001388. https://doi.org/10.1371/journal.pbio.1001388

Lewontin, R. C. (1974). *The genetic basis of evolutionary change* (Vol. 560). Columbia University Press New York.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences [Number: 7357 Publisher: Nature Publishing Group]. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Lohmueller, K. E., Albrechtsen, A., Li, Y., Kim, S. Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A. F., Grarup, N., Jørgensen, T., Jiang, T., Witte, D. R., Sandbæk, A., Hellmann, I.,

Lauritzen, T., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLOS Genetics*, *7*(10), e1002326. https://doi.org/10.1371/journal.pgen.1002326

Losos, J. B. (2009, August 15). Evolutionary biology as a historical science. In J. Losos (Ed.), *Lizards in an evolutionary tree: Ecology and adaptive radiation of anoles* (p. 0). University of California Press. https://doi.org/10.1525/california/9780520255913.003.0001

Macpherson, J. M., Sella, G., Davis, J. C., & Petrov, D. A. (2007). Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in drosophila. *Genetics*, *177*(4), 2083–2099. https://doi.org/10.1534/genetics.107.080226

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the adh locus in drosophila. *Nature*, *351*(6328), 652–654. https://doi.org/10.1038/351652a0

McVean, G. A., & Cardin, N. J. (2005). Approximating the coalescent with recombination [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Miles, A., Bot, P. I., Rodrigues, M. F., Ralph, P., Harding, N., Pisupati, R., & Rae, S. (2020). Cggh/scikit-allel: V1. 3.2. *Zenodo*.

Mo, Z., & Siepel, A. (2023). Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data [Publisher:

Public Library of Science]. *PLOS Genetics*, *19*(11), e1011032. https://doi.org/10.1371/journal.pgen.1011032

Murphy, D. A., Elyashiv, E., Amster, G., & Sella, G. (2022). Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements (M. Nordborg, Ed.) [Publisher: eLife Sciences Publications, Ltd]. *eLife*, *11*, e76065. https://doi.org/10.7554/eLife.76065

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297–304. https://doi.org/10.1093/genetics/156.1.297

Nam, K., Munch, K., Mailund, T., Nater, A., Greminger, M. P., Krützen, M., Marquès-Bonet, T., & Schierup, M. H. (2017). Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proceedings of the National Academy of Sciences*, *114*(7), 1613–1618. https://doi.org/10.1073/pnas.1605660114

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, *154*(2), 931–942. https://doi.org/10.1093/genetics/154.2.931

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab]. *Genome Research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Ohta, T., & Kimura, M. (1974). SIMULATION STUDIES ON ELECTROPHORETICALLY DETECTABLE GENETIC VARIABILITY IN a FINITE POPULATION. *Genetics*, *76*(3), 615–624. https://doi.org/10.1093/genetics/76.3.615

Orr, H. A. (2003). The distribution of fitness effects among beneficial mutations. *Genetics*, *163*(4), 1519–1526. https://doi.org/10.1093/genetics/163.4.1519

Pavlidis, P., Živković, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, *30*(9), 2224–2234. https://doi.org/10.1093/molbev/mst112

Phung, T. N., Huber, C. D., & Lohmueller, K. E. (2016). Determining the effect of natural selection on linked neutral divergence across species. *PLOS Genetics*, *12*(8), e1006199. https://doi.org/10.1371/journal.pgen.1006199

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences (K. Veeramah, P. J. Wittkopp & I. Gronau, Eds.). *eLife*, *7*, e36317. https://doi.org/10.7554/eLife.36317

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., ... Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, *160*(3), 1179–1189. https://doi.org/10.1093/genetics/160.3.1179

Ragsdale, A. P., Nelson, D., Gravel, S., & Kelleher, J. (2020). Lessons learned from bugs in models of human history. *The American Journal of Human Genetics*, *107*(4), 583–588. https://doi.org/10.1016/j.ajhg.2020.08.017

Ralph, P., Thornton, K., & Kelleher, J. (2020). Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, *215*(3), 779–797. https://doi.org/10.1534/genetics.120.303253

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs [Publisher: Public Library of Science]. *PLOS Genetics*, *10*(5), e1004342. https://doi.org/10.1371/journal.pgen.1004342

Rodrigues, M. F., Kern, A. D., & Ralph, P. L. (2024). Shared evolutionary processes shape landscapes of genomic variation in the great apes. *Genetics*, iyae006. https://doi.org/10.1093/genetics/iyae006

Rodrigues, M. F., & Ralph, P. L. (2021). *Vignette: Parallelizing SLiM simulations in a phylogenetic tree — PySLiM manual*. Retrieved September 24, 2021, from https://tskit.dev/pyslim/docs/latest/vignette_parallel_phylo.html

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020, October 9). Temporal graph networks for deep learning on dynamic graphs. https://doi.org/10.48550/arXiv.2006.10637

Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., & Sella, G. (2011). Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. *PLoS genetics*, *7*(2), e1001302. https://doi.org/10.1371/journal.pgen.1001302

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understanding human evolution [Number: 10 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *13*(10), 745–753. https://doi.org/10.1038/nrg3295

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history [Number: 12 Publisher: Nature Publishing Group]. *Nature Reviews Genetics*, *16*(12), 727–740. https://doi.org/10.1038/nrg4005

Schrider, D. R. (2020). Background selection does not mimic the patterns of genetic diversity produced by selective sweeps. *Genetics*, *216*(2), 499–519. https://doi.org/10.1534/genetics.120.303469

Schrider, D. R., & Kern, A. D. (2016). S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLOS Genetics*, *12*(3), e1005928. https://doi.org/10.1371/journal.pgen.1005928

Schrider, D. R., & Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, *34*(8), 1863–1877. https://doi.org/10.1093/molbev/msx154

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*(3), 1207–1223. https://doi.org/10.1534/genetics.116.190223

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference [Publisher: Public Library of Science]. *PLOS Computational Biology*, *12*(3), e1004845. https://doi.org/10.1371/journal.pcbi.1004845

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005).

