Using genealogies to study the genomic basis of species divergence

by

Arka Pal

November, 2025

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:
Ilaria Caiazzo, Chair
Nicholas H. Barton
Beatriz Vicoso
Graham Coop

The thesis of Arka Pal, titled *Using genealogies to study the genomic basis of species divergence*, is approved by:

Supervisor: Nicholas H. Barton, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Beatriz Vicoso, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Graham Coop, University of California, Davis, USA

Signature: _____

Defence Chair: Ilaria Caiazzo, ISTA, Klosterneuburg, Austria

Signature: _____

[Signed page is on file]

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Arka Pal
November, 2025

[Signed page is on file]

Abstract

Understanding the mechanisms underlying speciation is a central aim of evolutionary biology. A persistent challenge in the field is to identify loci that contribute to reproductive isolation, while disentangling signals of selection from demography, linkage and intrinsic genomic features. Traditional population genomic approaches that rely on site-based statistics in arbitrary fixed windows face inherent limitations, as they conflate historical and contemporary processes of divergence and overlook haplotype structure. Recent advances in whole-genome sequencing and methods to infer ancestral recombination graphs (ARGs) now offer the opportunity to study genealogical relationships explicitly, revealing how lineages coalesce and recombine through time. By directly analysing haplotype clustering by species or phenotype and their patterns of coalescence, ARG-based methods show promise for diagnosing sweeps, identifying barrier loci maintained under divergent selection amid gene flow, and tracing their evolutionary history.

In this thesis, I explore the utility of genealogical approaches for studying species divergence. In chapter 2, I propose a conceptual framework for defining haplotype blocks through the structure of the ARG, using simulations and empirical data to highlight how genealogical processes generate rich and often overlooked haplotypic patterns.

In chapter 3, I examine the genomic basis of a key evolutionary innovation in marine snails *Littorina*. These snails offer a unique opportunity to study an innovation because they include a very recent transition from egg-laying to live bearing, yet snails with the different reproductive modes are not reciprocally monophyletic. I exploited this by using topology clustering in ARG-derived local genealogical trees to pinpoint narrow genomic regions or haplotype blocks that carry swept alleles, thus revealing that the transition from egg-laying to live-bearing involves multiple, live-bearer-specific sweeps.

Chapter 4 establishes a population-scale, phased genomic resource for *Antirrhinum majus*, using cost-effective haplotagging, then optimizes imputation from low-coverage data against high-accuracy KASP sequencing to maximize sequence completeness with modest accuracy trade-offs against a traditional short-read sequence pipeline. A hybrid phasing strategy combines molecular phasing with statistical phasing to generate phased whole genome sequences of 1084 *Antirrhinum* individuals at a fraction of long-read sequencing costs.

In chapter 5, I analyse hybridising populations from two replicate hybrid zones to find a parallel genetic basis of flower colour, amidst the noise in genomic differentiation landscape driven by variation in demographic history. While outlier genome scans of F_{ST} failed to dissect the causes of differentiation, ARG-based topology clustering revealed a reuse of colour associated haplotypes across hybrid zones. In addition to the biological insight, this chapter also presents a comparison of the latest ARG inference tools, showing that signals of

topological clustering qualitatively agree between methods, despite differences in the tree sequences.

Next, in chapter 6, by leveraging ~1000 individuals in one of the hybrid zones, I integrated genome-wide association studies of floral pigmentation with genealogical inference, to test for additional colour loci, and confirm the effect of previously described loci. This work demonstrates that flower colour variation is driven by a small number of large effect loci, while also hinting at the presence of a new candidate regulatory factor.

Finally in chapter 7, in a preliminary analysis, I begin to dissect the genomic island of speciation around *Rosea/Eluta* to understand its evolutionary origins. My results show that it consists of 5 highly divergent loci, each of which is associated with flower colour. Using patterns of coalescence in genealogical trees, I find evidence of staggered selective sweeps and a persistent localized barrier to gene flow within an otherwise permeable genome.

