

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Hidden Genomic Architectures Shaping Variation in *Arabidopsis thaliana*

Permalink

<https://escholarship.org/uc/item/9sb8d8rx>

ISBN

9798193265382

Author

Lensink, Mariele Elizabeth

Publication Date

2026-09-08

Supplemental Material

<https://escholarship.org/uc/item/9sb8d8rx#supplemental>

Peer reviewed|Thesis/dissertation

Hidden Genomic Architectures Shaping Variation in *Arabidopsis thaliana*

By

MARIELE LENSINK
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR IN PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Grey Monroe, Co-Chair

Daniel Kliebenstein, Co-Chair

Jeffrey Ross-Ibarra

Committee in Charge
2026

Acknowledgments

Thank you to my parents, for your unwavering support, for lending me your hope and strength when I could not muster my own, and for championing my aspirations. I know it would be the same even if this dissertation never happened.

To my friends, thank you for making this PhD less lonely, inspiring me with your own resilience, and making me laugh so hard I genuinely fear urinary incontinence.

To my mentors. Dan Kliebenstein, for teaching me how to think and to not be afraid to keep asking questions. For your patience and availability, and for always looking out for me. Grey Monroe, for letting me start a small zoo in the office, and for your contagious excitement about new scientific horizons. David Cummings, for inspiring me to not only be a good scientist, but one with humility, compassion, and a commitment to serving others.

To my postdocs. Kevin, for always treating me like I have something valuable to offer, even as a junior scientist. Chaehee, for your constant encouragement and for being an example of steadfastness, reliability, and perseverance. Ritu, for your compassionate ear, your wisdom, and your openness in demystifying graduate school.

Thank you to the ladies of the prayer group at my mom's church for keeping me on your prayer list for 6 years. Please engrave my name in stone on it - I humbly request a permanent residency. And thank you to all of the women who have championed me, encouraged me to advocate for myself, and do the uncomfortable work of being an inconvenience sometimes.

This dissertation is also dedicated to the memory of Dr. Ryan Botts, whose mentorship put me on a path I never would have believed I was capable of finishing. This dissertation would not exist without you. Thank you for seeing my potential, for teaching me how to code, and for putting a stack of articles about imposter syndrome on my desk as a 20-year-old undergraduate. Thank you for your servant's heart, your deep compassion for students, and your goodness. Your legacy extends far beyond your life.

Abstract

Genetic variation provides the material for evolution, but the variation available to evolutionary processes is itself shaped by biological mechanisms. Mutations determine how genetic differences enter populations, while regulatory systems determine how those differences are expressed. In this dissertation, I use *Arabidopsis thaliana* to examine these two stages: the generation of genetic variation through mutation and the translation of genetic differences into gene expression plasticity.

I investigate the genetic architecture of transcriptomic plasticity in response to salicylic acid using a recombinant inbred population. Although the parental accessions exhibited similar transcriptional responses, their recombinant progeny showed extensive transgressive segregation, with most expression quantitative trait loci acting in *trans*. These results reveal substantial regulatory variation hidden beneath similar parental phenotypes and suggest that trans-regulatory contributions to plasticity may be underestimated in natural populations. I then investigate how mutation-rate heterogeneity contributes to population genetic variation using forward evolutionary simulations and approximate Bayesian computation. Across genomic regions and modeling conditions, inference consistently supported lower mutation rates in genic sequence. Together, these studies span two fundamental stages in the production of evolutionary variation: how genetic differences arise and how those differences are expressed. They demonstrate that observed patterns of genomic and phenotypic variation depend on the underlying molecular and genetic architectures that generate, regulate, and structure the variation on which evolution acts.

Index/Table of contents

Acknowledgments.....	ii
Abstract.....	iii
Index/Table of contents.....	iv
Chapter 1. Trans-regulatory loci shape natural variation of gene expression plasticity in Arabidopsis.....	1
Abstract.....	1
1. Introduction.....	2
2. Materials and Methods.....	6
Experimental design and data collection.....	6
Transcriptome analysis.....	7
eQTL analysis.....	9
Determining <i>cis</i> - vs <i>trans</i> -eQTL position and treatment conditionality.....	10
Calculating eQTL effect size.....	11
Sequence comparison of Bay-0 and Sha to explore potential mechanistic origins of eQTLs.....	11
3. Results.....	12
Increased influence of genotype and plasticity on transcription in RILs.....	12
Greater transcriptome variation in RILs.....	13
<i>Trans</i> -regulation as a major driver of gene expression plasticity.....	16
Plasticity is largely controlled by <i>trans</i> -regulatory loci of small to moderate effect size...	18
Large-effect <i>trans</i> -eQTL.....	20
4. Discussion.....	23
5. Declarations.....	26
6. Supplemental Tables and Figures.....	27
7. References.....	29
Chapter 2. Mutation Rate Heterogeneity Shapes Population-Level Genetic Variation in Arabidopsis.....	39
Abstract.....	39
1. Introduction.....	40
Mutation, repair, and spatial heterogeneity.....	40
The mutation-selection identifiability problem.....	42
Existing approaches and their limitations.....	42
Mutation accumulation (MA) experiments.....	43
Population genetic inference approaches.....	43
Simulation-based inference approaches.....	44
Study aims and approach.....	44
2. Materials and Methods.....	45
Genomic region selection.....	45
Forward evolutionary simulations.....	46

Empirical data processing and summary statistic calculation.....	48
Approximate Bayesian computation inference.....	50
Simulation-based parameter recovery.....	52
3. Results.....	53
Empirical inference supports higher intergenic mutation rates.....	54
The relative mutation-rate contrast is robust to simulation scaling.....	57
Pseudo-observed tests establish greatest confidence in mutation-rate inference.....	58
Genic hypomutatin is robust to demographic assumptions.....	62
Summary.....	63
4. Discussion.....	64
Inferred mutation-rate contrast is consistent with gene-body hypomutation.....	66
Parameter relationships may contribute to recovery hierarchy.....	68
Conclusions.....	69
5. Declarations.....	70
Acknowledgments.....	70
Data availability.....	70
Code availability.....	70
6. Supplemental Tables and Figures.....	70
7. References.....	77

Chapter 1. *Trans*-regulatory loci shape natural variation of gene expression plasticity in *Arabidopsis*

Mariele Lensink, Grey Monroe, Daniel J Kliebenstein, *Trans*-regulatory loci shape natural variation of gene expression plasticity in *Arabidopsis*, *Genetics*, Volume 230, Issue 4, August 2025, iyaf116, <https://doi.org/10.1093/genetics/iyaf116>

As first author, I led the study from its initial development through publication. I contributed to the development of the research question and study design with a co-author and independently conducted the analyses, interpreted the results, prepared the figures, and wrote the manuscript. The empirical data used in the study were generated by collaborators.

Abstract

Organisms regulate gene expression in response to environmental cues, a process known as plasticity, to adjust to changing environments. Research into natural variation and the evolution of plasticity frequently studies cis-regulatory elements with theory suggesting they are more important evolutionarily than trans-regulatory elements. Genome-wide association (GWA) studies have supported this idea, observing a predominance of cis-loci affecting plasticity. However, studies in structured populations provide a contrasting image, raising questions about the genetic architecture of natural variation in plasticity. To circumvent potential statistical difficulties present in GWA studies, we mapped loci underlying transcriptomic plasticity in response to salicylic acid (SA) using recombinant inbred lines generated from 2 random *Arabidopsis thaliana* accessions. We detected extensive transgressive segregation in the SA response, suggesting that plasticity to salicylate in *Arabidopsis* is polygenic. Most loci (>75%) underlying this variation act in trans, especially for loci influencing plasticity. Trans-acting loci were enriched in genome hotspots, with predominantly small-effect

sizes distributed across many genes. This could potentially explain their under-discovery in GWA studies. This work reveals a potentially important role for trans-acting loci in plastic expression responses, with implications for understanding plant adaptation to different environments.

1. Introduction

Plasticity can be defined as the capacity of an individual genotype to change a trait in response to external stimuli (Pigliucci et al. 2006). Plastic responses allow organisms to optimize their phenotypes to respond to and persist in their environment. As such, plasticity can be evolutionarily advantageous by enabling a species to persevere at a new phenotypic optimum until genetic variation arises that generates a more adaptive phenotype. This model largely depends on the strength of selection and the time scale of evolutionary change (Ghalambor et al. 2007; Levis and Pfennig 2016). Furthermore, genetic variation in plasticity has been shown to be advantageous in novel environments and increase the likelihood of evolutionary rescue (Walter et al. 2023). Understanding how genetic variation in plasticity is genomically distributed will provide valuable insight into the regulation of response to environmental stimuli and inform strategies to increase plant resilience to changing climates.

While plasticity affects a wide array of physiological and morphological traits, gene expression is a key tool for studying how plasticity evolves and varies. Gene expression is inherently dynamic, considered to be more directly linked to genetic changes than morphological phenotypes, and relatively easy to quantify en masse with transcriptomics (Schlichting and Smith 2002; de Nadal et al. 2011). Thus, studies on a

gene's expression plasticity can yield more detailed mechanistic insight into how plasticity variation occurs than may be possible with a physiological or morphological trait. These benefits have led many studies to use transcriptomics to measure and map plasticity in naturally variable populations. The identified genomic loci that explain variation in transcription are called expression quantitative trait loci, or eQTL. In the context of a single gene, plasticity can be cis-regulated by genomic elements adjacent to or within the gene of interest, or trans-regulated by loci unlinked to the gene of interest. Natural variation in plasticity is theorized to be predominantly controlled by cis-regulators, as cis effects typically have less pleiotropic effects, are predicted to experience relaxed or balancing selection (as opposed to purifying selection for trans-regulators), and are more influential in gene expression divergence (Hoekstra and Coyne 2007; Metzger et al. 2017; Brown and Kelly 2022; Vande Zande et al. 2022). In contrast, trans-regulatory mutations tend to be more pleiotropic with potentially deleterious effects. The resulting theory that cis-acting loci are under weaker purifying selection than trans-acting eQTL is supported by genome-wide association (GWA) studies of natural populations that identify cis-eQTLs as the primary identifiable basis for plasticity variation (Mai et al. 2021; Sun et al. 2023; Xu et al. 2023). Consequently, research regarding the natural variation in plasticity and its evolution often focuses on cis-acting changes (Massouras et al. 2012; He et al. 2016, 2021; Pruunsild and Bading 2019; Dardiry et al. 2023; Takou et al. 2024).

In contrast to GWA populations, many studies using structured populations [e.g. recombinant inbred lines (RILs) originating from a biparental cross] find mainly

trans-acting regulation of gene expression and plasticity. In recombinant inbred strains of *Caenorhabditis elegans* and *Zea mays*, a significantly larger proportion of trans-acting genes had plasticity eQTL in contrast to cis-acting genes (Li et al. 2006, 2013). In inbred mice, expression plasticity was also largely attributed to trans-acting loci (Ballinger et al. 2023). Supporting these observations was work on a small *Arabidopsis thaliana* (*Arabidopsis* hereafter) collection of accessions that showed the majority of plasticity variation links to coexpressed gene networks, suggesting mainly trans-eQTL (van Leeuwen et al. 2007). The difference between GWA and biparental populations establishes a discord regarding whether variation in gene expression plasticity is predominantly controlled by cis or trans factors.

One possible explanation for the difference in the relative proportions of trans and cis-acting eQTL between structured and GWA populations is ascertainment bias, caused by the nature of cis vs trans gene regulation. Cis-eQTL frequently has larger effects on a transcript's expression compared with trans loci. GWA studies, leveraging the diversity of alleles in natural populations, are generally powered to detect loci with moderate to large effects but are less effective at detecting loci with smaller effects. This inherent bias in GWA studies favors associations with moderate-frequency alleles that exhibit large effects (Myles et al. 2009; Sham and Purcell 2014), and detected rare alleles are likely to have an overestimated effect size (Göring et al. 2001). In selfing species such as *Arabidopsis*, population structure and the trait of interest are more likely to be confounded, complicating the detection of true associations in GWA studies (Zhao et al. 2007). In contrast, structured populations, by having fewer genotypes per locus

and balanced allele frequencies, are better suited to find these smaller effect loci (Gibson 2012; Korte and Farlow 2013; Marigorta et al. 2018; Liu et al. 2019). The advantages and disadvantages of GWA and QTL studies are comprehensively discussed in Josephs et al. (2017).

Previous work in *Arabidopsis* has followed this trend; GWA of transcript expression in specific conditions and their plastic response has largely found and focused on cis-regulation (Zhang et al. 2011; Lasky et al. 2014; He et al. 2016, 2021; Zan et al. 2016). To map the architecture of plasticity and assess the relative contributions of cis- and trans-eQTL within *Arabidopsis*, we used the structured Bayreuth (Bay-0) × Shahdara (Sha) *Arabidopsis* RIL population. This RIL population was treated with the defense signaling hormone, salicylic acid (SA), or received control treatment, and the transcriptome was measured at 28 h post-treatment.

SA was selected for plasticity analysis because it is a key hormone regulating plant defense, and its exogenous application induces responses that mirror endogenous signaling during pathogen attack. In *Arabidopsis*, SA is biosynthesized by the isochorismate pathway involving ICS1, EDS5, PBS3, and EPS1 (Peng et al. 2021). SA then interacts with NPR1 and/or NPR3/4 to regulate sets of targets, including TGACG-Binding transcription factors, typically by Sumoylation or Ubiquitination of the targets (Zhang and Li 2019; Peng et al. 2021). The ability to synthesize and respond to SA is highly conserved across land plants and essential for defense against pathogens (Spoel and Dong 2024). SA is also highly interconnected with other defense pathways,

such as the jasmonic acid pathway, creating numerous feedback and feedforward loops that can regulate its basal and response behavior (Peng et al. 2021; Spoel and Dong 2024). While SA-mediated defense responses are highly conserved across land plants, the specific biosynthetic and regulatory pathways can vary across plant families (Peng et al. 2021). In this way, SA treatment provides a defined treatment that can induce plasticity responses similar to those in plant-pathogen interactions while minimizing physiological shifts. It should be noted that, because SA is an endogenous plant phytohormone, it exists in small quantities in untreated plants. As a result, our control group does not explicitly represent an absence of signal. The Bay-0 × Sha population was chosen as these are diverse representatives of Arabidopsis with species typical SA responses (West et al. 2006, 2007; van Leeuwen et al. 2007). The 28-h time point was chosen to maximize the plasticity response of gene expression while minimizing any physiological change (van Leeuwen et al. 2007). Using this transcriptome data, we mapped plasticity eQTL controlling variation in the SA response and showed that they predominantly function in trans and with small effects.

2. Materials and Methods

Experimental design and data collection

We acquired transcript abundance data for the Bay-0 × Sha parental and F8 RIL lines (Loudet et al. 2002) from a previous microarray experiment (West et al. 2007). In this experiment, the parents and 211 RILs were grown in a randomized design and treated with either a control silwet L77 (a surfactant), or 0.3 mM SA in silwet L77 (West et al. 2007). The control and SA-treated plants were grown together and tissue was collected

28 h after treatment (van Leeuwen et al. 2007). Growth and treatment protocols for plants are fully described in West et al. (2007). The entire experiment was independently replicated twice. The transcriptomic analysis of the SA treatment has not been previously reported or compared with the control treatment to investigate plasticity.

The microarray used was the ATH1 Genome Array where each of 24,000 transcripts is measured using 11 probes per transcript, with each probe having 25 bases (ThermoFisher Scientific). The 11 independent and non-overlapping probes per gene are predominantly in the 3'UTR of the gene and designed using the Arabidopsis genome sequence to maximize the ability to specifically measure an individual transcript compared with any closely related sequences. Any potential cross-hybridizing probe sets are identified in the original design. This minimizes potential cross-signal caused by related sequences for other genes (Redman et al. 2004).

Transcriptome analysis

Using the expression estimate for each transcript across all the genotypes, we tested the distribution of variance linked to experimental replicates, treatment, genotype, and their interaction in both the parental accessions and RILs. This was done using a linear mixed model within the R package “R/car” (Fox and Weisberg 2018), with experimental replicate as a random effect while genotype and treatment were modeled as fixed effects. The sum of squares for each factor for each transcript was calculated using a type III ANOVA. These variance results were then clustered into 10 profiles using kmeans to better visualize the major trends in variation in gene expression attributed to

the model's effects. Samples of 1,000 transcripts from each cluster were randomly selected and plotted, independently in parents and RILs (Figure 1.1a).

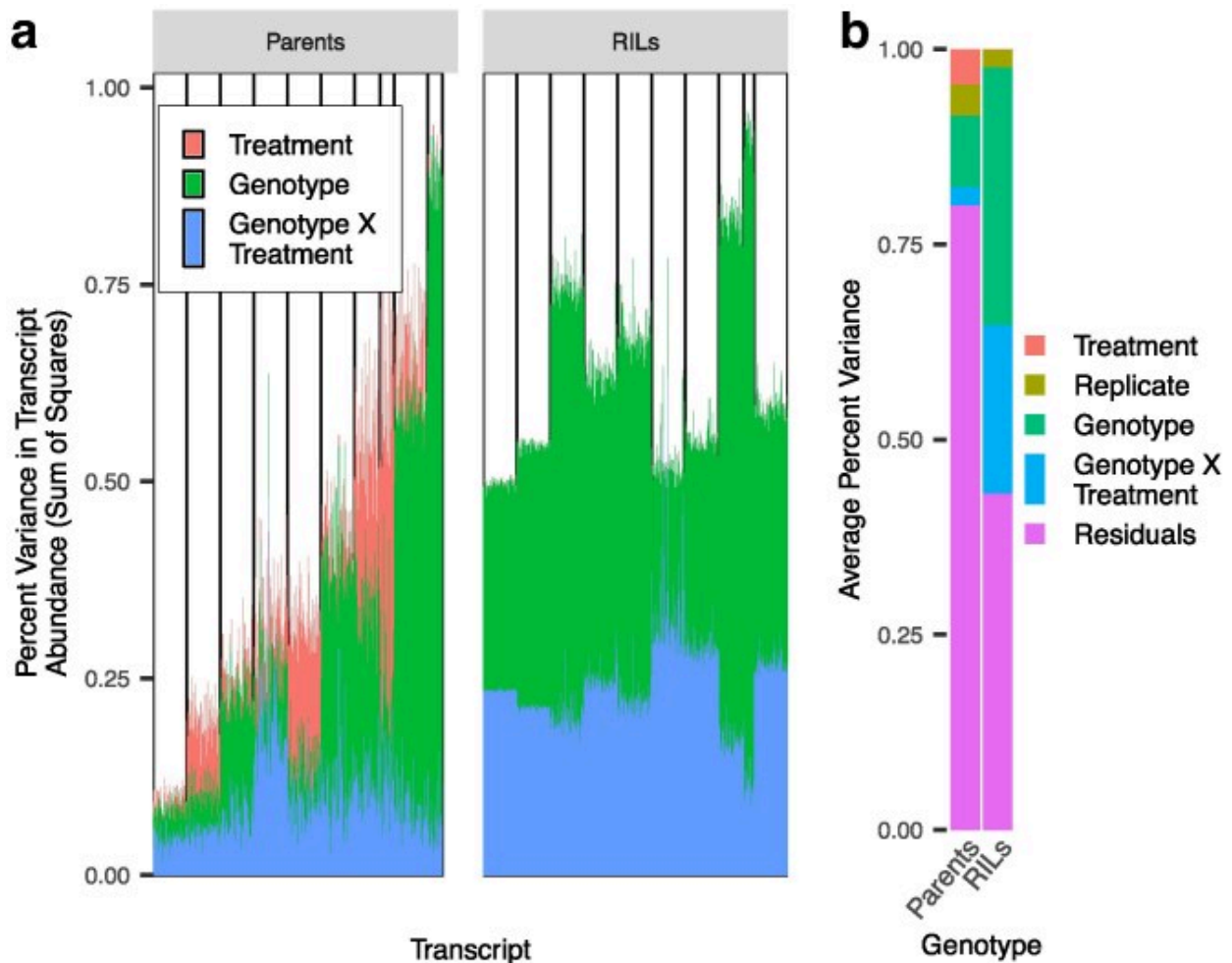


Figure 1.1. Changes in gene expression across environments are attributed to separate experimental variables between parent and RIL genotypes. a) Variance in transcript abundance (sum of squares/TSS) attributable to each experimental variable for parents and RILs. Results from the type III ANOVA for each transcript were clustered into 10 variance profiles using k means, and 1,000 transcripts were sampled from each cluster for visualization. b) Average proportion of variation (sum of squares) across transcripts explained by each experimental variable.

The average expression for each transcript within each treatment was estimated by calculating the mean value of every transcript for every individual across biological replicates. The mean values were then used to visualize how the transcriptome varies across recombinant inbred and parental lines via a principal component analysis (PCA)

using R/stats. This was repeated after removing transcripts that contained cis-eQTLs to test if the pattern was altered.

eQTL analysis

The average value per transcript per treatment per genotype was then used for eQTL mapping. The RILs were previously genotyped using a combination of genetic markers and the map was obtained from (Loudet et al. 2002; West et al. 2006). Using this map, which has an average marker distance of 0.17 cM, we conducted eQTL analyses with 210 of the 211 RILs due to missing genotype information for RIL 417. For this analysis, we used the R/QTL2 package using Haley–Knott regression (Broman et al. 2019). eQTL analysis was performed on transcript values obtained using either silwet (control) or SA samples and as a mathematically derived value called “delta.” Delta represents an estimate of the plasticity phenotype obtained by calculating the difference in each transcript abundance across the 2 treatments divided by the average transcript abundance across the 2 treatments: $(SA - control) / [(SA + control) / 2]$. This was calculated independently for each genotype for each transcript. These 3 eQTL analyses, being SA, control, and delta, will hereafter be referred to as “expression phenotypes.” QTL locations were estimated for each transcript for each expression phenotype within R/QTL2 using a significance threshold of 2 LOD. This threshold was previously shown by permutations to be a suitable balance of type I and type II error rates (West et al. 2006, 2007).