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Simonsen, K. L., Churchill, G. A., & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, *141*(1), 413–429. https://doi.org/10.1093/genetics/141.1.413

Slotte, T. (2014). The impact of linked selection on plant genomic variation. *Briefings in Functional Genomics*, *13*(4), 268–275. https://doi.org/10.1093/bfgp/elu009

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene [Publisher: Cambridge University Press]. *Genetics Research*, *23*(1), 23–35. https://doi.org/10.1017/S0016672300014634

Smith, N., & Eyre-Walker, A. (2002). Adaptive protein evolution in drosophila. *Nature*, *415*(6875), 1022–1024. https://doi.org/10.1038/4151022a

Smith, T. C. A., Arndt, P. F., & Eyre-Walker, A. (2018). Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLOS Genetics*, *14*(3), e1007254. https://doi.org/10.1371/journal.pgen.1007254

Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., & Myers, S. R. (2021). Inferring population histories for ancient genomes using genome-wide genealogies. *Molecular Biology and Evolution*, *38*(9), 3497–3511. https://doi.org/10.1093/molbev/msab174

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples [Number: 9 Publisher: Nature

Publishing Group]. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Stankowski, S., Chase, M. A., Fuiten, A. M., Rodrigues, M. F., Ralph, P. L., & Streisfeld, M. A. (2019). Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. *PLOS Biology*, *17*(7), e3000391. https://doi.org/10.1371/journal.pbio.3000391

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., Bustamante, C. D., Hammer, M. F., & Wall, J. D. (2016). The time scale of recombination rate evolution in great apes. *Molecular Biology and Evolution*, *33*(4), 928–945. https://doi.org/10.1093/molbev/msv331

Torres, R., Stetter, M. G., Hernandez, R. D., & Ross-Ibarra, J. (2020). The temporal dynamics of background selection in nonequilibrium populations. *Genetics*, *214*(4), 1019–1030. https://doi.org/10.1534/genetics.119.302892

Torres, R., Szpiech, Z. A., & Hernandez, R. D. (2018). Human demographic history has amplified the effects of background selection across the genome. *PLOS Genetics*, *14*(6), e1007387. https://doi.org/10.1371/journal.pgen.1007387

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in anopheles gambiae. *PLoS Biology*, *3*(9). https://doi.org/10.1371/journal.pbio.0030285

Uricchio, L. H., & Hernandez, R. D. (2014). Robust forward simulations of recurrent hitchhiking. *Genetics*, *197*(1), 221–236. https://doi.org/10.1534/genetics.113.156935

van Doren, B. M., Campagna, L., Helm, B., Illera, J. C., Lovette, I. J., & Liedvogel, M. (2017). Correlated patterns of genetic diversity and

differentiation across an avian family. *Molecular Ecology*, *26*(15), 3982–3997. https://doi.org/10.1111/mec.14083

Wakely, J. (2016, April 22). *Coalescent theory: An introduction* [Google-Books-ID: x30RAgAACAAJ]. Macmillan Learning.

Wang, J., Street, N. R., Park, E.-J., Liu, J., & Ingvarsson, P. K. (2020). Evidence for widespread selection in shaping the genomic landscape during speciation of populus. *Molecular Ecology*, *29*(6), 1120–1136. https://doi.org/10.1111/mec.15388

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., & Wright, S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of capsella grandiflora. *PLoS Genetics*, *10*(9). https://doi.org/10.1371/journal.pgen.1004622

Williamson, S., & Orive, M. E. (2002). The genealogy of a sequence subject to purifying selection at multiple sites. *Molecular Biology and Evolution*, *19*(8), 1376–1384. https://doi.org/10.1093/oxfordjournals.molbev.a004199

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes [Publisher: American Association for the Advancement of Science]. *Science*, *375*(6583), eabi8264. https://doi.org/10.1126/science.abi8264

Won, Y.-J., & Hey, J. (2005). Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, *22*(2), 297–307. https://doi.org/10.1093/molbev/msi017

Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A. W., & Kelleher, J. (2023, November 4). A general and efficient representation of ancestral recombination graphs [Pages: 2023.11.03.565466 Section: New Results]. https://doi.org/10.1101/2023.11.03.565466

Y C Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, *221*(1), iyac044. https://doi.org/10.1093/genetics/iyac044

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., & Palamara, P. F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits [Number: 5 Publisher: Nature Publishing Group]. *Nature Genetics*, *55*(5), 768–776. https://doi.org/10.1038/s41588-023-01379-x

Zhen, Y., Huber, C. D., Davies, R. W., & Lohmueller, K. E. (2021). Greater strength of selection and higher proportion of beneficial amino acid changing mutations in humans compared with mice and drosophila melanogaster. *Genome Research*, *31*(1), 110–120. https://doi.org/10.1101/gr.256636.119