Together, these chapters add to the increasing pool of studies using genealogical approaches to complement and extend site-based statistics to use haplotype structures in speciation research. By tracking haplotypes directly and connecting genealogical clustering to population processes, ARG-based inference promises to provide new insights into how local selective pressures, demographic history, and long-term barriers interact to shape the genomic architecture of divergence. By underscoring the value of ARGs in revealing the fine-scale origins and maintenance of biodiversity, this thesis presents cautious optimism about the benefits of using genealogical inference to learn more than what site-based statistics could tell us.

Acknowledgements

I feel incredibly lucky to have had Nick Barton as my PhD advisor. From population genetics to baking—you have always been kind and extremely generous with your thoughts, time, patience, and guidance. I genuinely cannot thank you enough, so, a mild concern: the mantra — 'Speciation is easy' — slightly worries me.

I am equally grateful to Sean. You never always showed up with your enthusiasm, ideas, comments, and encouragement from the early days of a lost grad student to the last-minute manuscript edits. Thank you for your unwavering support.

I want to thank the numerous mentors and collaborators who guided my work: Frank for countless insights into genomic analyses and Graham for making sure my time in Davis was both productive and fun. I also thank Beatriz for her thoughtful feedback and words of encouragement during every one of my progress reviews.

My thesis wouldn't exist without the 'true heroes', the numerous field interns and volunteers who, year on year, walked the 'dangerous' mountain roads in search of the snapdragons. Thank you for your hard work and company during the long days in Pyrenees, especially, Agnese, Anna, Helena, Mariona and Alba. Special thanks to Eva and David, for giving us the best field home anyone could ask for.

I'm thankful to ISTA—colleagues, IT, HPC, GSO, admin, HR and grant offices, and Vlad—for making all things, scientific and bureaucratic, run smoothly. I also acknowledge the following funding agencies and research grants – ERC (Advanced Grant: Haplotype Structure 101055327), and Austrian Science Fund (FWF) (Snapdragon Speciation P32166).

I am grateful to be part of the 'Bartonoid' family. You all have been, and I am sure will continue to be, fantastic colleagues and friends. First and foremost, Sofia, you are most certainly the best officemate one can hope for. I have cherished sharing every high and low of my grad school experience with you, both in the office and out of it. Thank for you being my buddy, my friend, and awesome fieldwork mate. Dasha, I have learnt so much from you, both personally and professionally; thanks for answering countless questions about bioinformatics, coding, Illustrator and most importantly, how to navigate PhD. Rosina, I admire your patience and passion in everything you do. I have learnt not only about ecology but also about leadership, from our shared time in the Pyrenees. Hilde, thanks for the comments on my thesis and manuscripts, scientific discussions and showing up in Chelsea spontaneously. I also want to thank Parvathy, Louise, Gemma and Hila among many others for their collaboration, friendship, comments, lunch breaks, the occasional pint and, most importantly, walking up to the bridge to get coffee with me.

I equally thank the 'Coopons', who has made my time in Davis so memorable. Importantly, you made me feel at home. First, thank you Gabrielle for the scientific discussions, friendship, the Delta of Venus lifestyle, a memorable trip to Mexico and the days

after. Thanks to James for the endless enthusiasm in ARGs. Thanks to Jeff for being a great colleague and friend, and sharing our musical interests, although we never got to jamming together.

I would not have made it through this PhD without the friendships I built along the way. Clara and Antony, you have given me a solid support system for the last decade, being there and never failing to show up through all the ups and downs. Natalia, your friendship has meant so much to me, thanks for every single dinner, bike ride, trip, conversation and more. Thank you, Stephan, you have seen it all and been there every time I ranted, celebrated, and partied, but most importantly, for being my favourite flatmate. Galien, Anna and Gianluca, the last five years wouldn't have been the same if you weren't there for all the parties, the weekends, the dinners, the trips, film nights and cocktail evenings. Nishchal, thanks for giving me that unfiltered friendship, and currently being my flatmate while I toil through writing my thesis when you distract me. Mario, Stefano and Charlotte, thanks for always being a phone call away, although I have failed to make that call in the last years and plan to change that. This list is endless, but I am incredibly grateful to Lukas, Dom, Ozzy, Peiping, Juancho, Lisa, Dagny, Leni, Fabienne and so many more for their friendship, love, and support. My time in Vienna would not have been the same without you.