The distribution of eQTLs across the genome was plotted using a sliding window with a 10 cM range and a step size of 1, in accordance with West et al. (2007). For the sake of

downstream analysis, transcript names were converted to the standard Arabidopsis Genome Initiative naming system. Changes in genome predictions and gene annotations led to the removal of 1,201 genes that were measured by the microarray but are not present in the current genome annotation model. This left 88,729 eQTLs. To identify potential eQTL hotspots, permutation analysis was used to estimate a global significance threshold for QTL position within each expression phenotype (control, SA, and delta). The 95th percentile value was used for the threshold to indicate significantly enriched eQTL hotspot regions.

Determining cis- vs trans-eQTL position and treatment conditionality

To determine cis-acting vs trans-acting QTL, the position of every gene in the *A. thaliana* genome was interpolated into the genetic map using the Single Feature Polymorphism marker positions from West et al. (2007). The threshold for calling an eQTL cis was if the eQTL for a transcript was located 10 cM in either direction of the transcript's genomic position based on which was the more conservative window size in favor of calling cis-eQTL according to West et al. (2007). To assess an eQTLs treatment effect, we compared the map positions for eQTLs for a transcript across the control, SA, and delta expression phenotypes. If there was an eQTL for a transcript in 1 or more of these conditions that mapped within 5.5 cM, it was considered to be the same QTL. The increased distance of 5.5 cM was chosen as a more conservative value for this measure to account for the small population size, which limits the recombination and resolution of the genomic map. The eQTL was then collapsed, and the average position was recorded as the shared map position for further analysis. Previous work has shown that the presence of SNPs within the probes for the different transcripts on the microarray

had minimal impact on the identification of eQTLs (West et al. 2006, 2007; van Leeuwen et al. 2007; Chan et al. 2010).

Calculating eQTL effect size

The effect size for each eQTL for each transcript was estimated using linear regression analysis for which the predictor variable for each model was the genotype of a given eQTL across the RILs, and the response variable was the transcript abundance across the same RILs within a specific expression phenotype (control, SA or delta). The R-squared value ascribed to the genotype from each model was then used to represent the effect size of each eQTL for each transcript in each expression phenotype (control, SA, or delta).

Sequence comparison of Bay-0 and Sha to explore potential mechanistic origins of eQTLs

To explore the molecular mechanisms that might influence the presence and magnitude of eQTLs, we used the TAIR10 Arabidopsis Col-0 reference genome, along with Bay-0 and Shahdara assemblies from Lian et al. (2024). We analyzed synteny-informed gene copy number variations using GENESPACE (Lovell et al. 2022). For each Arabidopsis accession, we quantified the number of paralogs associated with transcripts exhibiting eQTLs across different expression phenotypes. An eQTL was considered to be located in a hotspot if it fell within 5 cM of the salicylate or jasmonate hotspot regions on chromosomes 2 and 5, respectively (refer to Supplementary Table 1.1). Due to differing gene nomenclature in the Bay-0 and Shahdara assemblies, we used the Columbia reference to identify potential trans-eQTLs caused by the presence or absence of

paralogs within 10 cM of the eQTL loci. Additionally, we excluded 1,000 control eQTL transcripts from further comparative analysis after discovering they were fragments not representing whole genes in the TAIR assembly. The location of paralogs in centimorgans was then inferred using the same genetic map as previously described.

3. Results

Increased influence of genotype and plasticity on transcription in RILs

To understand the genetic and environmental factors controlling transcriptional variation, we modeled the relative contributions of genetics, treatment (plasticity/environment as the SA vs control application), and the interaction of genetics and treatment (i.e. genotype $\times$ environment/plasticity) to variation in transcript abundance. This was done independently within the Parents and RIL genotypes across transcripts. In this analysis, variation between the control and SA treatments is an estimate of the plastic response to SA treatment. For each transcript, we performed linear modeling and used the results to calculate the proportion of variance for each factor. We found distinct differences between the parents and RILs in the relative contributions of genotype, treatment, and genotype $\times$ treatment in influencing the expression of transcripts. First, most of the transcript variance in the RILs is associated with genotype and genotype-treatment while the parents' transcript variance is predominantly in the treatment term (Figure 1.1a and b). The transcripts in the parents exhibit diverse expression patterns, with some being entirely genotype-dependent, others predominantly plasticity-dependent (the difference between SA and control), and some having no contribution from any of the experimental components. The parents also exhibit higher variability in the contribution

of residuals to expression than the RIL population (Figure 1.1a). In contrast to these variable contributions in the parents, the same transcripts in the RILs were all influenced by genotype and genotype by treatment at similar levels (Figure 1.1a). The shift where some transcripts in the parents are mainly influenced by treatment to the same transcripts being mainly influenced by genotypes $\times$ treatment in the RILs suggests the potential for multiple polymorphic loci with opposing effect directions influencing SA response in the parents. Upon crossing and recombination, it is possible that the alleles at these loci shuffle to create new combinations in the RILs. If this is true, there should be transgressive segregation in SA responses (plasticity) in the RILs where the RILs have SA responses not found in the parental accessions.

Greater transcriptome variation in RILs

In light of the differences in how genotype and plasticity influence transcripts between the parents and RILs (delta of SA response), we analyzed the overall pattern of individual transcriptomes across RILs and parents to understand general trends better. We performed PCA across all individuals first using their whole transcriptome. Results from this PCA are consistent with the idea that there is transgressive segregation for the transcriptomes within the RILs. For SA and silwet treatments, along with delta (representing a change in expression across the environment, or plasticity), the RILs exhibited more transcriptome diversity than the parent lines (Figure 1.2). This pattern implies that the generation of new genotypes also generates new expression patterns that are not present within the parental genotypes. To further visualize this, we queried 4 individual transcripts classically associated with SA responses. These 4 are commonly used genes across the literature that act as biomarkers for measuring response to SA.

As such, we chose them to illustrate how expression can vary across treatment between the parental accessions and the RILs. PR1 and WRKY54 are downstream response genes typically upregulated in response to SA (Wang et al. 2006; Vlot et al. 2009). PDF1.2b produces an antimicrobial peptide in response to jasmonate, but is generally expected to be suppressed by salicylate (Manners et al. 1998). Finally, PAD3 is involved in camalexin biosynthesis and is expressed combinatorially with salicylate, jasmonate, and other signals (Ferrari et al. 2007). Plotting each gene's change in expression in response to treatment across individuals showed that for each of these genes, the plasticity responses of the RILs exceed the range set by the parents (Figure 1.3), which is not unexpected in segregating hybrid populations (Rieseberg et al. 1999). Interestingly, this led to shifts in the sign and magnitude of SA response across RILs with genes typically induced by SA showing no change or even being repressed (Figure 1.3). The increased variation in the global transcriptome and for expression response of known SA response genes between RILs and parental lines is evidence for transgressive segregation for gene expression plasticity within this population. Directly mapping loci and identifying opposing allelic effects between loci can validate this model.

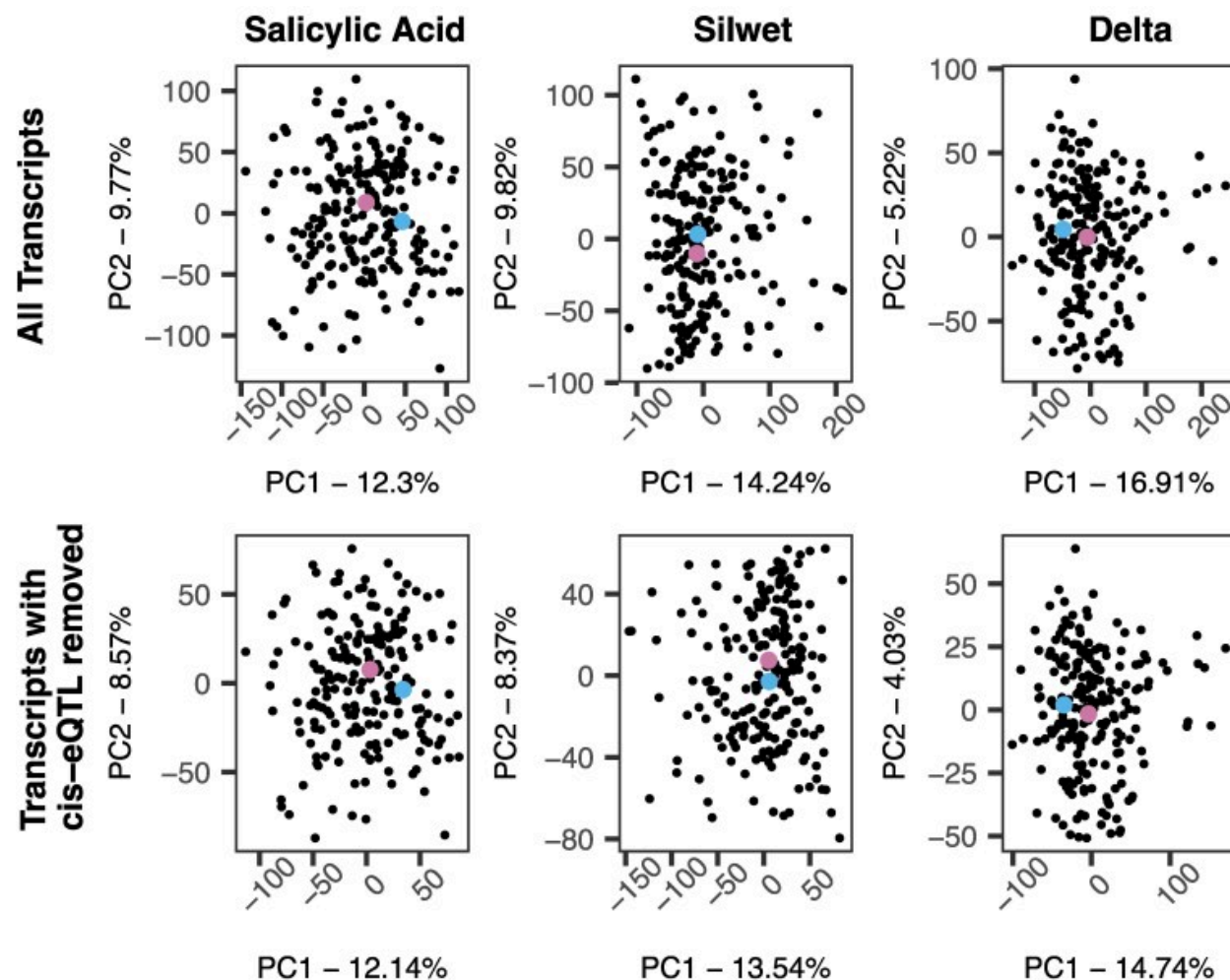


Figure 1.2. Transcriptome variation across recombinant inbred and parent lines. PCA reveals high variation in transcript abundance in RILs compared with parental lines in Bay-0 (blue) and Sha (pink). Analysis using the entire transcriptome is shown on the top row while transcriptomes without genes containing *cis*-eQTL is shown on the bottom row.

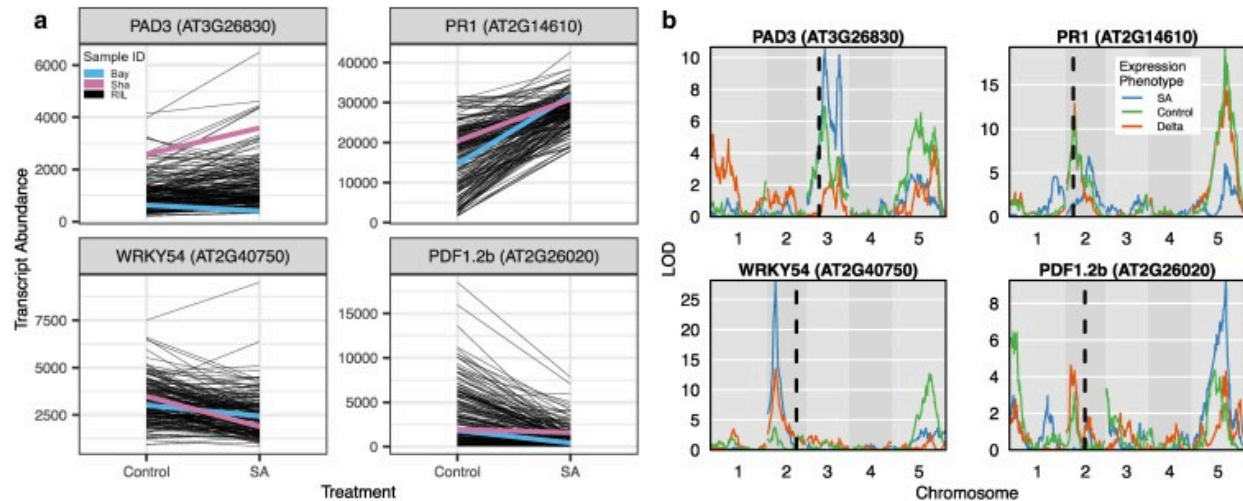


Figure 1.3. The RIL mean expression and plasticity response of classic SA response genes exceeds the range set by parent lines. a) Transcript abundance of marker genes involved in SA responses across environmental treatment groups. Lines represent abundance in individual RILs (black) and the parent lines Bay-0 (blue) and Shahdara (pink). b) The panels to the right show the corresponding eQTL map for each transcript for each expression phenotype (SA treatment in blue, control in green, and delta in orange).

Trans-regulation as a major driver of gene expression plasticity

To identify the genetic loci involved in the regulation of gene expression in the context of both environment-specific expression and plasticity, we conducted 3 quantitative trait loci analyses on the following expression phenotypes for each transcript: expression in SA, silwet (control), and delta. We mapped eQTL for each expression phenotype and used the identified eQTL to quantify cis- and trans-regulatory eQTLs and map hotspot positions. In total, we detected 8,007 cis and 30,911 trans-eQTLs for SA, 7,844 cis and 29,398 trans-eQTLs for control, and 1,278 cis and 11,291 trans-eQTLs for delta. There was an average of 1.54 trans-eQTL per transcript and 0.70 cis-eQTL per transcript across all 3 expression phenotypes. For all expression phenotypes, most eQTL were trans-acting (~80%) with a slightly higher fraction in the delta phenotype (90%), indicating that regulation of plasticity is highly polygenic (Figure 1.4a).

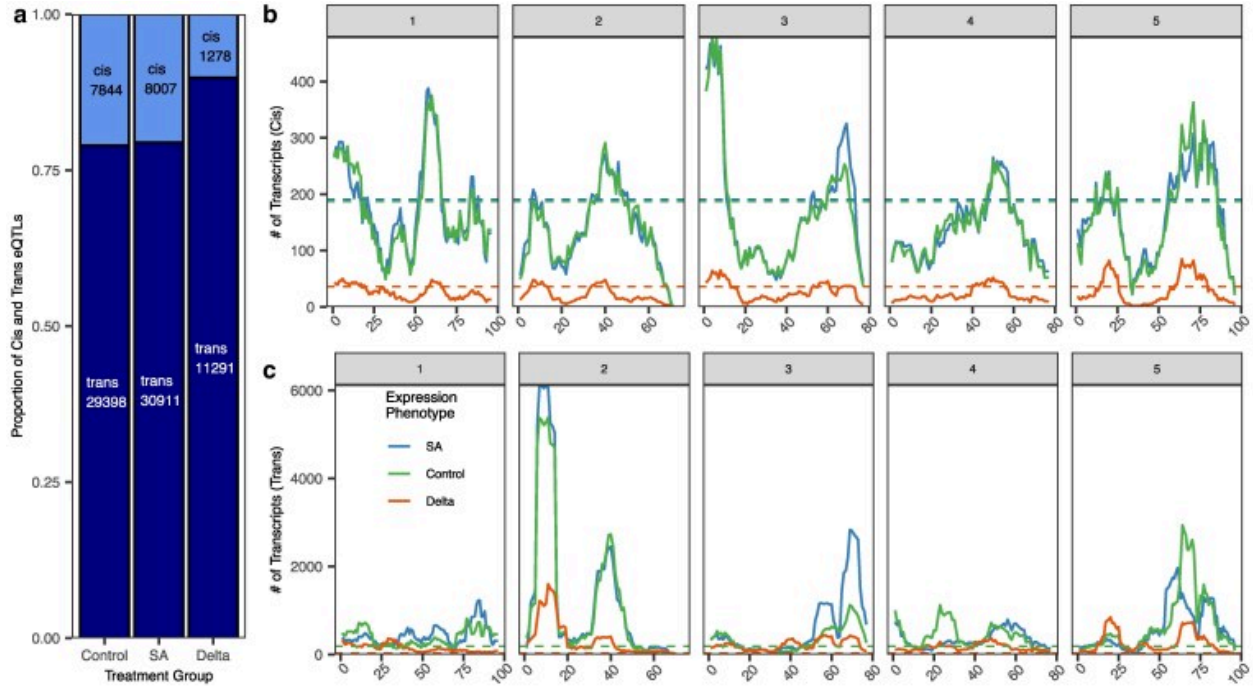


Figure 1.4 The RIL mean expression and plasticity response of classic SA response genes exceeds the range set by parent lines. a) Transcript abundance of marker genes involved in SA responses across environmental treatment groups. Lines represent abundance in individual RILs (black) and the parent lines Bay-0 (blue) and Shahdara (pink). b) The panels to the right show the corresponding eQTL map for each transcript for each expression phenotype (SA treatment in blue, control in green, and delta in orange).

The trans-eQTL for all 3 expression phenotypes (SA, control, and delta) display similar distributions across the genome, albeit at different magnitudes. The major hotspots are located on chromosomes 2, 3, and 5, although there are secondary hotspots existing on every chromosome (Figure 1.4b). This general pattern confirms the previous eQTL map (West et al. 2007), performed on the same control expression data. In contrast, the distribution of cis-eQTLs does not display any hotspot and follows the pattern of gene density across the genome. This co-localization of eQTL hotspots across the expression phenotypes implies a shared genetic architecture between plasticity (delta) and within-treatment expression variation (SA and control). It is clear that most QTLs are trans-acting, especially for plasticity, and are similarly distributed across the genome for the different expression phenotypes (Figure 1.4). The presence of multiple hot spots for

trans-eQTL, particularly for plasticity, indicates that multiple loci influence a large number of transcripts. Supporting the observation of transgressive segregation, these trans hotspots often have alleles with opposing directional effects for the same transcript (Figure 1.3a). This suggests that Bay and Sha parents have variation at a set of loci controlling SA responses and that the allelic variation is balanced such that the parents have the same transcriptome responses. When the allelic variation at these loci is shuffled, the parents can have dramatically different SA responses. Given that the Bay and Sha parents were randomly chosen with respect to SA signaling, the presence of transgressive variation suggests that Arabidopsis accessions may hide underlying genetic variation in SA responses more broadly (Koornneef et al. 2008).

Plasticity is largely controlled by trans-regulatory loci of small to moderate effect size

We sought to characterize the effect size and relative abundance of cis and trans-eQTLs existing in and across the 3 expression phenotypes (control, SA, and delta). To estimate the phenotypic effects of each eQTL, we used a linear model for every eQTL detected for each transcript in each expression phenotype. We used the proportion of variance explained (R-squared) by genotype to estimate the eQTL's effect size. To not double count eQTLs for transcripts, we compared eQTLs for each transcript across the 3 expression phenotypes and grouped them as a single eQTL if 2 eQTLs regulating the same transcript existed within 5.5 cM of each other across the 3 expression phenotypes. This showed that eQTLs influencing plasticity are mostly of low to moderate effect size, both when the eQTL is exclusive to plasticity response, and when the eQTL is shared across all expression phenotypes. Additionally, eQTLs that influenced expression plasticity (delta, SA-delta, control-delta, and SA-control-delta)

were less frequent than eQTLs found for only the non-plasticity expression phenotypes. Finally, the plasticity-associated eQTLs had greater proportions of trans- to cis-QTL, and low effect sizes compared with eQTLs that solely affected the accumulation in SA, control, or SA-control (Figure 1.5; Supplementary Figure 1.3).

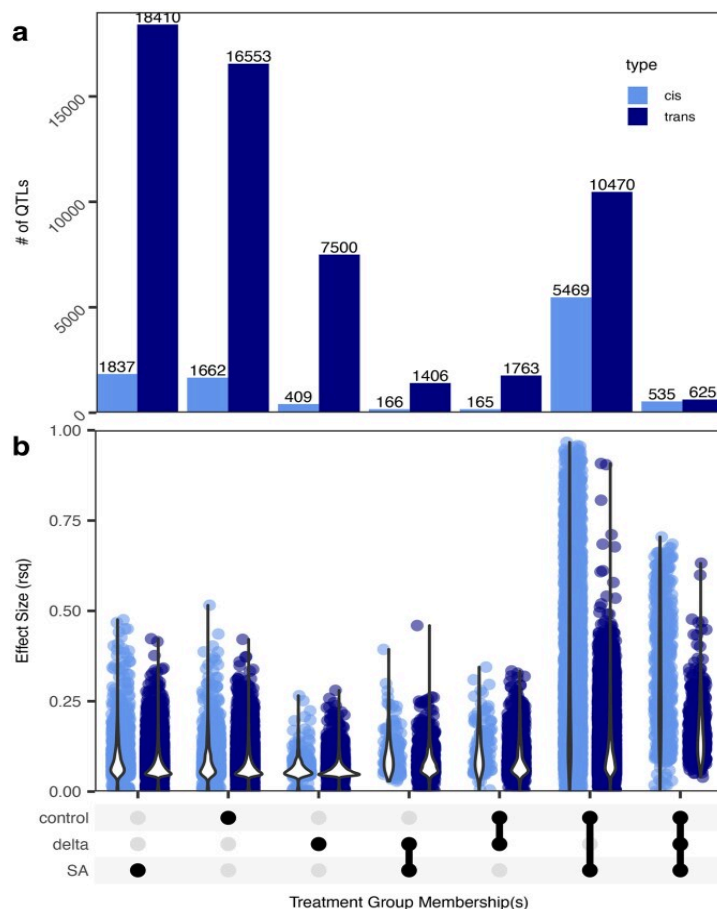


Figure 1.5. Quantity and effect size of *cis* and *trans*-eQTL according to membership across treatment groups. *Cis* (dark gray) and *trans* (light gray) eQTLs present in each of the 3 treatments. a) Number of eQTLs present in SA, silwet, delta, or a combination of treatment groups. b) Effect size of eQTLs according to treatment group membership. Effect size is represented by an *R*-squared value. In cases where an eQTL exists in more than 1 group, the mean *R*-squared value is plotted.