Finally, thanks to my parents for giving me the support in everything I do, and my dear brother Remo, for being adventurous and pulling the not-so-rare drama and shenanigans, but mostly, for being the best sibling one could ask for.

PS: I thank anyone who attempt to read my thesis, cover to cover, within the first third of the 21st century. There is a simple scavenger hunt leading to a snapdragon fieldwork memory. If the reader can find the answer, I will surely thank them with either a snapdragon illustration or a beverage of their choice. Here's the first clue –

Summer fieldwork sun.
Some flowers choose where to bloom.
Begin at the start.

About the author

Arka Pal completed a BS with Biology Major at the Indian Institute of Science, Bangalore, and a double Masters in Evolutionary Biology from LMU Munich and Uppsala University. He joined ISTA, first as an ISTern at 2015, again for a research internship in 2018, and continuing as a PhD student since 2019. His research interest broadly involves studying evolutionary processes that promote speciation. To do so, he is keen on building and using tools that infer genealogical history of a population, that can help study important biological processes ranging from speciation to epidemiology. While at ISTA, he was actively involved in the annual fieldwork involving long term evolutionary study of a plant population in the Pyrenees. In his PhD dissertation, he worked on two different study systems (marine snails, *Littorina* and the common snapdragon, *Antirrhinum*), and has published two of his thesis chapters as lead author in *Molecular Ecology* and contributed to one more in *Science*. He has also presented his research at several international conferences, including ESEB 2023 in Prague, PopGroup 2024 in St. Andrews, SMBE 2024 in Mexico, and has been invited for seminars in Stanford University and UC Berkeley among others. He was awarded the Godfrey Hewitt Mobility Award from the European Society of Evolutionary Biology (ESEB) in 2024 to spend 3 months at University of California Davis as a visiting scholar. Alongside his PhD, he has held the position of track representative for Biology Track in the ISTA graduate school. He has also been actively involved in science communication, organising the Vienna chapter of the popular science festival, *Pint of Science* from 2020–2024. When not chasing the elusive snapdragon, he likes tall mountains, biking along rivers, obscure indie films, and all music from jazz to punk.

List of collaborators and publications

The chapters in this thesis have either been published or written in view of an eventual publication. As such, there is considerable repetition in the introduction of each chapter. Additionally, much of the work presented in this thesis is inevitably the result of collaboration with others, as detailed below.

Chapter 2 is published as:

> Shipilina, D.*, **Pal, A.***, Stankowski, S.*, Chan, Y.F.[§] and Barton, N.H.[§], 2023. On the origin and structure of haplotype blocks. *Molecular Ecology*, *32*(6), pp.1441-1457. https://doi.org/10.1111/mec.16793
> * Joint first authors, § Joint senior authors

All authors jointly conceived the idea, and wrote the manuscript. Nick Barton performed the theoretical simulations to explore the structure of haplotype blocks under neutrality and selective sweeps. Guided by the definition of haplotype blocks, I implemented a way to empirically identify edges as haplotype blocks from genealogical tree inference, demonstrated in an example of a selective sweep in *Heliconius* butterflies. Daria Shipilina and Sean Stankowski masterminded most of the visualisations in the chapter, with some menial contributions from me.

Chapter 3 is also published:

> Stankowski, S., Zagrodzka, Z.B., Garlovsky, M.D., **Pal, A.**, Shipilina, D., Castillo, D.G., Lifchitz, H., Le Moan, A., Leder, E., Reeve, J., Johannesson, K., Westram, A.M. and Butlin, R.K., 2024. The genetic basis of a recent transition to live-bearing in marine snails. *Science*, *383*(6678), pp.114-119. https://doi.org/10.1126/science.adi2982

The *Littorina* project was mostly led by Sean Stankowski and Roger Butlin at University of Sheffield. Much of the work, from sampling and sequencing to data analyses was led by them and other co-authors in the publication. A key methodological advancement presented in the paper, *TwisstNTern*, was conceived by Sean. My sole contribution is the genealogical analysis and identification of haplotype blocks that carry alleles involved in live-birth in *Littorina*, seen in Figure 3 of the publication. I am grateful to Sean, Roger and colleagues to have me as part of the team, and let me explore the utilities of ARGs in speciation related questions, that has since become a central theme of my thesis.