It was initially puzzling that there were eQTL that linked to a transcript in SA but not in control, or vice versa, yet had no signal upon a plasticity/delta effect. However, plotting the effect of these eQTL on specific transcripts showed that they are likely a power-to-detect issue and likely have similar effects in both conditions and were under the significance threshold in the other condition (Figure 1.6). In agreement with this, the eQTL effects for a given transcript showed the same magnitude and direction of change

in expression across SA and control, indicating similar effects in both SA and control and that the original observation was caused by false negative eQTL. In contrast, transcript eQTLs linked to SA-delta and control-delta effects had much stronger plasticity effects with a greater divergence in the slope of the response between the 2 genotypes (Figure 1.6).

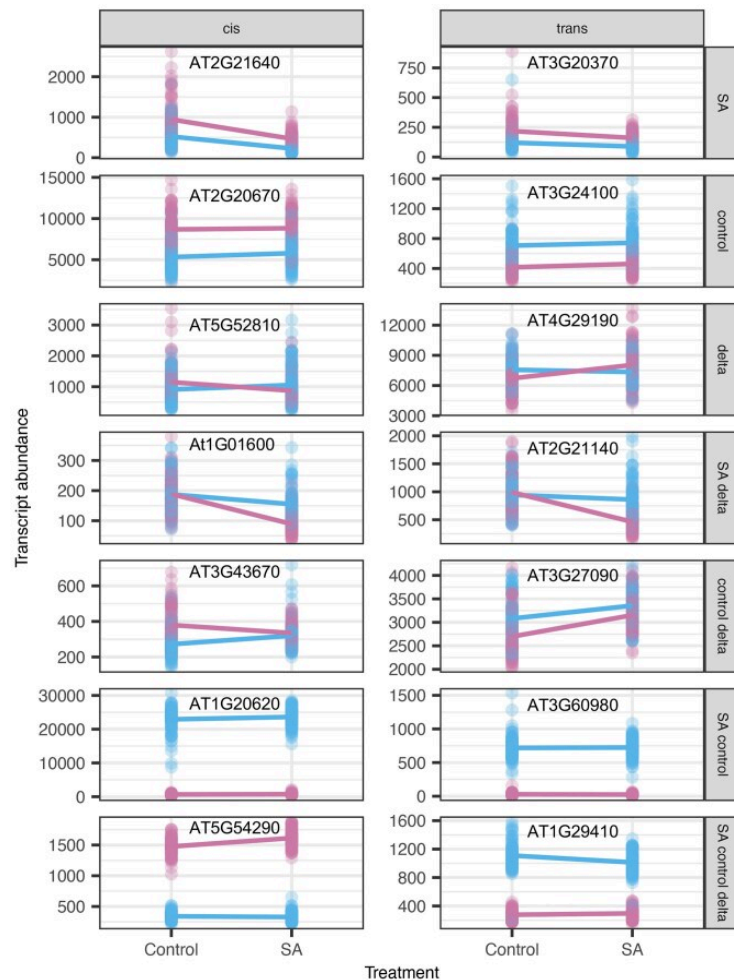


Figure 1.6. Expression across environments of high-effect size eQTL according to treatment group membership. The transcript abundance for the highest effect size eQTL across type (*cis* or *trans*) and treatment group combination was plotted. Individual genotypes are colored according to the genotype (Bayreuth in blue or Shahdara in pink) at the given eQTL. Mean expression across individuals within a given genotype is represented by a solid line.

Large-effect trans-eQTL

While a majority of trans-eQTLs are small-effect size, we did identify a number of large-effect trans-eQTL, and we wanted to understand what these may represent (Figure 1.7). To better resolve these patterns and potential underlying mechanisms, we

grouped all the eQTLs by their cis/trans status, presence in a hotspot, and presence of a paralog at the locus. One observation from this was that eQTLs with high-effect sizes often involve genes with multiple paralogs or copy number variation between accessions, especially those shared across multiple expression phenotypes. This agrees with previous observations that large-effect cis-eQTL are often associated with structural variation at the gene and these show large expression differences (Figure 1.5; Li et al. 2013). The same pattern also held for large-effect trans-eQTLs where there was a paralog to the transcript mapping to the trans-eQTL. For example, the transcript AT1G2490 has a high-effect trans-eQTL across all 3 phenotypes, likely due to a paralog at the trans-eQTL that is present in Bay-0 but absent in Sha (Supplementary Table 1.1). Similarly, AT3G60980 has a high-effect trans-eQTL in SA and control, where Bay-0 retains an extra paralog at the trans position, but Sha has lost the copy (Supplementary Table 1.1).

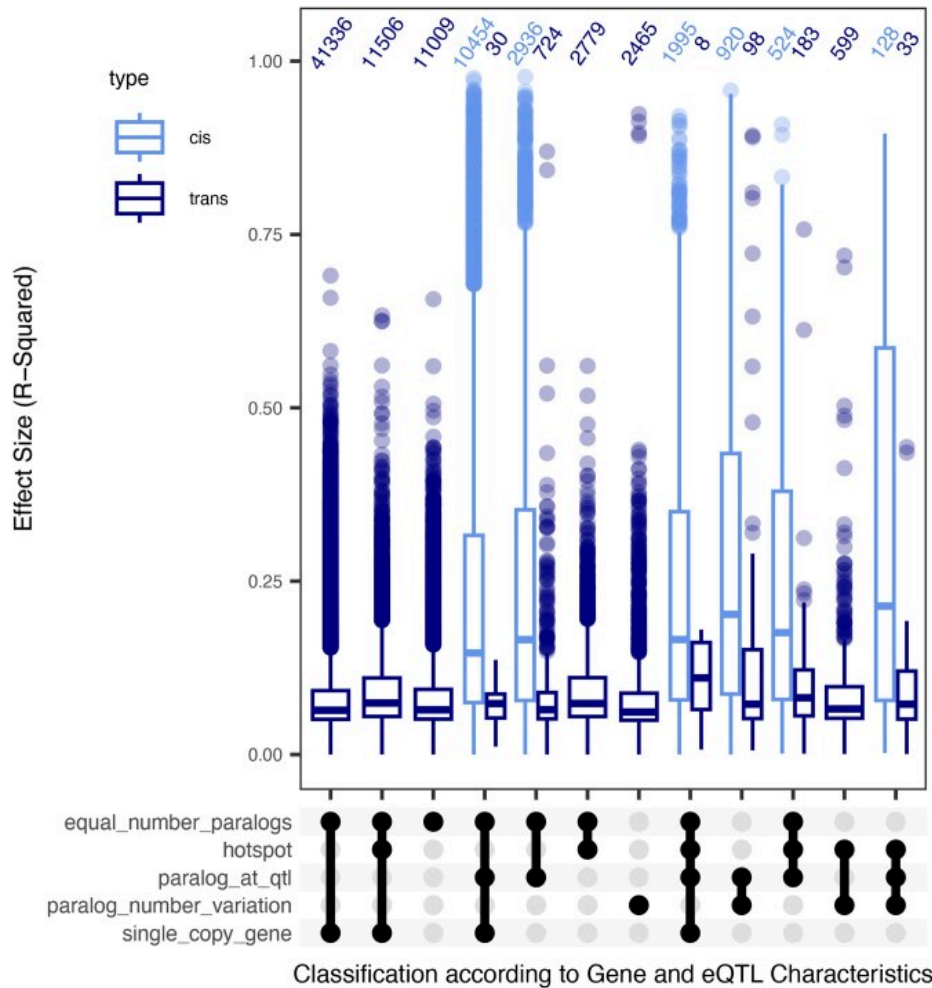


Figure 1.7. Distribution of QTL effect sizes according to genomic attributes of the transcript. Effect size of *cis* and *trans* eQTLs categorized according to the location of eQTL as well as paralog information of the transcript they control.

Querying this list further identified genes where *cis* structural variation in the gene family at 1 locus links to a *trans* effect at the other members of the gene family. There are several possible mechanisms for this. First, it may be cross-recognition of the paralog by the microarray leading to a *trans* signal. Another option is that in the phosphoribosylanthranilate isomerase gene family in *Arabidopsis*, paralog variation at 1 locus creates a silencing mechanism decreasing expression of the other gene family members, as previously described (Bender 2004). Additionally, some genes are known

to move laterally in the genome creating an appearance of a trans signal (West et al. 2006). There were a number of large-effect trans-eQTLs that have no connection to paralogues or paralogue variation (Figure 1.7). This showed that trans hotspots contain large-effect eQTLs affecting a number of genes. These were distributed between trans-eQTLs that were both within and outside of genomic hotspots. Interestingly, the presence of large-effect trans-eQTLs within hotspots associated with plasticity suggests that pleiotropic trans-eQTLs are not obligated to have small effects. Assessing some of these genes showed that they included key components of the SA response and regulatory machinery (Supplementary Figures 1.1 and 1.2; Supplementary Table 1.1). Thus, while the preponderance of effects at these loci is modest to small, there can be large trans-eQTLs that modulate the regulatory machinery directly influencing the plasticity.

4. Discussion

Transcriptomic plasticity to SA treatment within the Bay $\times$ Sha-0 RIL population appears to be highly influenced by trans-regulatory eQTL. These plasticity eQTLs, because they are predominantly trans, tend to have a smaller effect than the cis-eQTLs (Supplementary Figure 1.3). GWA analyses are typically underpowered to find small-effect loci, suggesting a potential for ascertainment bias in expression plasticity studies against these plasticity trans-eQTL. GWA is typically powered to identify moderate to larger effect loci, which in this case are cis-eQTL largely not associated with plasticity. This power bias could lead to an under-detection of trans-eQTL in GWA or other natural variation studies. Furthermore, previous studies have demonstrated that trans-regulatory variation can influence adaptive evolution within species (Rhoné et al.

2017; Lopez-Arboleda et al. 2021). For example in *Arabidopsis*, genes that show subpopulation-specific association are more likely to be exclusively trans-regulated (Lopez-Arboleda et al. 2021). This raises the potential that GWA studies have significant amounts of undetected trans-regulatory loci influencing gene expression and suggests that incorporating structured populations into plasticity studies will be key to better resolving the regulatory landscape of plasticity variation and helping to develop better models of how plasticity variation may evolve.

SA signaling is considered to be a highly conserved core component of plant responses to pathogen attack. This model was supported by previous studies identifying minimal variation in SA responses across *Arabidopsis* accessions (Koornneef et al. 2008). In agreement, we observed minimal variation in the SA transcriptome response between the Bay and Sha parents (Figures 1.1 and 1.2). In contrast, the RIL progeny displayed extensive genotype $\times$ SA treatment variation across nearly the entire transcriptome, and this mapped to a defined set of trans-eQTL hotspots (Figures 1.1 and 1.4). These hotspots pleiotropically influence the SA response of 100–1,000 s of transcripts (Figure 1.4). The population size likely prevents any ability to see transgressive variation within cis variants due to the rarity of within-promoter recombination. It should be noted that the potential for cis-regulatory elements to generate transgressive phenotypes has been demonstrated in tomatoes with the use of CRISPR/Cas-9 technology, (Rodríguez-Leal et al. 2017). Because these hotspot loci had opposing directional effects on the response to SA, they created a situation whereby the parental accessions had largely the same SA response while the RIL progeny showed extensive transgressive

segregation in SA response (Figure 1.3). This shows that the minimal phenotypic plasticity to SA in *Arabidopsis* accessions hides genetic variation underlying the SA response, further supporting the need for structured populations in assessing plasticity variation within a species.

Previous research has observed that transgression is relatively common across plants and influences a diverse set of traits. The likelihood for transgressive segregation can increase for traits under constraining selection like a key defense response and also within selfing plant populations such as *Arabidopsis* (Rieseberg et al. 1999). Population structure limitations on gene flow can also create partitions that allow for the evolution of transgressive traits. Differentiating between these models would require the development of a number of structured populations across the species to better sample the potential for transgression to be a common aspect of typically assumed conserved signaling pathways. This is probably necessary to fully understand the genetic architecture of plasticity because extensive transgressive segregation across *Arabidopsis* accessions would influence the interpretation of GWA (Jimenez-Gomez et al. 2011). Transgressive variation and constraining selection would hide causal genetic variation that influences the trait and may be under locally adaptive selective pressures.

Observations from this structured population suggest that GWA studies of transcriptomic plasticity may need careful assessment. (1) Plasticity eQTL are biased toward small-effect trans loci which could go undetected in GWA studies. (2) Large-effect plasticity trans-eQTL can be linked to paralog variants, suggesting they are

actually cis in causation. (3) Plasticity can display transgressive segregation that can constrain GWA ability to identify causal loci. While this is from a single structured population, it was randomly chosen with regard to the traits in question and the parents represent a broad sampling of Arabidopsis genetic diversity. While it will take additional structured populations to assess this potential more broadly, it does suggest there is a need to conduct these structured population studies more broadly to better understand the results from GWA collections.

5. Declarations

Data availability

The microarray data set used in this study has been deposited at EBI ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) under nos. E-TABM-126 and E-TABM-224. Detailed information on the trans-eQTLs with the highest effect sizes is found in Supplementary Table 1.1.

Code availability

Code for this research is maintained at https://github.com/mariele-lensink/plasticity_eQTL.

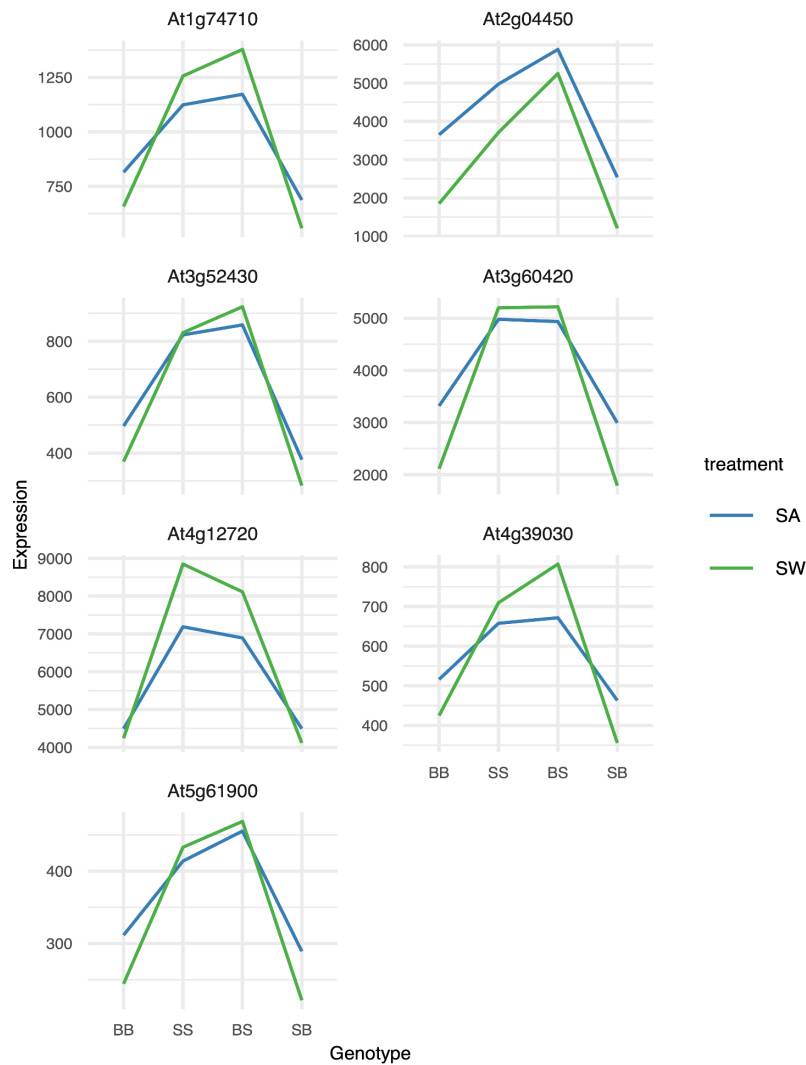
Funding

Funding for this work was provided to DK by US National Science Foundation awards IOS 2434545 and MCB 1906486.

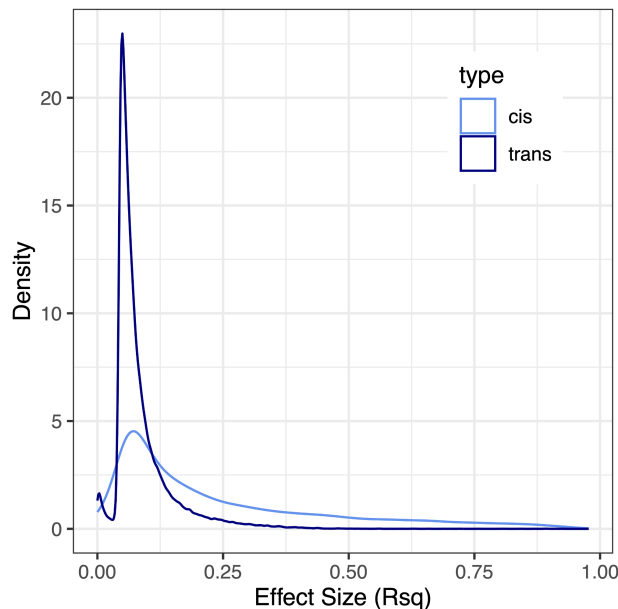
6. Supplemental Figures and Tables



Supplementary Figure 1.1.
Expression of transcripts with SA hotspot eQTLs on Chromosome 2 under salicylate and silwet treatment for parental and recombinant genotypes The x axis is the genotype of RILs at the location of the SA and JA hotspots. For example, BB represents RILs with the Bay-0 genotype at both locations, where SB is RILs with the Sha genotype at the chromosome 2 hotspot and Bay-0 at the chromosome 5 hotspot.



Supplementary Figure 1.2.
Expression of transcripts with
JA hotspot eQTLs on
Chromosome 5 under
salicylate and silwet treatment
for parental and recombinant
genotypes The x axis is the
 genotype of RILs at the
 location of the SA and JA
 hotspots. For example, BB
 represents RILs with the Bay-0
 genotype at both locations,
 where SB is RILs with the Sha
 genotype at the chromosome 2
 hotspot and Bay-0 at the
 chromosome 5 hotspot.



Supplementary Figure 1.3. Density distribution
of cis and trans eQTL effect sizes.

Effect size is represented as r-square value for cis (dark blue) and trans (light blue) eQTLs.