In **Chapters 4**, the idea of haplotagging 1084 snapdragon genomes was conceived by Nick Barton and Frank Chan at MPI Tubingen (currently at University of Groningen). Sean Stankowski led the sampling efforts in 2021 with further help by me and field interns, Mariona Vinyeta and Helena Ramirez. Sean and I performed sample processing, while Marek Kucka prepared sequencing libraries with Sean. I performed all the bioinformatic pipeline and genomic data analyses to understand accuracy and robustness of imputation and phasing, with supervision from Sean, Frank and Nick. David Field, Nick and Frank acquired the funding for this project.

Moving on, **Chapters 5** has again been published as:

> **Pal, A.**, Shipilina, D., Le Moan, A., McNairn, A.J., Grenier, J.K., Kucka, M., Coop, G., Chan, Y.F., Barton, N.H., Field, D.L. and Stankowski, S., 2025. Genealogical analysis of replicate flower colour hybrid zones in Antirrhinum. *Molecular Ecology*, p.e70067. https://doi.org/10.1111/mec.70067

Samples were collected during the annual fieldwork at Planoles between 2017 and 2019. I performed all formal data analyses albeit a couple exceptions, with primary supervision from Sean Stankowski, Nick Barton, Frank Chan and Graham Coop. Daria Shipilina inferred neighour-joining trees for the comparison of topology weights between tree inference methods in Figure 3, while Alan Le Moan performed the demographic modelling in Figure S3 and Table in the paper. Sean and I wrote the original manuscript together, while all authors commented on the manuscript.

In **Chapters 6 and 7**, I did all the formal data analysis, visualisation and writing except the following. Sean Stankowski and I manually scored flower photographs together, that was then used as the phenotype for GWAS. Additionally, Sean came up with the digital scoring system, *SnapPallete*. Parvathy Surendranadh performed the sigmoid cline fitting presented in Figure 2 of Chapter 6. Much of this work was primarily supervised by Sean and Nick Barton with additional comments from Frank Chan, Thomas Ellis and Magnus Nordborg. Graham Coop was instrumental in conceiving and shaping the genealogical analyses presented in Chapter 7.

Table of contents

Chapter 1
Introduction

1.1 | Understanding the genomic landscape of speciation—Aims and challenges

A central goal in evolutionary biology is to understand the genetic mechanisms underlying speciation (Coyne and Orr, 1998). Speciation is a continuous process where reproductive barriers accrue over time, restricting gene flow between populations (Feder et al., 2012; Seehausen et al., 2014; Stankowski and Ravinet, 2021; Wu, 2001). Key to this goal is the question: how do barriers to gene flow shape genomic variation during speciation, and what does this reveal about the number, distribution, and effect sizes of loci that underpin reproductive isolation? The resulting patchwork of genomic divergence defines a landscape that is the product of multiple, sometimes competing, evolutionary forces as well as intrinsic features of the genome (Ravinet et al., 2017; Wolf and Ellegren, 2017).

Over the past two decades, advances in sequencing technologies have ushered an era of evolutionary genetics devoted to identifying candidate genes and genomic regions involved in speciation—examining not only the build-up of reproductive isolation, but also the role of adaptation, population structure and gene flow among other processes in shaping species divergence. To this end, the field has employed a diverse toolkit: hybrid zone analyses that directly measure gene flow and selection in nature (Sobel and Streisfeld, 2015; Surendranadh et al., 2025), manipulative experiments that link genotype to phenotype (Bradshaw and Schemske, 2003), genome-wide association studies and admixture mapping to identify causal variants of traits that affect fitness (Buerkle and Lexer, 2008; Hooper et al., 2024; Lindtke et al., 2013; Todesco et al., 2020) as well as genome-wide scans of differentiation that identify loci contributing to reproductive isolation and adaptive divergence (Burri et al., 2015; Ellegren et al., 2012; Le Moan et al., 2024; Nadeau et al., 2012). These approaches reveal a dynamic interplay between selection, gene flow, and genomic structure, underscoring the varied routes by which populations can evolve into distinct species.