7. References

- Ballinger MA, Mack KL, Durkin SM, Riddell EA, Nachman MW. 2023. Environmentally robust cis-regulatory changes underlie rapid climatic adaptation. *Proc Natl Acad Sci U S A*. 120(39):e2214614120. doi: 10.1073/pnas.2214614120. - [DOI](#) - [PMC](#) - [PubMed](#)
- Bender J. 2004. DNA methylation of the endogenous PAI genes in *Arabidopsis*. *Cold Spring Harb Symp Quant Biol*. 69(0):145–154. doi: 10.1101/sqb.2004.69.145. - [DOI](#) - [PubMed](#)
- Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen S, Yandell BS, Churchill GA. 2019. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. *Genetics*. 211(2):495–502. doi: 10.1534/genetics.118.301595. - [DOI](#) - [PMC](#) - [PubMed](#)
- Brown KE, Kelly JK. 2022. Genome-wide association mapping of transcriptome variation in *Mimulus guttatus* indicates differing patterns of selection on cis-versus trans-acting mutations. *Genetics*. 220(1):iyab189. doi: 10.1093/genetics/iyab189. - [DOI](#) - [PMC](#) - [PubMed](#)
- Chan EKF, Rowe HC, Kliebenstein DJ. 2010. Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics*. 185(3):991–1007. doi: 10.1534/genetics.109.108522. - [DOI](#) - [PMC](#) - [PubMed](#)
- Dardiry M, Eberhard G, Witte H, Rödelsperger C, Lightfoot JW, Sommer RJ. 2023. Divergent combinations of cis-regulatory elements control the evolution of phenotypic plasticity. *PLoS Biol*. 21(8):e3002270. doi: 10.1371/journal.pbio.3002270. - [DOI](#) - [PMC](#) - [PubMed](#)

- de Nadal E, Ammerer G, Posas F. 2011. Controlling gene expression in response to stress. *Nat Rev Genet.* 12(12):833–845. doi: 10.1038/nrg3055. - [DOI](#) - [PubMed](#)
- Ferrari S, Galletti R, Denoux C, De Lorenzo G, Ausubel FM, Dewdney J. 2007. Resistance to *Botrytis cinerea* induced in *Arabidopsis* by elicitors is independent of salicylic acid, ethylene, or jasmonate signaling but requires PHYTOALEXIN DEFICIENT3. *Plant Physiol.* 144(1):367–379. doi: 10.1104/pp.107.095596. - [DOI](#) - [PMC](#) - [PubMed](#)
- Fox J, Weisberg S. 2018. *An R Companion to Applied Regression*. Thousand Oaks (CA): Sage Publications.
- Ghalambor CK, McKay JK, Carroll SP, Reznick DN. 2007. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. *Funct Ecol.* 21(3):394–407. doi: 10.1111/j.1365-2435.2007.01283.x. - [DOI](#)
- Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet.* 13(2):135–145. doi: 10.1038/nrg3118. - [DOI](#) - [PMC](#) - [PubMed](#)
- Görling HH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet.* 69(6):1357–1369. doi: 10.1086/324471. - [DOI](#) - [PMC](#) - [PubMed](#)
- He F, Arce AL, Schmitz G, Koornneef M, Novikova P, Beyer A, De Meaux J. 2016. The footprint of polygenic adaptation on stress–responsive cis-regulatory divergence in the *Arabidopsis* genus. *Mol Biol Evol.* 33(8):2088–2101. doi: 10.1093/molbev/msw096. - [DOI](#) - [PubMed](#)

- He F, Steige KA, Kovacova V, Göbel U, Bouzid M, Keightley PD, Beyer A, De Meaux J. 2021. Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. *Nat Commun.* 12(1):3376. doi: 10.1038/s41467-021-23558-2. - [DOI](#) - [PMC](#) - [PubMed](#)
- Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution.* 61(5):995–1016. doi: 10.1111/j.1558-5646.2007.00105.x. - [DOI](#) - [PubMed](#)
- Jimenez-Gomez JM, Corwin JA, Joseph B, Maloof JN, Kliebenstein DJ. 2011. Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS Genet.* 7(9):e1002295. doi: 10.1371/journal.pgen.1002295. - [DOI](#) - [PMC](#) - [PubMed](#)
- Josephs EB, Stinchcombe JR, Wright SI. 2017. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytol.* 214(1):21–33. doi: 10.1111/nph.14410. - [DOI](#) - [PubMed](#)
- Koornneef A, Leon-Reyes A, Ritsema T, Verhage A, Den Otter FC, Van Loon LC. 2008. Kinetics of salicylate-mediated suppression of jasmonate signaling reveal a role for redox modulation. *Plant Physiol.* 147(3):1358–1368. doi: 10.1104/pp.108.121392. - [DOI](#) - [PMC](#) - [PubMed](#)
- Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods.* 9(1):29. doi: 10.1186/1746-4811-9-29. - [DOI](#) - [PMC](#) - [PubMed](#)

- Lasky JR, Des Marais DL, Lowry DB, Povolotskaya I, McKay JK, Richards JH, Keitt TH, Juenger TE. 2014. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. *Mol Biol Evol*. 31(9):2283–2296. doi: 10.1093/molbev/msu170. - [DOI](#) - [PMC](#) - [PubMed](#)
- Levis NA, Pfennig DW. 2016. Evaluating “plasticity-first” evolution in nature: key criteria and empirical approaches. *Trends Ecol Evol*. 31(7):563–574. doi: 10.1016/j.tree.2016.03.012. - [DOI](#) - [PubMed](#)
- Li L, Petsch K, Shimizu R, Liu S, Xu WW, Ying K, Yu J, Scanlon MJ, Schnable PS, Timmermans MC, et al. 2013. Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genet*. 9(1):e1003202. doi: 10.1371/journal.pgen.1003202. - [DOI](#) - [PMC](#) - [PubMed](#)
- Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC, et al. 2006. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet*. 2(12):e222. doi: 10.1371/journal.pgen.0020222. - [DOI](#) - [PMC](#) - [PubMed](#)
- Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R. 2024. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. *Nat Genet*. 56(5):982–991. doi: 10.1038/s41588-024-01715-9. - [DOI](#) - [PMC](#) - [PubMed](#)
- Liu X, Li YI, Pritchard JK. 2019. Trans effects on gene expression can drive omnigenic inheritance. *Cell*. 177(4):1022–1034.e6. doi: 10.1016/j.cell.2019.04.014. - [DOI](#) - [PMC](#) - [PubMed](#)

- Lopez-Arboleda WA, Reinert S, Nordborg M, Korte A. 2021. Global genetic heterogeneity in adaptive traits. *Mol Biol Evol.* 38(11):4822–4831. doi: 10.1093/molbev/msab208. - [DOI](#) - [PMC](#) - [PubMed](#)
- Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F. 2002. Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis*. *Theor Appl Genet.* 104(6):1173–1184. doi: 10.1007/s00122-001-0825-9. - [DOI](#) - [PubMed](#)
- Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife.* 11:e78526. doi: 10.7554/eLife.78526. - [DOI](#) - [PMC](#) - [PubMed](#)
- Mai NT, Mai CD, Van Nguyen H, Le KQ, Duong LV, Tran TA, To HT. 2021. Discovery of new genetic determinants of morphological plasticity in rice roots and shoots under phosphate starvation using GWAS. *J Plant Physiol.* 257:153340. doi: 10.1016/j.jplph.2020.153340. - [DOI](#) - [PubMed](#)
- Manners JM, Penninckx IA, Vermaere K, Kazan K, Brown RL, Morgan A, Maclean DJ, Curtis MD, Cammue BP, Broekaert WF. 1998. The promoter of the plant defensin gene PDF1.2 from *Arabidopsis* is systemically activated by fungal pathogens and responds to methyl jasmonate but not to salicylic acid. *Plant Mol Biol.* 38(6):1071–1080. doi: 10.1023/A:1006070413843. - [DOI](#) - [PubMed](#)
- Marigorta UM, Rodríguez JA, Gibson G, Navarro A. 2018. Replicability and prediction: lessons and challenges from GWAS. *Trends Genet.* 34(7):504–517. doi: 10.1016/j.tig.2018.03.005. - [DOI](#) - [PMC](#) - [PubMed](#)

- Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, Dermitzakis ET, Stone EA, Jensen JD, Mackay TF, et al. 2012. Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet.* 8(11):e1003055. doi: 10.1371/journal.pgen.1003055. - [DOI](#) - [PMC](#) - [PubMed](#)
- Metzger BPH, Wittkopp PJ, Coolon JD. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among *Saccharomyces* species. *Genome Biol Evol.* 9(4):843–854. doi: 10.1093/gbe/evx035. - [DOI](#) - [PMC](#) - [PubMed](#)
- Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES. 2009. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell.* 21(8):2194–2202. doi: 10.1105/tpc.109.068437. - [DOI](#) - [PMC](#) - [PubMed](#)
- Peng Y, Yang J, Li X, Zhang Y. 2021. Salicylic acid: biosynthesis and signaling. *Annu Rev Plant Biol.* 72(1):761–791. doi: 10.1146/annurev-arplant-081320-092855. - [DOI](#) - [PubMed](#)
- Pigliucci M, Murren CJ, Schlichting CD. 2006. Phenotypic plasticity and evolution by genetic assimilation. *J Exp Biol.* 209(12):2362–2367. doi: 10.1242/jeb.02070. - [DOI](#) - [PubMed](#)
- Pruunsild P, Bading H. 2019. Shaping the human brain: evolutionary cis-regulatory plasticity drives changes in synaptic activity-controlled adaptive gene expression. *Curr Opin Neurobiol.* 59:34–40. doi: 10.1016/j.conb.2019.04.003. - [DOI](#) - [PubMed](#)

- Redman JC, Haas BJ, Tanimoto G, Town CD. 2004. Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant J.* 38(3):545–561. doi: 10.1111/j.1365-313X.2004.02061.x. - [DOI](#) - [PubMed](#)
- Rhoné B, Mariac C, Couderc M, Berthouly-Salazar C, Ousseini IS, Vigouroux Y. 2017. No excess of cis-regulatory variation associated with intraspecific selection in wild pearl millet (*Cenchrus americanus*). *Genome Biol Evol.* 9(2):388–397. doi: 10.1093/gbe/evx004. - [DOI](#) - [PMC](#) - [PubMed](#)
- Rieseberg LH, Archer MA, Wayne RK. 1999. Transgressive segregation, adaptation and speciation. *Heredity (Edinb).* 83(4):363–372. doi: 10.1038/sj.hdy.6886170. - [DOI](#) - [PubMed](#)
- Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. 2017. Engineering quantitative trait variation for crop improvement by genome editing. *Cell.* 171(2):470–480.e8. doi: 10.1016/j.cell.2017.08.030. - [DOI](#) - [PubMed](#)
- Schlichting CD, Smith H. 2002. Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. *Evol Ecol.* 16(3):189–211. doi: 10.1023/A:1019624425971. - [DOI](#)
- Sham PC, Purcell SM. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet.* 15(5):335–346. doi: 10.1038/nrg3706. - [DOI](#) - [PubMed](#)
- Spoel SH, Dong X. 2024. Salicylic acid in plant immunity and beyond. *Plant Cell.* 36(5):1451–1464. doi: 10.1093/plcell/koad329. - [DOI](#) - [PMC](#) - [PubMed](#)
- Sun G, Yu H, Wang P, Lopez-Guerrero M, Mural RV, Mizero ON, Grzybowski M, Song B, Van Dijk K, Schachtman DP, et al. 2023. A role for heritable transcriptomic

- variation in maize adaptation to temperate environments. *Genome Biol.* 24(1):55. doi: 10.1186/s13059-023-02891-3. - [DOI](#) - [PMC](#) - [PubMed](#)
- Takou M, Bellis ES, Lasky J. 2024. Predicting gene expression responses to environment in *Arabidopsis thaliana* using natural variation in DNA sequence. bioRxiv 591174. 10.1101/2024.04.25.591174, preprint: not peer reviewed. - [DOI](#)
- Vande Zande P, Hill MS, Wittkopp PJ. 2022. Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. *Science*. 377(6601):105–109. doi: 10.1126/science.abj7185. - [DOI](#) - [PMC](#) - [PubMed](#)
- Van Leeuwen H, Kliebenstein DJ, West MA, Kim K, Van Poecke R, Katagiri F, Michelmore RW, Doerge RW, St. Clair DA. 2007. Natural variation among *Arabidopsis thaliana* accessions for transcriptome response to exogenous salicylic acid. *Plant Cell*. 19(7):2099–2110. doi: 10.1105/tpc.107.050641. - [DOI](#) - [PMC](#) - [PubMed](#)
- Vlot AC, Dempsey DA, Klessig DF. 2009. Salicylic acid, a multifaceted hormone to combat disease. *Annu Rev Phytopathol*. 47(1):177–206. doi: 10.1146/annurev.phyto.050908.135202. - [DOI](#) - [PubMed](#)
- Walter GM, Clark J, Terranova D, Cozzolino S, Cristaudo A, Hiscock SJ, Bridle J. 2023. Hidden genetic variation in plasticity provides the potential for rapid adaptation to novel environments. *New Phytol*. 239(1):374–387. doi: 10.1111/nph.18744. - [DOI](#) - [PubMed](#)
- Wang D, Amornsiripanitch N, Dong X. 2006. A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants.

- PLoS Pathog. 2(11):e123. doi: 10.1371/journal.ppat.0020123. - [DOI](#) - [PMC](#) - [PubMed](#)
- West MA, Kim K, Kliebenstein DJ, Van Leeuwen H, Michelmore RW, Doerge RW, St. Clair DA. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics. 175(3):1441–1450. doi: 10.1534/genetics.106.064972. - [DOI](#) - [PMC](#) - [PubMed](#)
- West MA, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, Clair DA, Michelmore RW. 2006. High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. Genome Res. 16(6):787–795. doi: 10.1101/gr.5011206. - [DOI](#) - [PMC](#) - [PubMed](#)
- Xu C, Song LY, Zhou Y, Ma DN, Ding QS, Guo ZJ, Li J, Song SW, Zhang LD, Zheng HL. 2023. Integration of eQTL and GWAS analysis uncovers a genetic regulation of natural ionomic variation in Arabidopsis. Plant Cell Rep. 42(9):1473–1485. doi: 10.1007/s00299-023-03042-5. - [DOI](#) - [PubMed](#)
- Zan Y, Shen X, Forsberg SKG, Carlborg Ö. 2016. Genetic regulation of transcriptional variation in natural *Arabidopsis thaliana* accessions. G3 (Bethesda). 6(8):2319–2328. doi: 10.1534/g3.116.030874. - [DOI](#) - [PMC](#) - [PubMed](#)
- Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res. 21(5):725–733. doi: 10.1101/gr.115337.110. - [DOI](#) - [PMC](#) - [PubMed](#)
- Zhang Y, Li X. 2019. Salicylic acid: biosynthesis, perception, and contributions to plant immunity. Curr Opin Plant Biol. 50:29–36. doi: 10.1016/j.pbi.2019.02.004. - [DOI](#) - [PubMed](#)

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, et al. 2007. An Arabidopsis example of association mapping in structured samples. PLoS Genet. 3(1):e4. doi: 10.1371/journal.pgen.0030004. - [DOI](#) - [PMC](#) - [PubMed](#)

Chapter 2: Approximate Bayesian Inference Supports Genic Hypomutation as a Driver of Genetic Diversity in *Arabidopsis thaliana*

Authors: Mariele Lensink, Vianney Ahn, Daniel Kliebenstein, Jeffrey Ross-Ibarra, Grey Monroe

As first author, I conceived and led the study from initial idea development through manuscript preparation. I developed the research questions and study design, conducted the simulations and analyses, interpreted the results, prepared the figures, and wrote the manuscript. The empirical genomic data used in the study were obtained from the 1001 Genomes Consortium

Abstract

Mutation rate varies across genomes due to heterogeneity in DNA damage, chromatin context, and repair activity, with important consequences for genome evolution. Across many taxa, mutation rates are lower in genic regions than in intergenic sequences. In *Arabidopsis thaliana*, experimental work has linked this pattern to spatial biases in DNA repair. This mechanistic association to functionally constrained genomic regions raises the possibility that reduced genic diversity at the population level reflects not only the removal of deleterious variants, but also a lower initial input of mutations, suggesting that mutation-rate heterogeneity itself may shape genomic patterns of variation and may be subject to evolutionary forces. However, in population-genomic data, the effects of heterogeneous mutation-rates are difficult to distinguish from those of natural selection and demography. Here, we develop a population-genetic modeling framework to investigate the consequences of intragenomic mutation-rate heterogeneity using forward evolutionary simulations with SLiM. We independently model genic and intergenic regions with distinct mutation rates and selection regimes. Using approximate Bayesian computation, we evaluate the relative support for mutation-rate variation and purifying selection as contributors to patterns of genetic variation comparable to those

observed in the Arabidopsis 1,001 Genomes dataset. Through this inference framework, we recover biologically plausible genic and intergenic mutation-rate estimates, and direct inference of their ratio strongly supports lower mutation rates in genic sequences, providing strong support for mutation-rate heterogeneity as an important contributor to population-genomic patterns of variation. Our findings highlight the importance of explicitly modeling both heterogeneous mutation processes and selective processes in population-genetic analyses, as variation in mutation rates can meaningfully influence inferred evolutionary patterns. Placed in the context of mutation accumulation experiments and experimentally characterized DNA repair mechanisms in plants, these results frame mutation-rate heterogeneity as shaping genome evolution and motivate integrative approaches that connect molecular mechanisms of mutation and repair with population-genetic patterns of diversity.

1. Introduction

Mutation, repair, and spatial heterogeneity

Mutation is the source of genetic variation and provides the raw material for evolution. However, mutations do not arise uniformly across the genome (Moore et al., 2021; Ness et al., 2015; Quiroz et al., 2023; Schuster-Böckler & Lehner, 2012; Weng et al., 2019). DNA is continuously exposed to damage arising from endogenous cellular processes, such as replication errors and spontaneous chemical changes, and from exogenous environmental sources such as ultraviolet radiation and elevated temperature (Belfield et al., 2021; Britt, 2004; Hitomi et al., 2012; Lu et al., 2021; Tuteja et al., 2009; Viguera et al., 2001; Watanabe et al., 2006; Willing et al., 2016; Zou et al., 2021). The frequency

and genomic distribution of this damage, as well as the likelihood that a lesion ultimately becomes a mutation, vary with local sequence and chromatin context. Regional mutation rates have been associated with nucleosome organization, histone modifications, and DNA methylation (X. Chen et al., 2012; C. Li & Luscombe, 2020; Monroe et al., 2022; Schuster-Böckler & Lehner, 2012; Yan et al., 2021; Fryxell & Moon, 2005; Monroe et al., 2022; Weng et al., 2019; Zhu et al., 2021). Replication timing and transcriptional activity also contribute to mutation-rate variation (C. Chen et al., 2017; Stamatoyannopoulos et al., 2009; Staunton et al., 2023; Mugal et al., 2009; Polak & Arndt, 2008; Staunton et al., 2023). Variation in the activity and targeting of DNA repair pathways can further alter the probability that damage is corrected before becoming a mutation (Belfield et al., 2018; Boyko et al., 2006; Giustozzi et al., 2022; Huang et al., 2018; Kimura et al., 2004; F. Li et al., 2013; Oztas et al., 2018; Quiroz et al., 2024; Supek & Lehner, 2015, 2017; Suter et al., 1997). Local mutation-rate heterogeneity has long been hypothesized to evolve in ways that reduce the deleterious mutation input in functionally important genomic regions (Martincorena & Luscombe, 2013). Consistent with this hypothesis, gene bodies in plants accumulate fewer mutations than surrounding intergenic regions, a pattern linked in part to preferential recruitment and activity of DNA repair machinery in functionally constrained genic regions (Monroe et al., 2022, 2025; Quiroz et al., 2024). If mutation rates vary systematically between genomic compartments, then patterns of standing polymorphism may reflect both heterogeneous mutation input and natural selection.

The mutation-selection identifiability problem

Ultimately, the genome-wide patterns of polymorphism are shaped by multiple interacting evolutionary processes: where mutations arise and persist following DNA repair, how purifying selection removes deleterious variants, recombination, and population structure (Charlesworth et al., 1993; Corbett-Detig et al., 2015; Ellegren & Galtier, 2016). Differences in the rates at which variants arise can influence evolutionary outcomes even when those variants are subsequently subject to natural selection (Stoltzfus & Yampolsky, 2009; Yampolsky & Stoltzfus, 2001). Within this broader context, reduced genic diversity in gene bodies can arise from lower genic mutation rates or stronger purifying selection removing deleterious variants, or both simultaneously (Charlesworth & Jensen, 2022; Liu & Zhang, 2022; Monroe et al., 2022). Critically, mutation-rate heterogeneity and purifying selection can generate similar population-genetic signatures (Charlesworth & Jensen, 2022; Monroe et al., 2022), therefore creating an identifiability problem. Observed reductions in genic polymorphism cannot easily be specifically attributed to mutation or selection. Many population-genetic inference frameworks implicitly assume that μ is constant. If this assumption is violated, mutation-rate heterogeneity may be incorrectly absorbed into estimates of selection or demographic history. Furthermore, when mutation-rate heterogeneity is not represented, it can increase uncertainty or lead to inaccurate estimates of demographic history and the distribution of fitness effects (Soni et al., 2024).

Existing approaches and their limitations

Determining whether genomic patterns of diversity arise from variation in mutation input, natural selection, or both requires approaches that can distinguish the generation of

new mutations from their subsequent evolutionary filtering. Three broad strategies can be used to study these processes: direct measurement of mutation through mutation-accumulation experiments, inference from patterns of standing polymorphism and divergence, and simulation-based approaches that allow mutation and selection to vary simultaneously.

Mutation-Accumulation (MA) Experiments

Mutation-accumulation (MA) experiments estimate mutation processes by propagating replicate lines through repeated severe bottlenecks, thereby reducing the efficacy of natural selection and allowing most spontaneous mutations to accumulate through genetic drift (Halligan & Keightley, 2009; Katju & Bergthorsson, 2019). Whole genome sequencing of MA lines can therefore provide direct estimates of spontaneous mutation rates and spectra, including in *Arabidopsis thaliana* (Monroe et al., 2022; Ossowski et al., 2010; Weng et al., 2019). However, there are drawbacks: strongly deleterious and lethal mutations remain subject to purifying selection even in MA designs; experiments are labor-intensive and time-consuming; studies are limited by the number of genotypes and environments; scaling across many environments and genetic backgrounds is difficult (Halligan & Keightley, 2009; Katju & Bergthorsson, 2019). As a result, MA studies provide valuable estimates of mutation input, but may not fully capture mutation-rate dynamics across natural populations.

Population genetic inference approaches

A second class of approaches infers evolutionary processes from patterns of standing polymorphism and divergence data in natural populations. These methods are primarily

designed to estimate selection, demographic history, and distributions of fitness effects (DFEs). However, they frequently assume homogeneous mutation rates across the genome, and in most cases, mutation-rate heterogeneity is not itself treated as a parameter of inference. Consequently, spatial differences in mutation rates may be confounded with signatures of purifying selection.

Simulation-based inference approaches

Simulation-based approaches provide a third strategy by allowing multiple evolutionary processes to vary simultaneously and evaluating which combination reproduces observed genomic patterns. In principle, these models can jointly represent mutation-rate heterogeneity, purifying selection, recombination and demography. However, few inference frameworks allow mutation and selection parameters to vary independently across genomic compartments. Recent work in plants suggests that gene-body hypomutation may be mechanistically controlled and may have evolutionary consequences (Belfield et al., 2018; Monroe et al., 2022; Quiroz et al., 2024). However, it remains difficult to experimentally separate and quantify how much of the contrast between genic and intergenic polymorphism levels is driven by mutation-rate heterogeneity vs purifying selection. A simulation-based inference framework is therefore needed to evaluate their contributions while accounting for the similar population-genetic signatures they can produce.

Study Aims and Approach

Our central question is whether reduced polymorphism in gene bodies is best explained by lower mutation input, stronger purifying selection, or a combination of both

processes. To address this question, we used forward evolutionary simulations and approximate Bayesian computation with random forest (ABC-RF) to evaluate the relative support for mutation-rate heterogeneity and purifying selection as explanations for differences in genic and intergenic polymorphism. Genic and intergenic sequence were modeled separately, with mutation rates, fraction of selected sites, and mean deleterious fitness effects allowed to vary independently and simultaneously between genomic compartments.

We relied on *Arabidopsis thaliana* because it has extensive population-genomic resources (1001 Genomes Consortium, 2016), experimentally characterized mutation-rate heterogeneity (Monroe et al., 2022; Weng et al., 2019), and mechanistic evidence of gene-body hypomutation (Belfield et al., 2018; Oztas et al., 2018; Quiroz et al., 2024).