Despite rapid progress, interpreting the genomic landscape of speciation remains a formidable challenge (Rockman, 2012). Genetic differentiation between species—often quantified by summary statistics such as F_{ST}—can arise from numerous processes beyond those directly related to reproductive isolation (Bierne, 2010; Charlesworth, 1998; Cruickshank and Hahn, 2014; Ravinet et al., 2017; Wolf and Ellegren, 2017). Local adaptation, demographic history (including bottlenecks and founder events), variation in mutation and recombination rates, and genetic drift can all generate patterns of divergence that mimic the genomic signatures of real barriers to gene flow. As a result, identifying barrier loci from

genomic data alone is fraught with ambiguity, particularly when evolutionary history is complex or when populations have experienced fluctuating gene flow over time. The persistence of these ambiguities highlights a broader issue: while genomic scans can efficiently flag candidate barrier loci, they often fail to distinguish causal relationships from correlations, or to disentangle the effects of selection from those of demography and genome structure (Schluter and Rieseberg, 2022). The fundamental difficulty is that genomic approaches cannot determine the causes of divergence. *This suggests that much of the landscape is shaped by more than is immediately visible; as we will see, understanding where approaches meet their limitations is key.*

1.2 | Limitations to traditional approaches

While traditional population-genomic approaches have yielded major insights into genomic landscapes, they come with several limitations. First, they rely heavily on site-level summaries that dilute information about linkage and haplotypes. It has long been known that evolution works on linked segments of the genome, with allele frequencies shifting in blocks rather than at isolated sites (Fisher, 1954; Slatkin, 1972). Yet much empirical analysis and theory reduce genomes to one or, at best, a few loci treated independently. In practice, most population-genomic workflows rely on site-wise summary statistics (e.g., F_{ST}, π, SFS) or, at best, on arbitrarily defined fixed windows that are effectively assumed not to recombine (Beeravolu et al., 2018; Nielsen, 2005; Pavlidis and Alachiotis, 2017; Ragsdale and Gutenkunst, 2017; Weir and Cockerham, 1984). Treating SNPs individually can therefore mask key signals and blur causal interpretation. A peak of differentiation at a single site, for example, could result from direct selection on that site, selection on a nearby linked variant, or demographic forces acting on the broader region; standard summaries offer little leverage to discriminate among these explanations. Although population-genetic theory can, in principle, model interactions among many loci, computations become rapidly intractable beyond a small number, and truly multilocus results typically rely on strong simplifying assumptions (Barton, 1986; Hudson and Kaplan, 1995; Roze, 2021). These tools have been invaluable, but their independence assumptions and arbitrary windowing constrain inference about the genomic basis of speciation.

The second limitation is that traditional genome scans conflate signals from processes operating at different timescales, obscuring patterns relevant to contemporary gene flow. Consider secondary contact after a phase of allopatry: drift and selection during isolation create heterogeneous divergence across the genome even in the absence of current barriers. Upon contact, introgression erodes differentiation in permeable regions, potentially leaving peaks where barrier loci reside, but this erosion proceeds gradually and may remain incomplete. As a consequence, genome-wide summaries confound historical divergence with doepresent-day exchange, complicating efforts to pinpoint regions that currently restrict gene flow. Relatedly, most statistics provide poor temporal resolution by offering a cross-sectional view of present-day genetic variation, limiting the ability to reconstruct when barriers arose

or in what order they accumulated—information crucial for distinguishing among speciation modes. Even sophisticated methods for estimating population density, structure, and gene flow face fundamental difficulties distinguishing among a multitude of plausible population structures (Richardson et al., 2016; Sousa et al., 2011), because DNA sequences are highly correlated and each dataset reflects a single historical realization of a stochastic process, which imposes hard limits on inference. Until analytical frameworks harness the full information embedded in linked sequence blocks—rather than relying primarily on site-level summaries—the ability to separate the footprints of selection from the confounding effects of demography and genome architecture, and to define what is truly inferable, will remain constrained. Together, these limitations blur inference about the number, location, timing, and effects of loci that restrict gene flow. *These limitations highlight the need for unified analytical workflows capable of scaling across genomes.*

1.3 | The promise of haplotype-based approaches and ancestral recombination graphs

The limitations of traditional site-based approaches have motivated the development of haplotype-based methods that can capture the correlated inheritance of linked variants (Garud et al., 2015; Hejase et al., 2020a; Leitwein et al., 2020; Messer and Petrov, 2013; Sabeti et al., 2002). Haplotypes—the specific combinations of alleles inherited together on the same chromosome—carry richer information about evolutionary history than individual SNPs. They encode the joint imprint of mutation, recombination, selection, and demography on shared genomic segments over time. Beyond increased power, recombination also provides an additional "clock": the size and decay of shared blocks and ancestry tracts carry timing information about recent events that mutation alone can miss, improving resolution for contemporary gene flow, selection, and demographic change (Harris and Nielsen, 2013).

The idea to track a continuous genome dates back to Fisher's work on junctions that delineate genomic blocks of different ancestry (Fisher, 1954). Theoretical development of the standard coalescent model (Kingman, 1982; Wakeley, 2009) led to the introduction of the ancestral recombination graph (ARG) (Hudson, 1990). ARGs capture both the genealogical relationships among sampled individuals and changes to those relationships along the genome due to recombination (Hudson, 1990; Wakeley, 2009). It is therefore the most comprehensive description of the coalescence and recombination events that occurred in the history of a population. Although long understood, it has been limited to only theoretical studies.

Recent advances in computational methods have made it feasible to infer ARGs from genome-wide sequence data, opening new opportunities for studying speciation processes. Recently, efficient methods have been devised for simulating genealogies and follow populations, both forwards (Haller et al., 2025) and backwards in time (Baumdicker et al., 2022). Moreover, several approaches are now available for reconstructing genealogies along

the genome, including *ARGweaver* (Rasmussen et al., 2014), *tsinfer+tsdate* (Kelleher et al., 2019; Wohns et al., 2022), *Relate* (Speidel et al., 2019), and *SINGER* (Deng et al., 2024). While these methods differ in their assumptions and computational approaches, they all aim to reconstruct the sequence of genealogical trees that describes relationships among sampled chromosomes across recombining genomic regions.

The availability of genome-wide genealogies enables new approaches to studying speciation that were previously impossible (Campagna et al., 2017; Hejase et al., 2022, 2020b; Hooper et al., 2024; Meyer et al., 2024; Rueda-M et al., 2024; Wang and Coop, 2022). For example, genealogical methods can identify regions where samples cluster by species or phenotype rather than by geographic proximity, providing direct evidence for barriers to gene flow (see Chapters 3, 5). They can also estimate coalescence times between diverging lineages, and help reconstruct the timing of evolutionary events and distinguish between recent gene flow and ancestral shared variation (Hejase et al., 2020b; Wang and Coop, 2022). Furthermore, genealogical approaches can identify signatures of selective sweeps by detecting regions with unusually shallow coalescence times within species and elevated coalescence times between species (Barton, 1998). In principle, all the properties of statistics calculated from SNP can be calculated from the structure of the genealogy on which they occur, assuming neutrality (Ralph et al., 2020); it would be most efficient to make inferences directly from the ARG. However, in practice, genealogies are not known with certainty. It is then unclear whether it is best to make inferences from the estimated genealogies, allowing for their uncertainty, or whether we should instead rely on summary statistics calculated directly from the SNP.