2. Materials and methods

Genomic region selection

Simulated genomic regions were based on annotated regions from the *Arabidopsis thaliana* TAIR10 reference genome (Lamesch, 2012). One region containing 100 annotated genes was selected from each of the five chromosomes, with regions chosen to avoid extreme recombination environments and to represent recombination rates characteristic of the broader genomic background (West et al., 2006). Because gene density varies across the genome, the resulting segments differed in physical length. Gene coordinates were obtained from the TAIR10 GFF3 annotation set, and genomic

intervals within each segment were partitioned into genic and intergenic compartments. The complete interval spanning the selected genes, including intervening intergenic sequence, was retained to preserve the local spatial organization of genomic features and patterns of polymorphism while reducing biases associated with unusually high- or low-recombination regions (Figure S2.1).

Forward evolutionary simulations

Forward evolutionary simulations were performed using SLiM with tree-sequence recording enabled (Haller et al., 2019; Haller & Messer, 2019). Simulations were designed to model patterns of polymorphism across genic and intergenic regions while allowing mutation rates and purifying selection parameters to vary independently between genomic compartments.

All six simulation parameters were varied across replicates: genic mutation rate, intergenic mutation rate, the fraction of deleterious sites in genic regions, the fraction of deleterious sites in intergenic regions, the mean deleterious fitness effect in genic regions, and the mean deleterious fitness effect in intergenic regions (Table 2.1). Mutation-rate priors were informed by empirical estimates of spontaneous mutation rates in *Arabidopsis thaliana* summarized across mutation-accumulation studies (Quiroz et al., 2023). Recombination rates were parameterized using genome-wide recombination estimates from *Arabidopsis thaliana* (Salomé et al., 2012). Deleterious fitness effects were modeled using compartment-specific gamma distributions informed by previous estimates of the distribution of fitness effects in *Arabidopsis thaliana* (Huber et al., 2018).

Parameter	Description	Prior range	Distribution
gμ	Genic mutation rate per site per generation	1×10^{-10} to 1×10^{-7}	Log-uniform
iμ	Intergenic mutation rate per site per generation	1×10^{-10} to 1×10^{-7}	Log-uniform
gd	Proportion of genic mutations modeled as deleterious	0.10 to 0.75	Uniform
id	Proportion of intergenic mutations modeled as deleterious	0.10 to 0.75	Uniform
gdfe	Mean selection coefficient for deleterious genic mutations	-0.100 to -0.001	Negative log-uniform
idfe	Mean selection coefficient for deleterious intergenic mutations	-0.100 to -0.001	Negative log-uniform

Table 2.1. Model parameters and prior distributions used for ABC inference. The table summarizes the six inferred parameters, their biological interpretations, prior ranges, and sampling distributions. Mutation rates are expressed per site per generation, while deleterious DFE means represent negative selection coefficients.

To improve computational tractability, a biologically effective population size of $N_e = 300,000$ was scaled by a factor of $Q = 50$, resulting in a simulated effective population size of $N_{sim} = 6,000$. Mutation rates, recombination rates, and selection coefficients were increased by the same factor to preserve key population-scaled quantities, including N_{ep} , N_{er} , and N_{es} . Although population-size rescaling is widely used to reduce the computational cost of forward evolutionary simulations, it does not preserve all evolutionary dynamics exactly, and its effects can depend on linkage, selection, genomic extent, mating system, and the degree of scaling applied (Dabi & Schrider, n.d.; Marsh et al., 2026). We therefore evaluated sensitivity to rescaling by repeating the chromosome 5 analysis using a reduced scaling factor. The chromosome 5 analysis was repeated using a reduced scaling factor of ($Q=25$) under the same scaling rules and simulation framework as the primary ($Q=50$) analysis.

Approximately 100,000 simulations were performed for each chromosome segment. Simulations were run for 60,000 generations, corresponding to 10N to provide an extended forward-time burn-in for selected variation before sampling. Tree-sequence recording was used throughout simulations to retain genealogical information for downstream recapitation and neutral mutation overlay (Haller et al., 2019).

To evaluate the sensitivity of the model inference to demographic assumptions, three demographic models were compared: a constant population size model, a recent bottleneck, and a recent expansion. All models shared an initial 60,000-generation period at $N_{sim} = 6,000$. The constant model remained at 6,000 whereas the bottleneck and expansion models underwent an instantaneous change in population size for an additional 500 generations. In the bottleneck model, population size changed to $N_{sim} = 1,000$, with subsequent recovery to $N_{sim} = 6,000$; in the expansion model, it increased to $N_{sim} = 18,000$ (Figure 2.5). The constant-size model was used as the primary demographic framework for parameter inference, while the bottleneck and expansion models were used as alternative recent demographic scenarios to test whether changes in population size could influence inferred mutation-rate and selection parameters. For the alternative demographic models, 100,000 simulations were generated under each scenario using the chromosome 1 segment.

Empirical data processing and summary statistic calculation

Population-genetic summary statistics describing genetic diversity and the site-frequency spectrum were calculated from publicly available whole-genome variation data generated by the *Arabidopsis thaliana* 1001 Genomes Project, consisting of 1,135

globally distributed natural accessions (1001 Genomes Consortium, 2016). The full 1,135-accession panel was used as the primary empirical target dataset for approximate Bayesian computation (ABC) inference.

Analyses were performed using TAIR10-aligned all-sites VCF files. Five focal genomic segments were analyzed, corresponding to the same chromosome regions used in simulations. Gene and intergenic coordinates were obtained from the TAIR10 genome annotation and used to classify each site as genic or intergenic.

Site-level missingness, allele counts, and nucleotide diversity were calculated using vcfTools v0.1.16 (Danecek et al., 2011). A site was considered callable if genotype data were available for at least 80% of the expected haplotypes across the 1,135 accessions. Callable-site counts were calculated separately for genic and intergenic regions and used as denominators for per-site summary statistics. Because an all-sites VCF was used, callable-site estimates included both invariant and polymorphic positions that passed the callability threshold.

Analyses were restricted to biallelic SNPs, and folded site-frequency spectra were calculated to avoid dependence on ancestral-state assignment. Invariant sites were retained for callable-site estimation but were excluded from segregating-site counts. For each genic and intergenic region, summary statistics included the number of callable sites, the number of segregating sites, nucleotide diversity (π), Watterson's θ , Tajima's D , singleton and doubleton counts, and folded site-frequency-spectrum proportions.

These statistics were calculated using the same analysis framework applied to simulated datasets and constituted the empirical target vector used for ABC inference. For the pooled five-segment analysis, genic and intergenic summary statistics were combined separately across segments. Site and variant counts were summed, and SNP, singleton, and doubleton densities were calculated using the total number of callable sites as the denominator. Nucleotide diversity was calculated as the callable-site-weighted mean of segment-level estimates. Folded site-frequency-spectrum counts were reconstructed from segment-level proportions and SNP counts before being combined across segments. Watterson's θ and Tajima's D were then recalculated using the pooled counts, callable sequence length, nucleotide diversity, and common sample size. As a sensitivity analysis, segment-level summary statistics were also combined using their simple arithmetic mean, thereby assigning equal weight to each segment.

Approximate Bayesian computation inference

Approximate Bayesian computation (ABC) was used to infer evolutionary parameters from empirical summary statistics by comparing observed genomic patterns to a reference table generated from forward simulations (Beaumont et al., 2002). For each simulation, the reference table included the parameter values used to generate the dataset and the corresponding summary statistics derived from the simulated genomic data.

ABC inference was performed using random forest ABC. Regression forests were fit separately for each inferred parameter using the `abcrf` R package (Raynal et al., 2019).

All six parameters varied simultaneously across the simulation reference tables, but ABC-RF inference was performed separately for each parameter using parameter-specific regression forests. Separate analyses were conducted for each of the five focal genomic segments, for alternative demographic models simulated on chromosome 1 (Figure 2.5), and for a pooled five-segment dataset that combined information across all genomic regions. For each analysis, simulated summary statistics were used as predictor variables and the corresponding generating parameter values were used as response variables. Random forest models were trained using 1,500 trees for each parameter. Posterior distributions for empirical target datasets were estimated from random forest weights, and posterior means, medians, 95% credible intervals, and normalized mean absolute error estimates were recorded. All ABC analyses used identical summary-statistic calculation procedures for simulated datasets to ensure direct comparability between the simulation reference tables and observed target vectors.

In addition to estimating genic and intergenic mutation rates separately, we directly inferred their ratio to quantify evidence for genic hypomutation. The response variable for this analysis was $\log_{10}(\text{genic}/\text{intergenic})$. Separate ratio forests were fitted for each chromosome segment and for the pooled, all-segment analysis. Each forest contained 500 trees. Posterior weights from the fitted forests were used to calculate posterior means, medians, and 95% credible intervals on the log scale, which were then transformed back to the mutation-rate-ratio scale. We also calculated the posterior

probability that $\text{genic/intergenic} < 1$, corresponding to a lower mutation rate in genic than in intergenic regions.

Simulation-based parameter recovery

The performance of the ABC inference framework was evaluated using a simulation-based parameter recovery analysis. For each simulation panel, including all chromosome-specific analyses, alternative demographic models, and pooled five-segment analyses, 100 pseudo-observed datasets were randomly sampled from the corresponding simulation reference table. The selected pseudo-observed simulation was removed from the reference table prior to inference.

Each pseudo-observed dataset was analyzed using the same summary statistics, transformations, and ABC-RF settings applied to empirical analyses. Because the true generating parameter values were known, posterior estimates could be directly compared with the underlying simulation parameters. Performance was assessed using mean absolute error, root mean squared error, posterior interval width, posterior contraction relative to the prior distribution, correlation between true and inferred values, and 95% credible interval coverage. Mutation-rate and deleterious-effect parameters were evaluated on log-transformed scales, whereas selected-site fractions were evaluated on their original linear scales.

Variable-importance scores were additionally summarized to identify which summary statistics contributed most strongly to parameter inference. These analyses provided an

assessment of parameter identifiability and the ability of the ABC-RF framework to recover known evolutionary parameters from simulated genomic data.

3. Results

We used an ABC random-forest (ABC-RF) workflow to determine which combinations of mutation and selection parameters produced patterns of genetic variation most similar to those observed in *Arabidopsis thaliana*. Forward simulations generated a reference dataset linking known parameter values to population-genetic summary statistics, and the ABC-RF models then used the empirical summary statistics to infer posterior distributions for the underlying parameters. We estimated six parameters: the mutation rate, the fraction of mutations modeled as deleterious, and the mean effect of deleterious mutations separately for genic and intergenic sequence (Figure 2.1; Table 2.1). This design allowed mutation and purifying selection to vary simultaneously and independently between genomic compartments. The primary analysis pooled summary statistics across five genomic segments while retaining separate genic and intergenic summaries. The segments were distributed across the five *Arabidopsis* chromosomes and differed in length and genic–intergenic composition (Figure S2.1; Table S2.1).

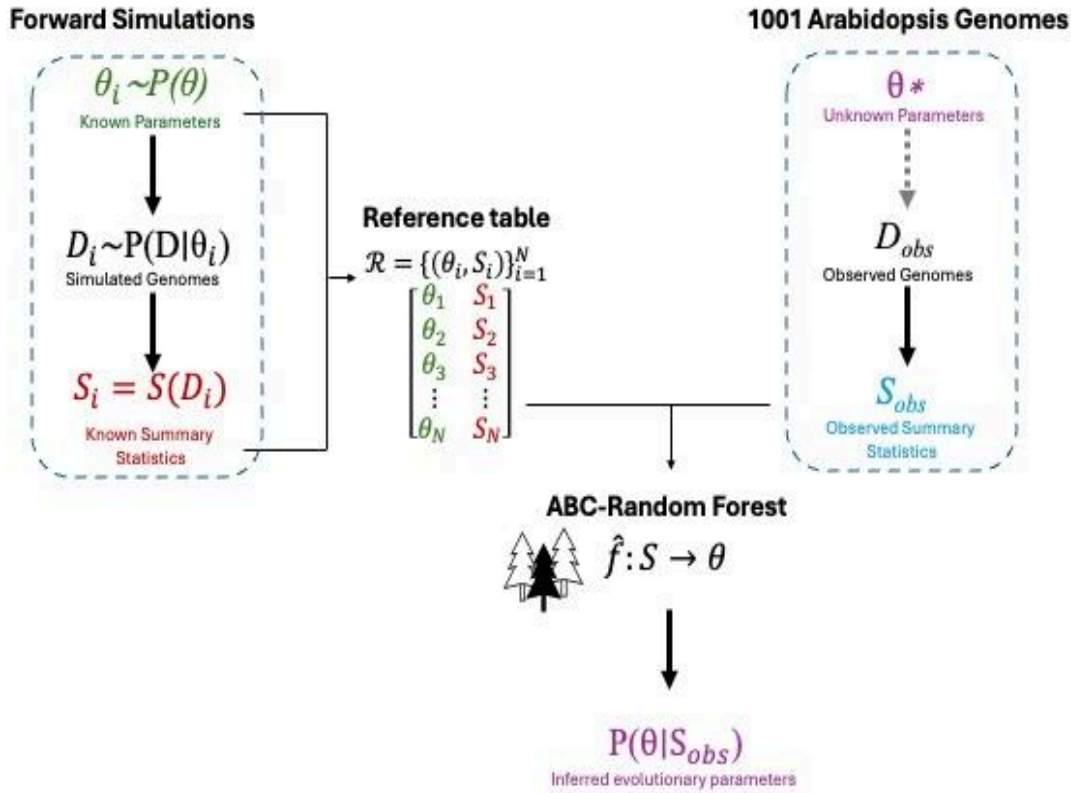


Figure 2.1. Overview of the simulation-based ABC random forest workflow. Parameters sampled from prior distributions were used in forward simulations to construct a reference table linking parameter values with summary statistics. ABC-RF compared this reference table with summary statistics from the *Arabidopsis thaliana* 1001 Genomes dataset to infer posterior parameter distributions.

Empirical inference supports higher intergenic mutation rates

The central question for the modelling was whether the reduced polymorphism observed in genic sequence is better explained by lower mutation input, or by stronger purifying selection. When the pooled five-segment ABC-RF model was applied to the empirical genic and intergenic summaries, the clearest compartment-level contrast was mutation rate, with lower inferred mutation input in genic than in intergenic sequence (Figure 2.2, Table S2.2). The genic posterior mean was 2.52×10^{-8} , approximately 29% lower than the intergenic mean of 3.53×10^{-8} ; the corresponding medians were 2.53×10^{-8} and 3.26×10^{-8} . Although the marginal 95% credible interval for the genic mutation

rate (1.94×10^{-9} to 5.76×10^{-8}) overlapped the interval for the intergenic rate (2.47×10^{-9} to 8.30×10^{-8}), the direction of the point-estimate contrast was repeated across the data: the genic posterior mean was lower than the intergenic mean in each of the five individual segments (Figure S2.2; Table S2.2). Comparison of these marginal intervals does not directly estimate uncertainty in the relative difference between the two mutation rates.

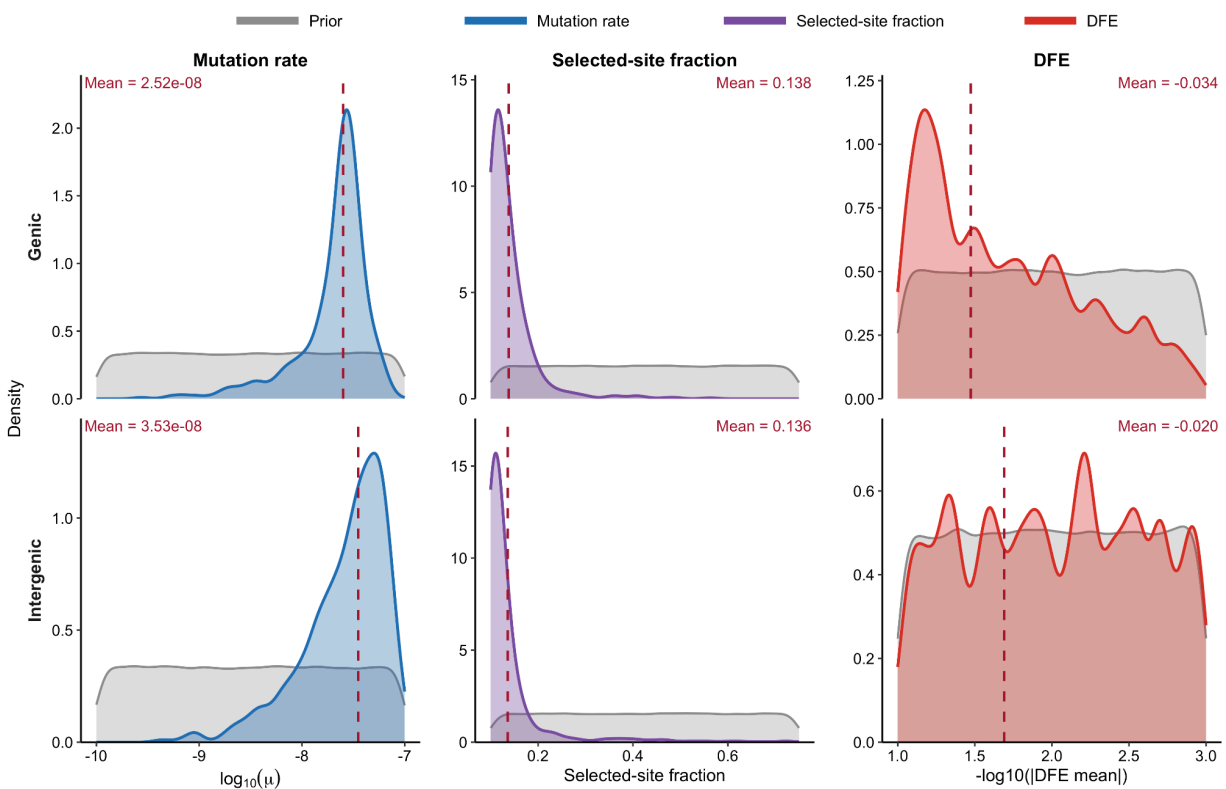


Figure 2.2. Prior and posterior distributions using the pooled five-segment analysis. Prior and ABC-RF posterior distributions are shown for the six parameters inferred from empirical summary statistics pooled across five chromosome segments. Columns show mutation rate, deleterious-mutation fraction, and DFE mean; rows show genic and intergenic parameters. Mutation rates are plotted as $\log_{10} \mu$, deleterious-mutation fractions on their original scale, and DFE means as $-\log_{10}(|\text{DFE mean}|)$. Colored densities show ABC-RF posteriors, gray densities show the corresponding priors, and dashed vertical lines and labels indicate posterior means. Density axes are scaled independently among parameter categories.

We therefore directly inferred the genic-to-intergenic mutation-rate ratio as a derived parameter using a separate ABC-RF analysis (Figure 2.3). In the pooled five-segment

analysis, the posterior median ratio was 0.58, with a 95% credible interval of 0.33-0.90 and a posterior probability of 0.991 that the ratio was below one. Thus, when the genic-intergenic contrast itself was treated as the parameter of inference, the posterior strongly supported lower mutation rates in genic sequence. The direction of this contrast was also reproduced across all five chromosome segments. Segment-specific posterior median ratios ranged from 0.60 to 0.85, and the posterior probability that the ratio was below one ranged from 0.75 to 0.95 (Figure 2.3).

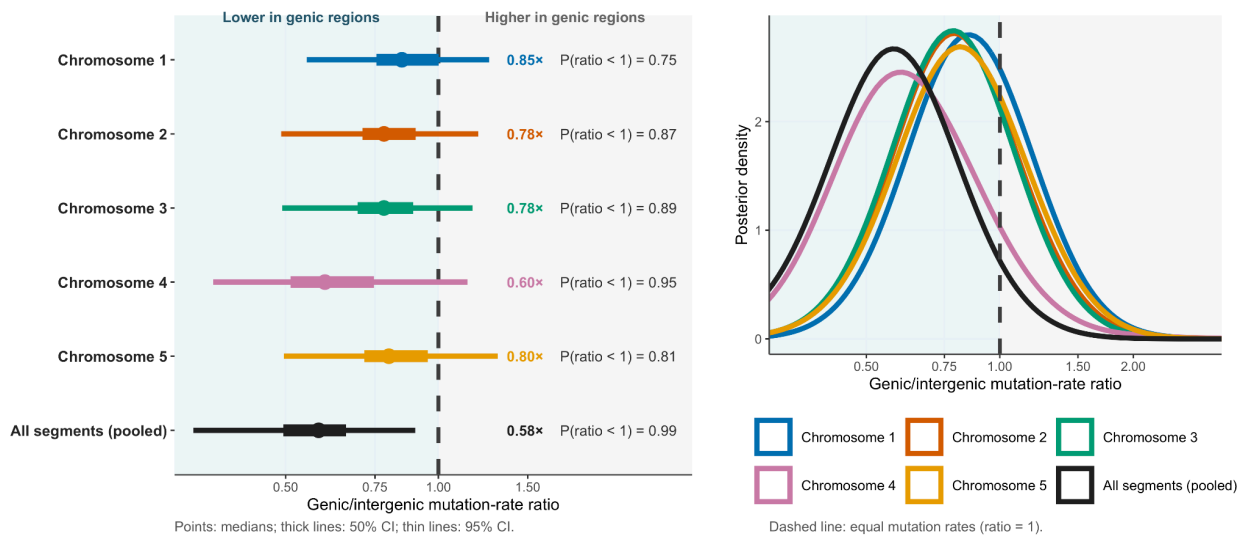


Figure 2.3. Chromosome-specific and pooled posterior estimates of the genic-to-intergenic mutation-rate ratio. (a) points show posterior medians, thick bars show 50% credible intervals, and thin bars show 95% credible intervals for each chromosome; labels report the median ratio and posterior probability that the ratio is below one. (b), weighted posterior density distributions are shown by chromosome. The dashed line denotes equal mutation rates in genic and intergenic regions (ratio = 1); values below one indicate lower inferred mutation rates in genic

The selection parameters did not show an equally consistent compartment-level contrast, indicating that the model provided little evidence for a systematic difference in purifying selection between genic and intergenic sequence. Pooled selected-site-fraction means were similar for genic and intergenic sequence (0.138 and 0.136, respectively), and both posteriors accumulated near the lower prior boundary of 0.10 (Figure 2.2; Table S2.2). Their segment-level ordering was inconsistent: the genic

mean was higher only in segment 4, whereas the intergenic mean was higher in segments 1,2, 3, and 5 (Figure S2.2; Table S2.2). Mean DFE posteriors were broad and overlapped extensively between genic and intergenic sequence. Although the pooled genic point estimate was more negative, the model did not estimate this difference precisely (Figure 2.2; Table S2.2).