1.4 | Integrating top-down and bottom-up approaches to understand speciation

Integrating bottom-up genome scans with top-down association mapping offers a way to address the limitations described in Section 1.2, particularly the dependence on site-level summaries and the blurring of historical and contemporary signals. By uniting these approaches, researchers can test causal mechanisms and directly evaluate how multiple barriers combine to limit gene flow. Studies that directly link candidate loci to phenotypes and fitness variation in nature provide the most robust evidence that loci are under divergent selection and generate barriers to gene flow (Bomblies and Peichel, 2022). A compelling motivation is the coupling hypothesis: strong reproductive isolation can emerge when multiple, initially independent barriers become coincident, creating spatially localized differentiation peaks or islands, that genome scans alone may misattribute to local adaptation or fail to link to the causal loci (Bierne et al., 2011). Framing studies around whether, where, and when barrier effects couple makes it necessary—not optional—to combine genomic scans, trait mapping, and explicit demographic context (Barrett and Hoekstra, 2011; Stinchcombe and Hoekstra, 2008).

However, integrating these approaches is challenging because they operate at different genomic and temporal scales and rely on different assumptions. QTL intervals are often broad, sometimes spanning large chromosomal segments due to limited recombination in lab-based crosses, whereas narrow F_{ST} outliers reflect finer-scale linkage breakdown over longer timescales and heterogeneous recombination, so apparent mismatches are expected even when causes are shared. Hybrid zones serve as powerful natural laboratories because they allow direct observation of ongoing gene flow, recombination, and introgression; recombination breaks down linkage disequilibrium to generate a fine-grained mosaic of haplotypes, thereby enhancing resolution for GWAS and admixture mapping, while restricted but persistent gene flow enables genome-wide scans to pinpoint loci that act as barriers to gene exchange (Barton and Hewitt, 1985; Gompert et al., 2017). However, this would still be a problem in hybrid zones with very strong reproductive isolation, since long ancestry tracts and persistent long-range linkage disequilibrium will still limit power of trait association by GWAS.

1.5 | Leveraging technological advances—Linked-read sequencing and haplotype reconstruction

Accurate inference of genealogies, haplotype-based inference from diploid data, and more generally, application of an integrative population genomic approach requires high-quality haplotype information across a large number of individuals (Browning and Browning, 2011; Pei et al., 2020). Traditional short-read sequencing excels at coverage and variant discovery but provides limited phase beyond short ranges (50–300bp), constraining haplotype-based inference across genomic distances (Browning and Browning, 2020). While long read sequencing (PacBio, Oxford Nanopore) offer greater read lengths (10–100Kb) that could significantly improve imputation, phasing, and structural variant detection, etc, it remains prohibitively expensive and lower throughput for large populations (Logsdon et al., 2020). Moreover, long read sequencing also typically requires high-quality, high molecular weight DNA, that can be challenging for non-model organisms. This has motivated linked-read sequencing strategies that preserve long-range linkage while retaining the throughput and cost profile of short-read platforms (Wang et al., 2019).

Haplotagging, a linked-read sequencing approach, provides a particularly promising solution for population-scale haplotype reconstruction (Meier et al., 2021). This method uses transposase-based barcoding to uniquely tag long DNA fragments (~10Kb) before sequencing, allowing short reads to be associated back to their originating molecules. Moreover, since long haplotypes can be observed directly, one can reliably use statistical methods to assemble these over much longer ranges, up to chromosomal scales (see Chapter 4). Thus, effectively long-read data can be obtained for essentially the same cost as with current short-read methods. This combination of long-range linkage information with high-throughput

sequencing makes *haplotagging* particularly suitable for studies requiring large sample sizes across diverse organisms.

1.6 | Study systems—Natural laboratories for speciation genomics

To address the challenges outlined above, this thesis leverages two well-established natural systems that provide complementary insights into different aspects of speciation: the marine snail *Littorina*, and the common snapdragon *Antirrhinum*. These systems represent different stages of the speciation continuum and offer distinct advantages for applying genomic approaches to study species formation.