The pooled estimates were not driven by a single chromosomal segment. For all six parameters, the pooled posterior means fell within the range of the five chromosomal segment-specific means, and an alternative analysis using the equal-weight mean of chromosomal segment-level summary statistics produced the same principal qualitative results (Figure S2.3; Table S2.2). Together with the direct ratio inference, these results identify lower genic mutation input as the most reproducible compartment-level difference, whereas the selection parameters provided little consistent evidence that either the fraction or mean effect of deleterious mutations differed systematically between genic and intergenic sequence.

The relative mutation-rate contrast is robust to simulation scaling

Because population-size rescaling can alter some outcomes of forward population-genetic simulations, we repeated the chromosome 5 analysis using a reduced scaling factor of $Q = 25$ and compared it with the primary $Q = 50$ analysis (Figure S2.8). Reducing the degree of scaling produced a moderate upward shift in the absolute mutation-rate estimates. The genic posterior median increased from 2.55×10^{-8} under $Q = 50$ to 3.33×10^{-8} under $Q = 25$, while the intergenic median increased

from 3.26×10^{-8} to 4.30×10^{-8} . The corresponding 95% credible intervals overlapped extensively between scaling treatments.

In contrast to the shift in absolute mutation-rate estimates, their relative genic-intergenic relationship was highly stable. The directly inferred mutation rate-ratio median was 0.80 under $Q = 50$ and 0.84 under $Q = 25$, a difference of less than 5%, with substantially overlapping posterior distributions (Figure S2.8). Thus, reducing the scaling factor shifted the absolute mutation-rate scale in both genomic compartments in the same direction while preserving the relative evidence for lower genic mutation rates.

Pseudo-observed tests establish greatest confidence in mutation-rate inference

To calibrate how much confidence to place in each empirical parameter estimate, we treated 100 simulated datasets as pseudo-observed data. Because each dataset was generated using a known set of mutation and selection parameter values, we could test how accurately the ABC-RF framework recovered the true underlying values for all six parameters. These tests revealed a clear hierarchy of inferential performance: genic and intergenic mutation rates were recovered nearly one-to-one, selected-site fractions were recovered with greater uncertainty and some shrinkage toward intermediate values, and mean DFE parameters were the least precisely resolved (Figure 2.4; Figure S2.4; Table S2.3).

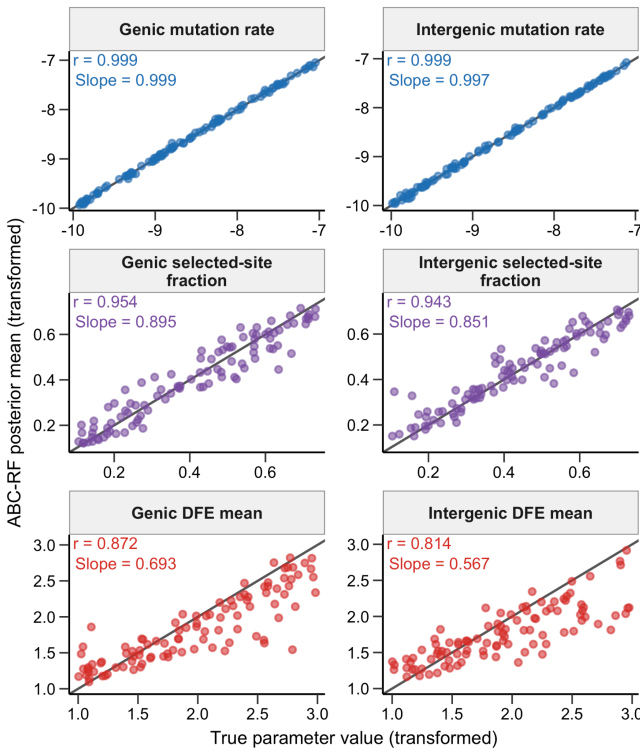


Figure 2.4. Recovery of model parameters using the pooled five-segment dataset. True simulated parameter values are plotted against ABC random-forest posterior means for 100 pseudo-observed datasets. The diagonal line denotes perfect recovery. Pearson's correlation coefficient (r) and the slope from regressing the posterior mean on the true value are reported for each parameter. Mutation rates are shown on the $\log_{10}$ scale, selected-site fractions on their original scale, and DFE means as $-\log_{10}(|\text{DFE mean}|)$.

The strongest performance was for genic and intergenic mutation rates. Posterior means closely tracked their generating values, with correlations of $r = 0.999$ and $r = 0.999$ and regression slopes of 0.999 and 0.997, respectively (Figure 2.4). Nominal 95% posterior-interval coverage was 1.00 for both the genic and intergenic mutation rates (Figure S2.4; Table S2.3). This near one-to-one recovery demonstrates that the model and summary statistics can identify mutation-rate variation even when the selection parameters vary concurrently.

Selected-site fractions showed intermediate recovery, with greater dispersion and shrinkage toward the center of the prior. Recovery was stronger for the genic fraction ($r = 0.954$; slope = 0.895) than for the intergenic fraction ($r = 0.943$; slope = 0.851), indicating that values near the extremes of the prior were systematically estimated

closer to the center (Figure 2.4). Mean DFE parameters showed the weakest recovery (genic: $r = 0.872$, slope = 0.693; intergenic: $r = 0.814$, slope = 0.567), consistent with their broad empirical posteriors. Their high 95% interval coverage of 0.96 should therefore be interpreted alongside the limited precision of those intervals (Figure 2.4; Figure S2.4; Table S2.3).

The recovery hierarchy was reproducible across individual chromosome segments and in both the pooled and equal-weight analyses. Mutation rates were the best-recovered parameters in each of the five individual chromosomal segment tests and in the equal-weight analysis, while selected-site fractions were intermediate and DFE means were least precise (Figure S2.5; Figure S2.6; Tables S2.3 and S2.4). Pooling most clearly improved the more difficult genic selected fraction and DFE parameters. These validation results establish that the repeated empirical mutation-rate contrast between genic and intergenic regions is the parameter class that the modeling framework most reliably estimates.

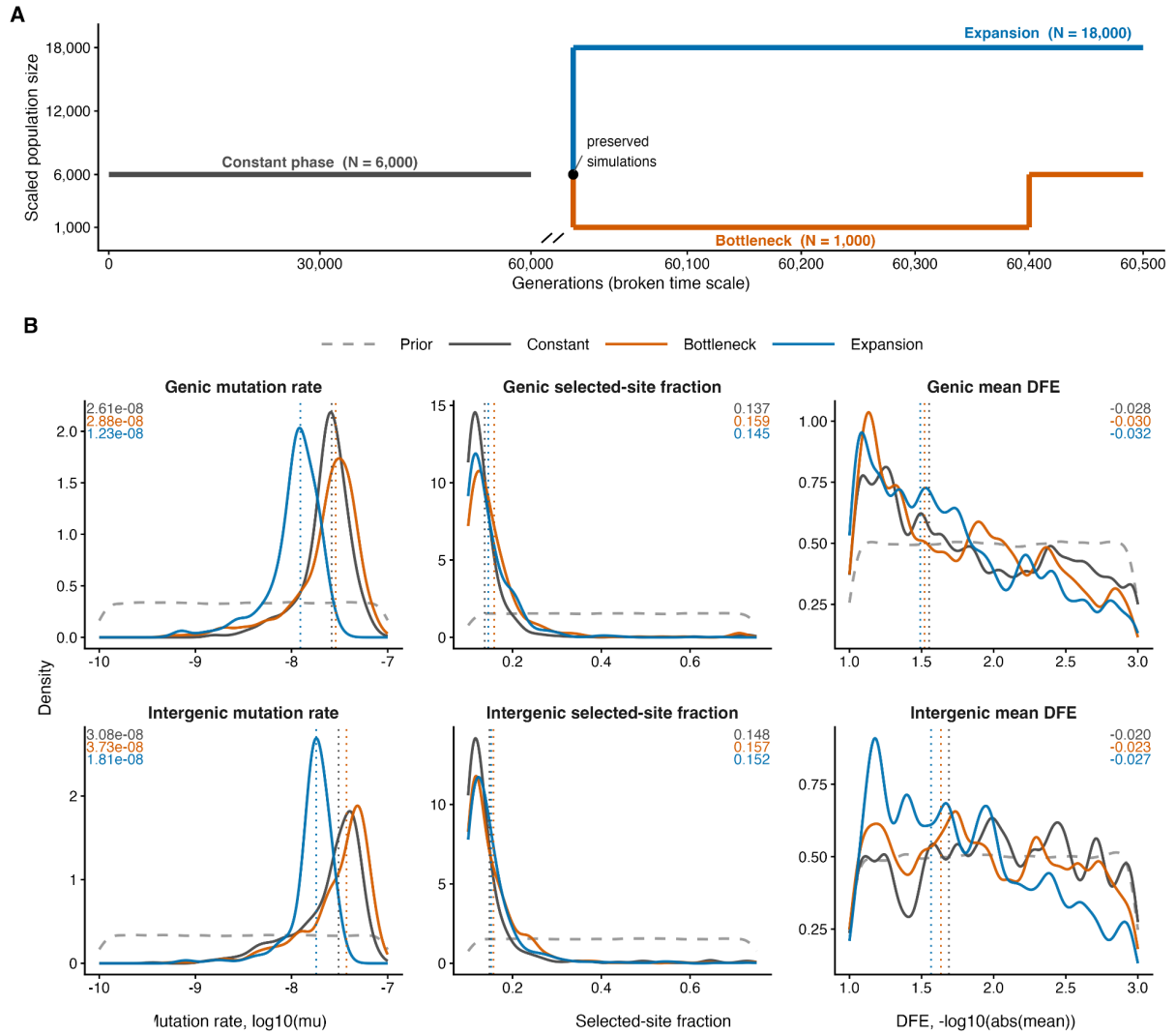


Figure 2.5. Demographic simulation design and sensitivity of parameter inference to demographic history. (a) Constant, bottleneck, and expansion histories used in the demographic sensitivity analysis. Simulations preserved after 60,000 generations at $N = 6,000$ provided the common starting point for the appended bottleneck and expansion treatments. Population sizes are simulation-scaled values, and the broken axis separates the constant phase from the subsequent 500-generation treatments. (b) Weighted ABC random-forest prior and posterior distributions for six parameters inferred from chromosome 1 under each demographic model. Rows distinguish genic and intergenic parameters, and columns show mutation rate, selected-site fraction, and mean DFE. Mutation rates are plotted as $\log_{10}(\mu)$, selected-site fractions on their original scale, and DFE means as $-\log_{10}(|\text{DFE mean}|)$. Dotted lines and colored labels report posterior means; density axes are scaled independently among panels.

Variable-importance patterns were consistent with the biological information expected to inform each parameter class (Figure S2.7). Genic and intergenic mutation rates were driven primarily by compartment-matched measures of variant abundance, including

singleton counts, SNP counts and densities, Watterson's θ , and doubleton counts. Deleterious-mutation fractions were informed more strongly by Tajima's D, nucleotide diversity, and the shape of the site-frequency spectrum, whereas DFE means depended primarily on low-frequency site-frequency-spectrum bins and singleton structure.

Genic hypomutation is robust to demographic assumptions

After establishing the primary genic-intergenic result, we tested whether it depended on the assumed demographic history by analyzing the chromosome 1 segment under different demographic models; constant-population, recent bottleneck, and recent expansion models (Figure 2.5A). Demography changed the centers and widths of some posterior distributions, but it did not change the mutation-rate difference: the genic estimate was lower than the intergenic estimate under all three models (Figure 2.5B; Table S2.5)

The constant and bottleneck models produced similar mutation-rate means. Genic estimates were 2.61×10^{-8} mutations per base pair per generation under the constant model and 2.88×10^{-8} mutations per base pair per generation under the bottleneck, while intergenic estimates were 3.08×10^{-8} and 3.73×10^{-8} mutations per base pair per generation, respectively. Expansion shifted both estimates downward, to 1.23×10^{-8} for genic and 1.81×10^{-8} for intergenic sequence. The 95% credible intervals for a given mutation-rate parameter overlapped across the three demographic models, indicating quantitative shifts in posterior mass rather than discrete model-specific estimates (Figure 2.5B; Table S2.5).

Selected-site-fraction estimates were similar across demographic models. The genic posterior means were 0.137, 0.159, and 0.145 under the constant, bottleneck, and expansion models, respectively; the corresponding intergenic means were 0.148, 0.157, and 0.152. Their ordering was not consistent across demographic models: the intergenic estimate was higher under the constant and expansion models, whereas the genic estimate was slightly higher under the bottleneck model. Thus, these analyses do not support a demography-independent genic–intergenic difference in selected-site fraction (Figure 2.5B; Table S2.5). DFE posteriors remained broad under all three demographic histories, providing little resolution of demographic or compartment-level differences. Overall, demography affected the quantitative parameter estimates but preserved the central result of a lower genic than intergenic mutation-rate point estimate.

Summary

Across pooled, chromosome segment-level, and demographic analyses, the genic mutation-rate point estimate was consistently lower than the intergenic estimate. Direct inference of the genic-to-intergenic mutation-rate ratio provided strong support for this contrast in the pooled analysis, with a posterior median of 0.58, a 95% credible interval of 0.33–0.90, and a posterior probability of 0.991 that the ratio was below one. The direction of the contrast was reproduced across all five chromosome segments and persisted under alternative demographic histories and simulation-scaling factors. Pseudo-observed validation further showed that mutation rates were the most accurately recovered parameter class. Selected-site fractions showed no stable genic–intergenic difference, and mean DFE effects were weakly resolved. Together,

these results identify lower genic mutation input as the strongest and most reproducible compartment-level signal in the empirical analysis.

4. Discussion

This study asked whether reduced polymorphism in gene bodies is better explained by lower mutation rates, stronger purifying selection, or a combination of both processes. The simulation framework allowed mutation rate, the fraction of mutations subject to purifying selection, and the mean effect of deleterious mutations to be simultaneously and independently varied between genic and intergenic sequences. Broad prior sampling included mutation-rate differences, selection differences, and combinations of both, allowing us to test which parameter combinations reproduced the empirical polymorphism patterns observed in the *A. thaliana* 1001 genomes data set. Across this parameter space, the most consistent compartment-level signal was a lower mutation-rate point estimate in genic sequence than in intergenic sequence. The pooled genic posterior mean was approximately 29% lower than the intergenic mean, and direct inference of the genic-to-intergenic mutation-rate ratio provided stronger evidence for this relative contrast. The same relationship occurred across all five chromosome segments and under alternative demographic models and persisted under the alternative simulation-scaling factor. Mutation rates were also the most accurately recovered of the six inferred parameters, providing greater confidence in the repeated genic-intergenic mutation-rate contrast.

The selection parameters did not show an equally consistent difference between genic and intergenic regions. Pooled selected-site fractions were nearly identical between

compartments, and their ordering varied among chromosome segments and demographic models. Mean DFE estimates pointed toward stronger deleterious effects in genic sequence, but their posteriors were broad and their pseudo-observed recovery was substantially weaker. Together, these results identify mutation-rate heterogeneity as the most strongly supported explanation for reduced genic polymorphism within the model tested here. They do not imply that purifying selection is absent; rather, the current summaries resolve a genic-intergenic mutation-rate contrast more reliably than they resolve compartment-specific differences in selection. The lack of a clear selection contrast should also not be interpreted as evidence that genic sequence lacks stronger functional constraint, because the intergenic category includes regulatory and other constrained noncoding elements that may also be subject to purifying selection (Haudry et al., 2013; Williamson et al., 2014).

This conclusion is comparative rather than mechanistic or absolute. The analysis does not establish what proportion of the diversity contrast is caused by mutation rate rather than selection, nor does it demonstrate that lower genic mutation rates act alone. However, direct ABC-RF inference of the genic-to-intergenic mutation-rate ratio allows the relative mutation-rate hypothesis to be evaluated without relying on overlap between the marginal credible intervals of the two mutation-rate parameters. The pooled ratio posterior strongly favored values below one, providing direct evidence for lower genic relative to intergenic mutation input within the model. The six underlying parameter-specific ABC-RF outputs are still marginal rather than joint posteriors, so they cannot directly characterize covariance among mutation and selection parameters

or quantify all combinations of mutation and selection that remain compatible with the data. The evidence instead answers the motivating question in a calibrated form: stronger purifying selection alone did not emerge as a comparably consistent explanation, whereas lower genic mutation rates were directly supported and recovered repeatedly across analyses and belonged to the parameter class that the framework estimates most reliably.

Inferred mutation-rate contrast is consistent with gene-body hypomutation

Many population-genetic analyses model variation in selective constraint while assuming that mutation input is uniform across the genome. Evidence that mutation rates vary with functional context challenges this simplification, because the same regions may both experience stronger selection and receive fewer mutations. The repeated inference of a lower genic mutation rate observed here is consistent with the experimental evidence for gene-body hypomutation in *Arabidopsis thaliana*. Direct work in *A. thaliana* has likewise reported reduced mutation occurrence in functionally constrained regions (Monroe et al., 2022). The present modeling framework reaches the same qualitative prediction from standing population-genomic variation while allowing mutation-rate and purifying selection parameters to vary simultaneously. Its main contribution is therefore not the simple observation of lower diversity in genes, but the recovery of a mutation-rate contrast when mutation and selection were allowed to contribute simultaneously. Consistent with this interpretation, a recent causal analysis similarly found that mutation-rate variation contributes alongside background selection and covaries with constrained-site density, suggesting that mutation input and functional

constraint may be biologically linked (Barroso et al., 2026). This parallel supports the broader relevance of the *Arabidopsis* pattern.

The absolute mutation-rate estimates were also biologically plausible and fell within the range reported for *Arabidopsis thaliana* (Beichman et al., 2024), although they should not be interpreted as definitive estimates of the contemporary spontaneous mutation rate. Inference from standing variation integrates mutations accumulated across many generations, genetic backgrounds, and environmental conditions, whereas mutation-accumulation experiments generally measure mutation under more controlled conditions. Moreover, the absolute estimates are conditional on the assumed effective population size, $N_e = 300,000$, because population-genetic data primarily inform population-scaled mutation rates. A different estimate of N_e would therefore shift the inferred per-generation rates, while being less likely to alter the relative genic–intergenic contrast because both compartments were analyzed under the same demographic assumptions.

The central mutation-rate contrast was unchanged across alternative demographic models and individual chromosome segments, indicating that it was not driven by a single demographic assumption or genomic region. Alternative demographic models shifted the absolute mutation-rate estimates, as expected because population growth changes genealogical branch lengths and the site-frequency spectrum (Aris-Brosou & Excoffier, 1996; Coventry et al., 2010; Keinan & Clark, 2012), and can therefore affect mutation-rate inference. However, these shifts did not change the central finding of a

lower genic than intergenic mutation rate. Direct inference of the mutation-rate ratio further showed that the direction of this contrast was reproduced across individual chromosome segments, while the pooled analysis concentrated substantially greater posterior support below a ratio of one.

The same qualitative contrast was also recovered under both population-scaling factors tested, indicating that the result was not specific to the primary scaling factor (Figure S2.8). Reducing the scaling factor from $Q = 50$ to $Q = 25$ shifted both genic and intergenic absolute mutation-rate estimates upward, but the directly inferred genic-to-intergenic ratio changed only from 0.80 to 0.84. This distinction suggests that the relative mutation-rate contrast is substantially more robust to simulation scaling than the absolute per-generation mutation-rate estimates.

Parameter relationships may contribute to recovery hierarchy

One possible explanation for the weaker recovery of the selection parameters is that selected-site fraction and mean DFE can have partially compensatory effects on observed summary statistics. Increasing the selected fraction raises the input of deleterious mutations, whereas a more negative mean DFE removes those mutations more efficiently. Different combinations of these parameters may therefore produce similar polymorphism patterns, making their individual effects more difficult to estimate precisely. Because the retained ABC-RF outputs were marginal rather than joint posteriors, they could not directly characterize covariance or compensatory relationships among parameters. Future work could address this partial confounding by incorporating additional summary statistics that respond differently to the fraction and

strength of deleterious mutations, or by using multivariate posterior approaches or dependence-aware simulation weights that retain covariance among selection parameters.

Conclusions

A flexible simulation framework separated mutation-rate signal from competing selection variation sufficiently to identify a reproducible genomic-compartment contrast. Lower genic mutation input was the most strongly supported explanation for reduced genic polymorphism within the tested model: direct inference of the genic-to-intergenic mutation-rate ratio strongly supported this contrast in the pooled analysis, and the same directional relationship recurred across chromosome segments, pooling approaches, demographic histories, and simulation-scaling factors. Mutation rates also belonged to the parameter class with the strongest pseudo-observed recovery. Purifying selection remains biologically important, but its compartment-specific fraction and mean effect were less consistently identified from the current summaries. The analysis therefore supports gene-body hypomutation, but without assigning a molecular mechanism, a causal percentage, or strict necessity. More broadly, the framework identifies both a repeatable biological signal and the present boundary between estimable and weakly estimable evolutionary processes.

5. Declarations

Acknowledgments

The research was conducted at the University of California Davis, which is located on land that was the home of the Patwin people for thousands of years (<https://diversity.ucdavis.edu/land-acknowledgement-statement>).

The authors acknowledge the High Performance Computing Core Facility at the University of California, Davis, for providing computational resources that have contributed to the research results reported in this paper.

Code availability

6. Supplemental Materials

Supplemental tables

Table S2.1. Genomic segment characteristics

Table S2.2. Empirical posterior estimates and segment sensitivity

Table S2.3. Pooled and equal-weight model-performance metrics

Table S2.4. Individual segment model-performance metrics

Table S2.5. Demographic posterior estimates and recovery

Supplemental Figures

Figure S2.1. Genomic composition of the simulated segments

Figure S2.2. Empirical posterior distributions across individual genomic segments

Figure S2.3. Empirical posterior distributions using equal-weight segment analysis

Figure S2.4. Posterior recovery for the pooled five-segment analysis

Figure S2.5. Parameter recovery across individual genomic segments

Figure S2.6. Posterior recovery using equal-weight averaging across five genomic segments

Figure S2.7. Summary-statistic importance in the pooled ABC-RF analysis

Figure S2.8. Robustness of chromosome 5 mutation-rate inference to simulation scaling

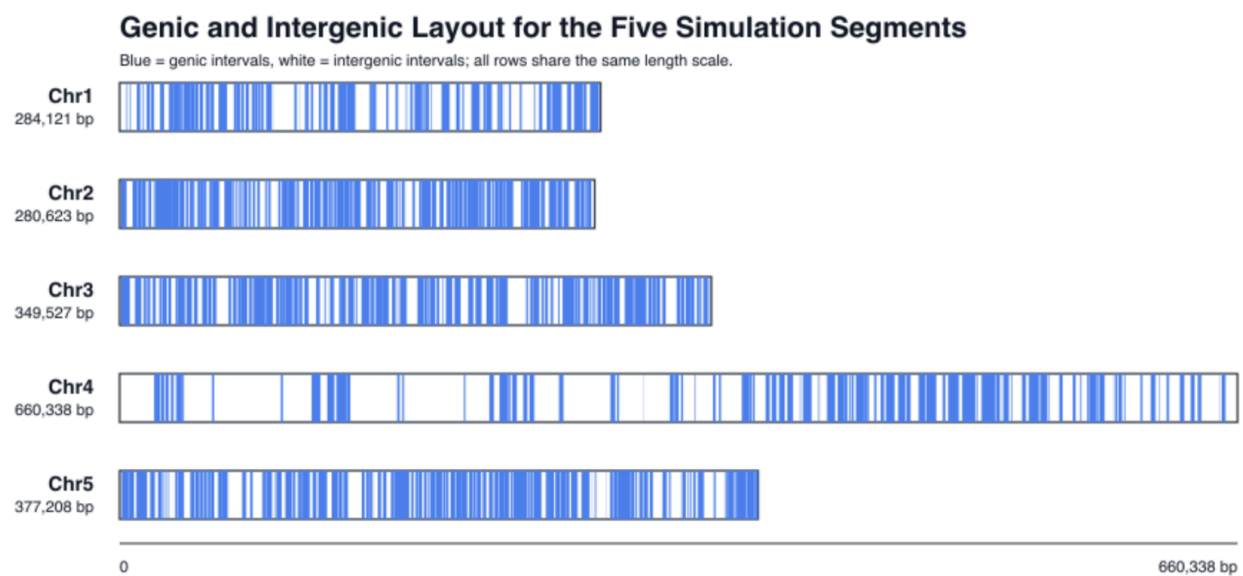


Figure S2.1. Genomic composition of the simulated segments. Blue and white intervals represent genic and intergenic sequence, respectively, across the five segments used for inference. All segments are shown on the same length scale.

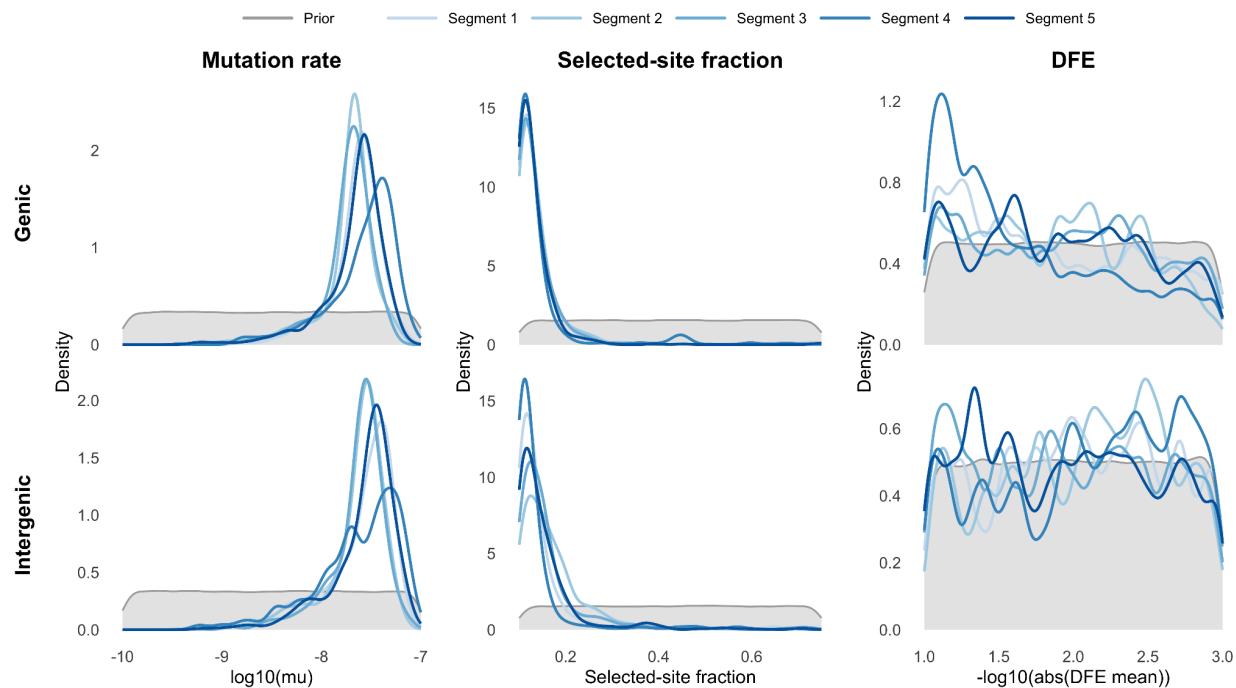


Figure S2.2. Empirical posterior distributions across individual genomic segments. Prior and ABC-RF posterior distributions for the six inferred parameters from five independently analyzed chromosome segments. Blue shades identify segments 1–5, and gray indicates the shared prior. Mutation rates are shown as $\log_{10}(\mu)$, selected-site fractions on their original scale, and DFE means as $-\log_{10}(\text{abs(DFE mean)})$.

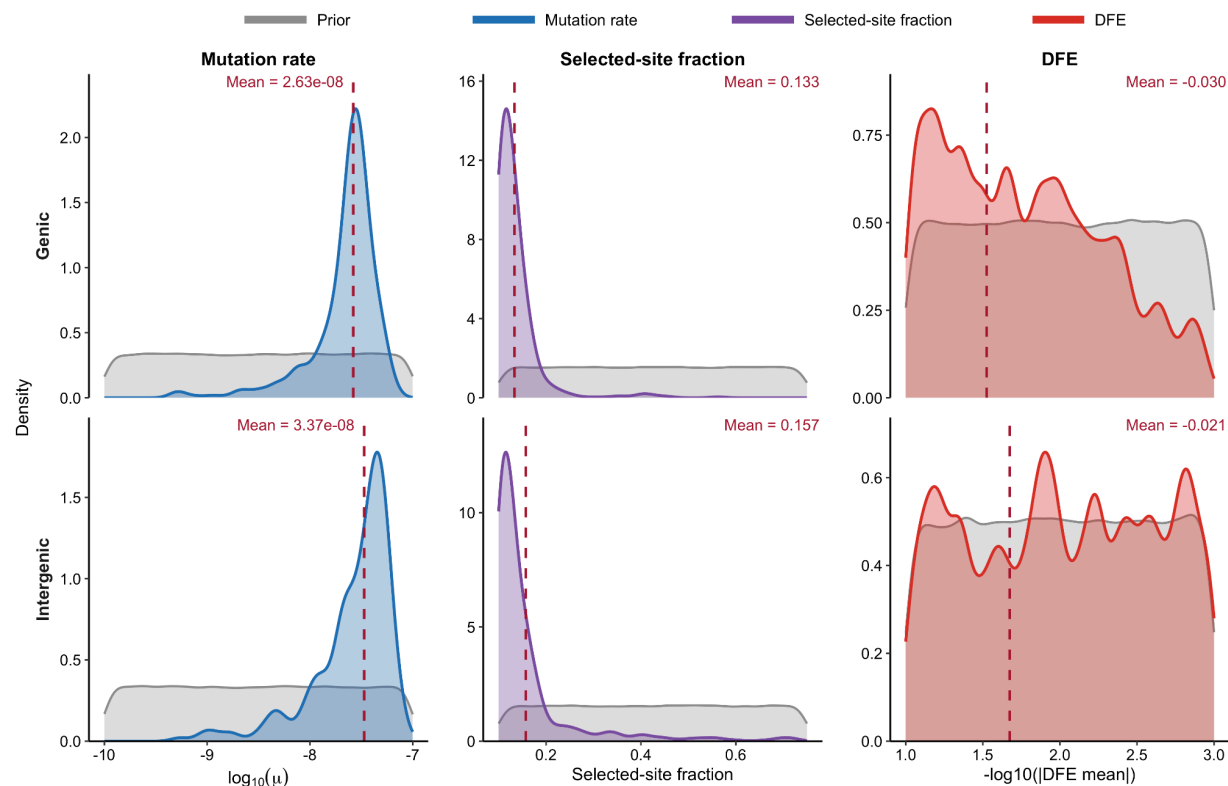


Figure S2.3. Empirical posterior distributions using equal-weight segment analysis. Prior and ABC-RF posterior distributions are shown for the six inferred parameters. Gray densities indicate the priors, colored densities indicate the posteriors, and dashed lines mark posterior means.

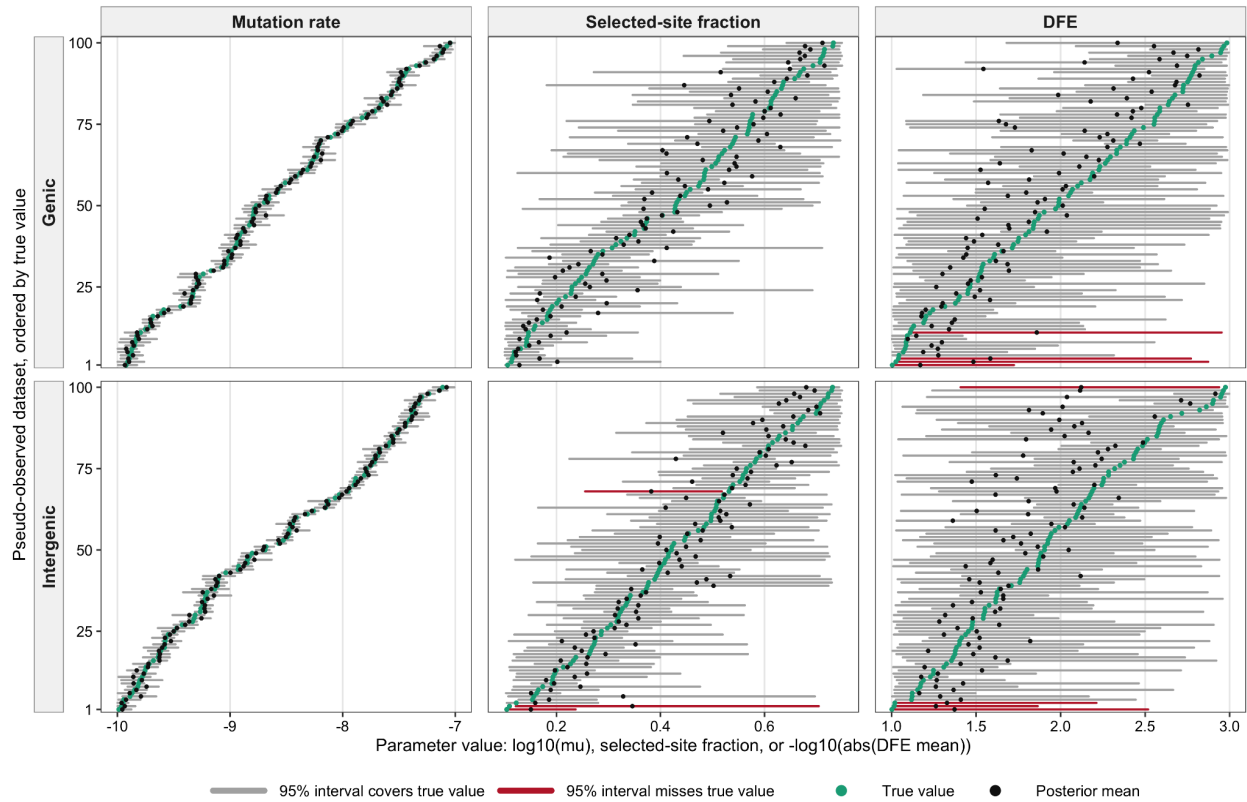


Figure S2.4. Posterior recovery for the pooled five-segment analysis. Gray and red lines show 95% posterior intervals that cover or miss the true generating value, respectively, across 100 pseudo-observed datasets. Green points indicate true values and black points indicate ABC-RF posterior means. Datasets are ordered by true value within each panel. Mutation rates are shown as $\log_{10}(\mu)$, selected-site fractions on their original scale, and DFE means as $-\log_{10}(\text{abs}(\text{DFE mean}))$.

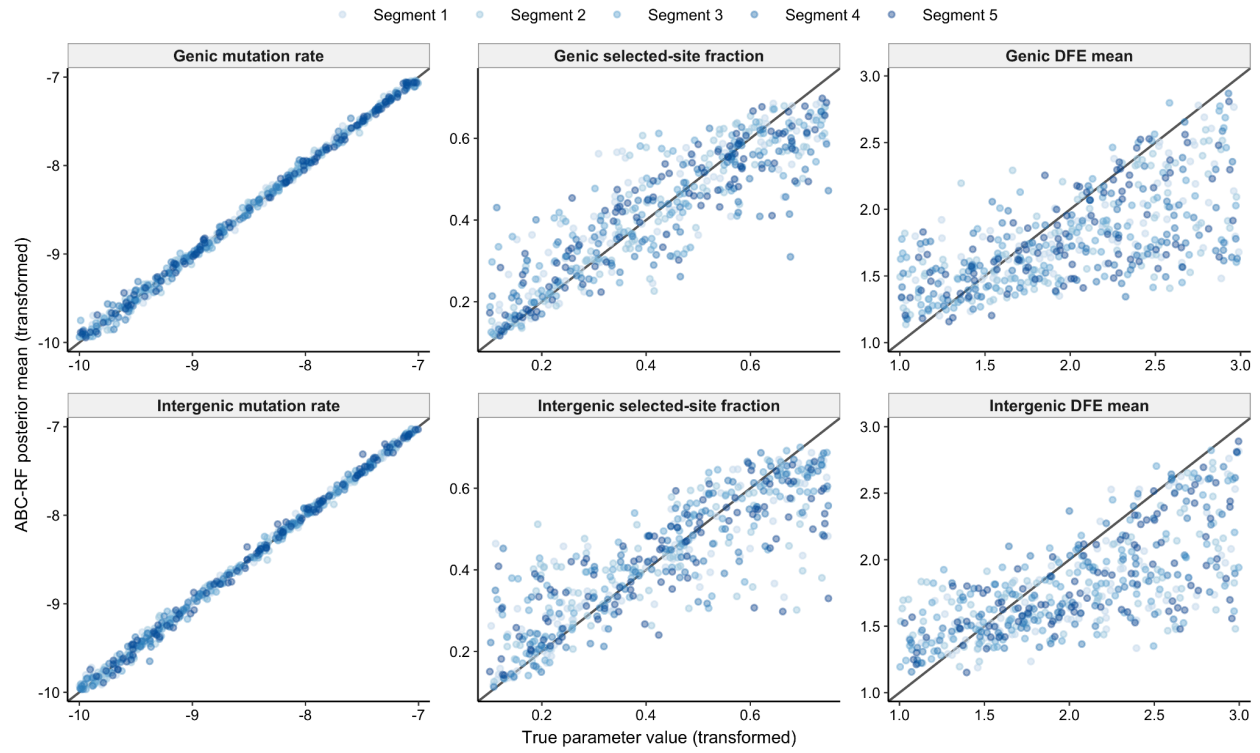


Figure S2.5. Parameter recovery across individual genomic segments. True generating values are plotted against ABC-RF posterior means for 100 pseudo-observed datasets from each of the five independently analyzed segments. The diagonal line denotes perfect recovery. Mutation rates are shown as $\log_{10}(\mu)$, selected-site fractions on their original scale, and DFE means as $-\log_{10}(\text{abs}(\text{DFE mean}))$.

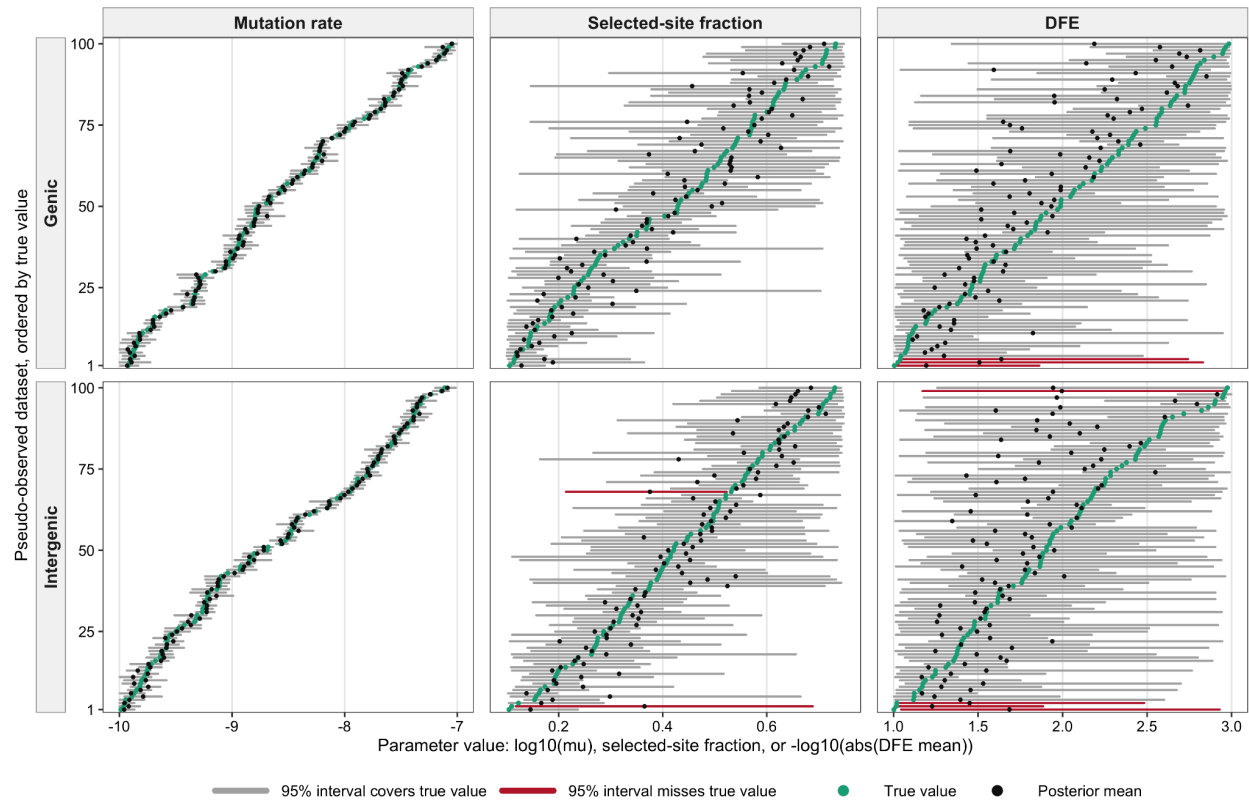


Figure S2.6. Posterior recovery using equal-weight averaging across five genomic segments. Gray and red lines show 95% posterior intervals that cover or miss the true generating value, respectively, across 100 pseudo-observed datasets. Green points indicate true values and black points indicate ABC-RF posterior means. Datasets are ordered independently by true value within each panel. Mutation rates are shown as $\log_{10}(\mu)$, selected-site fractions on their original scale, and DFE means as $-\log_{10}(\text{abs(DFE mean)})$. Segment-level summary statistics were combined using their simple arithmetic mean, giving each segment equal weight.

Summary-statistic importance in the pooled ABC-RF analysis

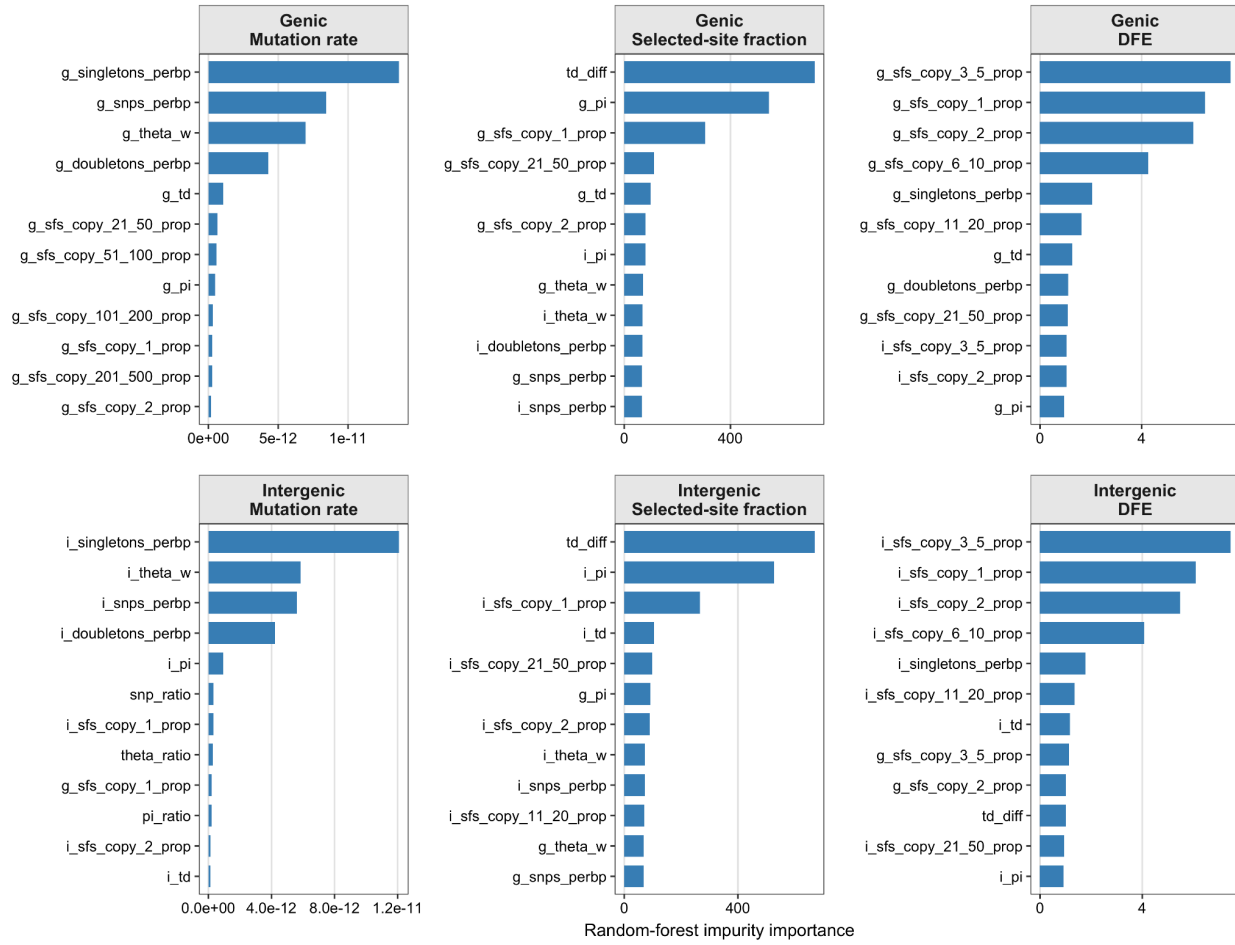


Figure S2.7. Summary-statistic importance in the pooled ABC-RF analysis. Random-forest variable importance for the 12 most informative summary statistics for each inferred parameter in the pooled five-segment analysis. Each panel represents a separately fitted parameter-specific forest and uses an independently scaled importance axis.

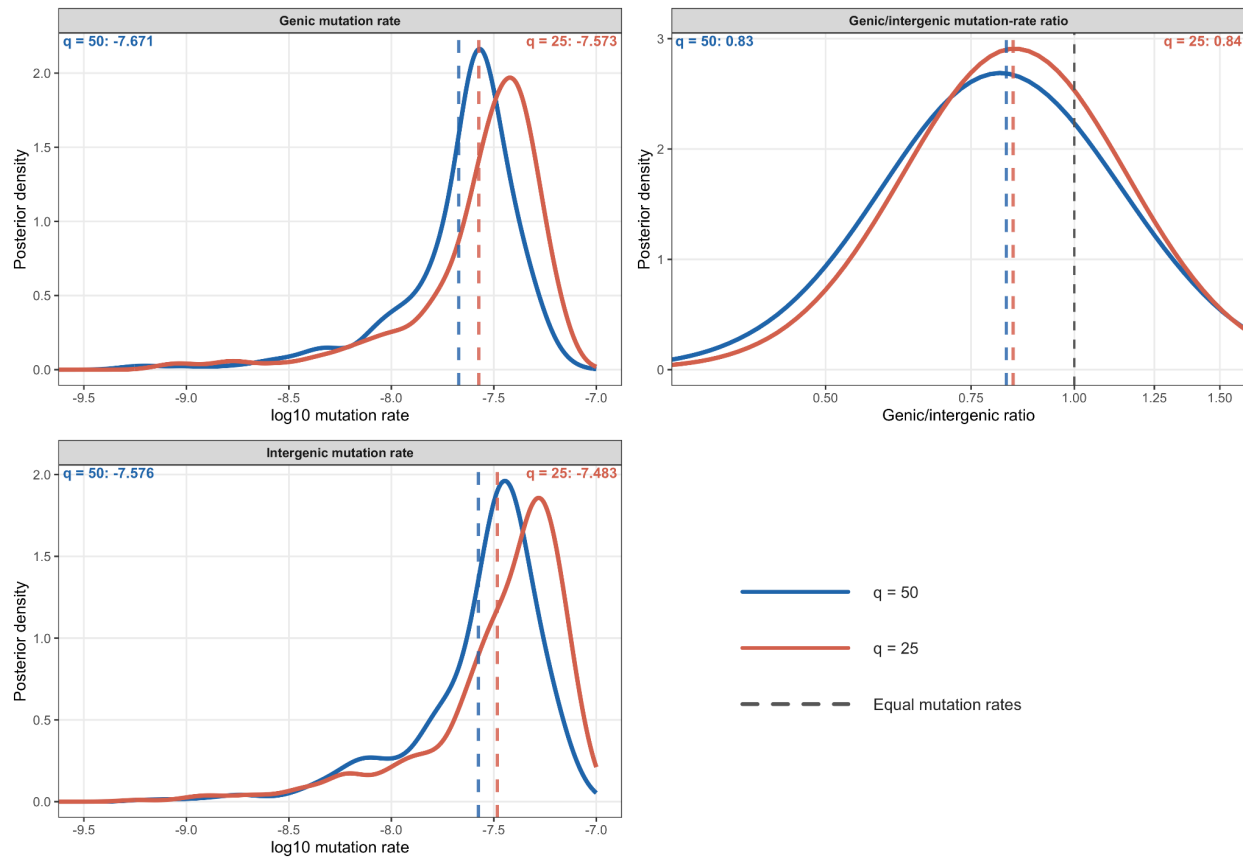


Figure S2.8. Robustness of chromosome 5 mutation-rate inference to simulation scaling. Posterior distributions under $q = 50$ (blue) and $q = 25$ (orange) are shown for genic mutation rate, intergenic mutation rate, and the directly inferred genic-to-intergenic mutation-rate ratio. Colored dashed lines indicate posterior means, and the gray dashed line denotes equal genic and intergenic mutation rates (ratio = 1).

7. References

- 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, & 1001 Genomes Consortium. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2), 481–491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Aris-Brosou, S., & Excoffier, L. (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution*, 13(3), 494–504. <https://doi.org/10.1093/oxfordjournals.molbev.a025610>
- Barroso, G. V., Collier, N. W., Di, C., Lohmueller, K. E., & Ragsdale, A. P. (2026). Causal inference clarifies the roles of background selection and mutation rate variation in shaping

- human genetic diversity. In *bioRxiv* (p. 2026.06.02.727906). bioRxiv.
<https://doi.org/10.64898/2026.06.02.727906>
- Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4), 2025–2035.
<https://doi.org/10.1093/genetics/162.4.2025>
- Beichman, A. C., Zhu, L., & Harris, K. (2024). The evolutionary interplay of somatic and germline mutation rates. *Annual Review of Biomedical Data Science*, 7(1), 83–105.
<https://doi.org/10.1146/annurev-biodatasci-102523-104225>
- Belfield, E. J., Brown, C., Ding, Z. J., Chapman, L., Luo, M., Hinde, E., van Es, S. W., Johnson, S., Ning, Y., Zheng, S. J., Mithani, A., & Harberd, N. P. (2021). Thermal stress accelerates *Arabidopsis thaliana* mutation rate. *Genome Research*, 31(1), 40–50.
<https://doi.org/10.1101/gr.259853.119>
- Belfield, E. J., Ding, Z. J., Jamieson, F. J. C., Visscher, A. M., Zheng, S. J., Mithani, A., & Harberd, N. P. (2018). DNA mismatch repair preferentially protects genes from mutation. *Genome Research*, 28(1), 66–74. <https://doi.org/10.1101/gr.219303.116>
- Boyko, A., Zemp, F., Filkowski, J., & Kovalchuk, I. (2006). Double-strand break repair in plants is developmentally regulated. *Plant Physiology*, 141(2), 488–497.
<https://doi.org/10.1104/pp.105.074658>
- Britt, A. B. (2004). Repair of DNA damage induced by solar UV. *Photosynthesis Research*, 81(2), 105–112. <https://doi.org/10.1023/b:pres.0000035035.12340.58>
- Charlesworth, B., & Jensen, J. D. (2022). Population genetic considerations regarding evidence for biased mutation rates in *Arabidopsis thaliana*. *Molecular Biology and Evolution*.
<https://doi.org/10.1093/molbev/msac275>
- Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4), 1289–1303.
<https://doi.org/10.1093/genetics/134.4.1289>

- Chen, C., Qi, H., Shen, Y., Pickrell, J., & Przeworski, M. (2017). Contrasting determinants of mutation rates in germline and Soma. *Genetics*, 207(1), 255–267.
<https://doi.org/10.1534/genetics.117.1114>
- Chen, X., Chen, Z., Chen, H., Su, Z., Yang, J., Lin, F., Shi, S., & He, X. (2012). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science (New York, N.Y.)*, 335(6073), 1235–1238. <https://doi.org/10.1126/science.1217580>
- 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>
- Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R., Wheeler, D. A., Sabo, A., Lusk, C., Weiss, K. G., Akbar, H., Cree, A., Hawes, A. C., Newsham, I., Varghese, R. T., ... Sing, C. F. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1(1), 131.
<https://doi.org/10.1038/ncomms1130>
- Dabi, A., & Schrider, D. (n.d.). *Population size rescaling significantly biases outcomes of forward-in-time population genetic simulations*. <https://doi.org/10.1101/2024.04.07.588318>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews. Genetics*, 17(7), 422–433. <https://doi.org/10.1038/nrg.2016.58>
- Fryxell, K. J., & Moon, W.-J. (2005). CpG mutation rates in the human genome are highly dependent on local GC content. *Molecular Biology and Evolution*, 22(3), 650–658.
<https://doi.org/10.1093/molbev/msi043>

- Giustozzi, M., Freytes, S. N., Jaskolowski, A., Lichy, M., Mateos, J., Falcone Ferreyra, M. L., Rosano, G. L., Cerdán, P., & Casati, P. (2022). Arabidopsis mediator subunit 17 connects transcription with DNA repair after UV-B exposure. *The Plant Journal: For Cell and Molecular Biology*, 110(4), 1047–1067. <https://doi.org/10.1111/tpj.15722>
- 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., & Keightley, P. D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, 40(1), 151–172. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173437>
- Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., Forczek, E., Joly-Lopez, Z., Steffen, J. G., Hazzouri, K. M., Dewar, K., Stinchcombe, J. R., Schoen, D. J., Wang, X., Schmutz, J., Town, C. D., Edger, P. P., Pires, J. C., Schumaker, K. S., ... Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8), 891–898. <https://doi.org/10.1038/ng.2684>
- Hitomi, K., Arvai, A. S., Yamamoto, J., Hitomi, C., Teranishi, M., Hirouchi, T., Yamamoto, K., Iwai, S., Tainer, J. A., Hidema, J., & Getzoff, E. D. (2012). Eukaryotic class II cyclobutane pyrimidine dimer photolyase structure reveals basis for improved ultraviolet tolerance in plants. *The Journal of Biological Chemistry*, 287(15), 12060–12069. <https://doi.org/10.1074/jbc.M111.244020>
- Huang, Y., Gu, L., & Li, G.-M. (2018). H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *The Journal of Biological Chemistry*,

- 293(20), 7811–7823. <https://doi.org/10.1074/jbc.RA118.002839>
- Huber, C. D., Durvasula, A., Hancock, A. M., & Lohmueller, K. E. (2018). Gene expression drives the evolution of dominance. *Nature Communications*, 9(1), 2750. <https://doi.org/10.1038/s41467-018-05281-7>
- Katju, V., & Bergthorsson, U. (2019). Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biology and Evolution*, 11(1), 136–165. <https://doi.org/10.1093/gbe/evy252>
- Keinan, A., & Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, N.Y.)*, 336(6082), 740–743. <https://doi.org/10.1126/science.1217283>
- Kimura, S., Tahira, Y., Ishibashi, T., Mori, Y., Mori, T., Hashimoto, J., & Sakaguchi, K. (2004). DNA repair in higher plants; photoreactivation is the major DNA repair pathway in non-proliferating cells while excision repair (nucleotide excision repair and base excision repair) is active in proliferating cells. *Nucleic Acids Research*, 32(9), 2760–2767. <https://doi.org/10.1093/nar/gkh591>
- Lamesch, P., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1), D1202–D1210. <https://doi.org/10.1093/nar/gkg076>
- Li, C., & Luscombe, N. M. (2020). Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nature Communications*, 11(1), 1363. <https://doi.org/10.1038/s41467-020-15185-0>
- Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W., & Li, G.-M. (2013). The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell*, 153(3), 590–600. <https://doi.org/10.1016/j.cell.2013.03.025>

- Liu, H., & Zhang, J. (2022). Is the Mutation Rate Lower in Genomic Regions of Stronger Selective Constraints? *Molecular Biology and Evolution*, 39(8).
<https://doi.org/10.1093/molbev/msac169>
- Lu, Z., Cui, J., Wang, L., Teng, N., Zhang, S., Lam, H.-M., Zhu, Y., Xiao, S., Ke, W., Lin, J., Xu, C., & Jin, B. (2021). Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. *Genome Biology*, 22(1), 160.
<https://doi.org/10.1186/s13059-021-02381-4>
- Marsh, J. I., Kaushik, S., & Johri, P. (2026). Effects of rescaling forward-in-time population genetic simulations. *Genetics*, 232(2), iyaf263. <https://doi.org/10.1093/genetics/iyaf263>
- Martincorena, I., & Luscombe, N. M. (2013). Non-random mutation: the evolution of targeted hypermutation and hypomutation. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 35(2), 123–130. <https://doi.org/10.1002/bies.201200150>
- Monroe, J. G., Lee, C., Quiroz, D., Lensink, M., Oya, S., Davis, M., Long, E., Bird, K. A., Pierce, A., Zhao, K., & Runcie, D. (2025). Convergent evolution of epigenome recruited DNA repair across the Tree of Life. In *eLife*. eLife. <https://doi.org/10.7554/elife.105016.1>
- Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, 602(7895), 101–105.
<https://doi.org/10.1038/s41586-021-04269-6>
- Moore, L., Cagan, A., Coorens, T. H. H., Neville, M. D. C., Sanghvi, R., Sanders, M. A., Oliver, T. R. W., Leongamornlert, D., Ellis, P., Noorani, A., Mitchell, T. J., Butler, T. M., Hooks, Y., Warren, A. Y., Jorgensen, M., Dawson, K. J., Menzies, A., O'Neill, L., Latimer, C., ... Rahbari, R. (2021). The mutational landscape of human somatic and germline cells. *Nature*, 597(7876), 381–386. <https://doi.org/10.1038/s41586-021-03822-7>
- Mugal, C. F., von Grünberg, H.-H., & Peifer, M. (2009). Transcription-induced mutational strand

- bias and its effect on substitution rates in human genes. *Molecular Biology and Evolution*, 26(1), 131–142. <https://doi.org/10.1093/molbev/msn245>
- Ness, R. W., Morgan, A. D., Vasanthakrishnan, R. B., Colegrave, N., & Keightley, P. D. (2015). Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Research*, 25(11), 1739–1749. <https://doi.org/10.1101/gr.191494.115>
- Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., & Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science (New York, N.Y.)*, 327(5961), 92–94. <https://doi.org/10.1126/science.1180677>
- Oztas, O., Selby, C. P., Sancar, A., & Adebali, O. (2018). Genome-wide excision repair in *Arabidopsis* is coupled to transcription and reflects circadian gene expression patterns. *Nature Communications*, 9(1). <https://doi.org/10.1038/s41467-018-03922-5>
- Polak, P., & Arndt, P. F. (2008). Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research*, 18(8), 1216–1223. <https://doi.org/10.1101/gr.076570.108>
- Quiroz, D., Lensink, M., Kliebenstein, D. J., & Monroe, J. G. (2023). Causes of Mutation Rate Variability in Plant Genomes. *Annual Review of Plant Biology*. <https://doi.org/10.1146/annurev-arplant-070522-054109>
- Quiroz, D., Oya, S., Lopez-Mateos, D., Zhao, K., Pierce, A., Ortega, L., Ali, A., Carbonell-Bejerano, P., Yarov-Yarovoy, V., Suzuki, S., Hayashi, G., Osakabe, A., & Monroe, G. (2024). H3K4me1 recruits DNA repair proteins in plants. *The Plant Cell*, 36(6), 2410–2426. <https://doi.org/10.1093/plcell/koae089>
- Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35(10), 1720–1728. <https://doi.org/10.1093/bioinformatics/bty867>

- Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity*, 108(4), 447–455. <https://doi.org/10.1038/hdy.2011.95>
- Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412), 504–507. <https://doi.org/10.1038/nature11273>
- Soni, V., Pfeifer, S. P., & Jensen, J. D. (2024). The effects of mutation and recombination rate heterogeneity on the inference of demography and the distribution of fitness effects. *Genome Biology and Evolution*, 16(2), evae004. <https://doi.org/10.1093/gbe/evae004>
- Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., & Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4), 393–395. <https://doi.org/10.1038/ng.363>
- Staunton, P. M., Peters, A. J., & Seoighe, C. (2023). Somatic mutations inferred from RNA-seq data highlight the contribution of replication timing to mutation rate variation in a model plant. *Genetics*, 225(2), iyad128. <https://doi.org/10.1093/genetics/iyad128>
- Stoltzfus, A., & Yampolsky, L. Y. (2009). Climbing mount probable: mutation as a cause of nonrandomness in evolution. *The Journal of Heredity*, 100(5), 637–647. <https://doi.org/10.1093/jhered/esp048>
- Supek, F., & Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550), 81–84. <https://doi.org/10.1038/nature14173>
- Supek, F., & Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell*, 170(3), 534–547.e23. <https://doi.org/10.1016/j.cell.2017.07.003>
- Suter, B., Livingstone-Zatchej, M., & Thoma, F. (1997). Chromatin structure modulates DNA repair by photolyase in vivo. *The EMBO Journal*, 16(8), 2150–2160.

- <https://doi.org/10.1093/emboj/16.8.2150>
- Tuteja, N., Ahmad, P., Panda, B. B., & Tuteja, R. (2009). Genotoxic stress in plants: shedding light on DNA damage, repair and DNA repair helicases. *Mutation Research*, 681(2-3), 134–149. <https://doi.org/10.1016/j.mrrev.2008.06.004>
- Viguera, E., Canceill, D., & Ehrlich, S. D. (2001). In vitro replication slippage by DNA polymerases from thermophilic organisms. *Journal of Molecular Biology*, 312(2), 323–333. <https://doi.org/10.1006/jmbi.2001.4943>
- Watanabe, K., Yamada, N., & Takeuchi, Y. (2006). Oxidative DNA damage in cucumber cotyledons irradiated with ultraviolet light. *Journal of Plant Research*, 119(3), 239–246. <https://doi.org/10.1007/s10265-006-0266-2>
- Weng, M.-L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., Weigel, D., & Fenster, C. B. (2019). Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics*, 211(2), 703–714. <https://doi.org/10.1534/genetics.118.301721>
- West, M. A. L., van Leeuwen, H., Kozik, A., Kliebenstein, D. J., Doerge, R. W., St Clair, D. A., & Michelmore, R. W. (2006). High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. *Genome Research*, 16(6), 787–795. <https://doi.org/10.1101/gr.5011206>
- 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), e1004622. <https://doi.org/10.1371/journal.pgen.1004622>
- Willing, E.-M., Piofczyk, T., Albert, A., Winkler, J. B., Schneeberger, K., & Pecinka, A. (2016). UVR2 ensures transgenerational genome stability under simulated natural UV-B in *Arabidopsis thaliana*. *Nature Communications*, 7(1), 13522. <https://doi.org/10.1038/ncomms13522>

- Yampolsky, L. Y., & Stoltzfus, A. (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evolution & Development*, 3(2), 73–83.
<https://doi.org/10.1046/j.1525-142x.2001.003002073.x>
- Yan, W., Deng, X. W., Yang, C., & Tang, X. (2021). The genome-wide EMS Mutagenesis bias correlates with sequence context and chromatin structure in rice. *Frontiers in Plant Science*, 12, 579675. <https://doi.org/10.3389/fpls.2021.579675>
- Zhu, X., Xie, S., Tang, K., Kalia, R. K., Liu, N., Ma, J., Bressan, R. A., & Zhu, J.-K. (2021). Non-CG DNA methylation-deficiency mutations enhance mutagenesis rates during salt adaptation in cultured Arabidopsis cells. *Stress Biology*, 1(1), 12.
<https://doi.org/10.1007/s44154-021-00013-2>
- Zou, X., Koh, G. C. C., Nanda, A. S., Degasperi, A., Urgo, K., Roumeliotis, T. I., Agu, C. A., Badja, C., Momen, S., Young, J., Amarante, T. D., Side, L., Brice, G., Perez-Alonso, V., Rueda, D., Gomez, C., Bushell, W., Harris, R., Choudhary, J. S., ... Nik-Zainal. (2021). A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer*, 2(6), 643–657. <https://doi.org/10.1038/s43018-021-00200-0>