First, the *Littorina* system provides an opportunity to study a recent evolutionary innovation—the transition from egg-laying to live-bearing—that has occurred within the past 100,000 years. This recent origin makes it possible to identify the specific genomic changes underlying a major phenotypic transition that has evolved repeatedly across the animal kingdom (Whittington et al., 2022). The system is particularly valuable because live-bearing is the only consistent difference between egg-laying and live-bearing populations, eliminating confounding effects of other phenotypic differences (Reid et al., 2012; Seshappa, 1947). Furthermore, the low genome-wide differentiation between reproductive modes provides a clean system to leverage genealogies for identifying specific genomic blocks that contain loci that contribute to this key innovation (Stankowski et al., 2020).

Second, the common snapdragon, *Antirrhinum majus* subspecies *majus*, offers a well-characterized hybrid-zone in which two varieties—*pseudomajus* and *striatum*—with strikingly different flower colour meet, hybridize, and maintain sharp phenotypic and genotypic clines despite overlapping ecological niches and shared pollinators (Khimoun et al., 2013; Surendranadh et al., 2025; Tavares et al., 2018; Whibley et al., 2006). The genetic basis of flower colour is resolved at key loci: magenta pigmentation is controlled by *Rosea*, *Eluta* and *Rubia*, while yellow pigmentation is controlled by *Sulfurea*, *Flavia*, *Cremosa* and *Aurina* (Bradley et al., 2025, 2017; Ono et al., 2006; Richardson et al., 2025; Schwinn et al., 2006; Tavares et al., 2018; Whibley et al., 2006). Prior analyses reveal elevated F_{ST} and steep clines at each of the colour loci (Surendranadh et al., 2025) while the rest of the genome is homogenised by gene flow (Ringbauer et al., 2018; Surendranadh et al., 2025).

The presence of replicate hybrid zones in different geographic locations makes it possible to apply genealogical approaches to distinguish between general patterns of divergence and location-specific demographic effects. Moreover, the tractable genetics of flower colour provides an ideal system for integrating top-down and bottom-up approaches to understand how known trait loci function as barriers to gene flow.

1.7 | Thesis overview

This thesis addresses the challenges and opportunities outlined above by developing and applying genealogical as well as integrative genomic approaches to study speciation in natural populations. The work is organized into seven main chapters that collectively advance our understanding of speciation processes by using whole genome sequence data.

Chapter 2 provides a theoretical foundation by developing a formal definition of haplotype blocks based on the structure of the ancestral recombination graph. This work addresses a fundamental conceptual issue in haplotype-based analyses by providing a rigorous framework for defining and interpreting haplotype structure. Using simulated examples, we demonstrate how different evolutionary processes—including neutrality and selective sweeps—generate distinct patterns of haplotype structure that can be detected using ARG-based methods. This chapter establishes the theoretical groundwork for the empirical applications in subsequent chapters.

Chapter 3 applies the genealogical framework developed in Chapter 2 to study the genetic basis of live-bearing in *Littorina* snails. Using whole-genome sequencing data from multiple populations, we identify genomic regions associated with reproductive mode through topology weighting of genealogical trees. The analysis reveals that live-bearing is associated with selection at multiple loci scattered throughout the genome, with evidence for selective sweeps. This chapter demonstrates how genealogical methods can provide insights into the genetic architecture and evolutionary history of key innovations that are difficult to obtain using traditional approaches.

Chapter 4 describes the development of a comprehensive methodological pipeline for processing linked-read sequencing data to generate high-quality, population-scale haplotype datasets. This chapter addresses the practical challenges of applying haplotagging to non-model organisms by optimizing approaches for genotype calling, imputation, and phasing. We benchmark the accuracy of different methods and provide guidelines for future studies seeking to apply linked-read sequencing to population genomic questions. The dataset generated through this pipeline provides the foundation for the analyses in Chapters 5-7.

Chapter 5 uses the *Antirrhinum* system to compare genealogical methods for studying speciation across replicate hybrid zones. We apply multiple approaches for inferring genealogical trees from genomic data and use topology weighting to identify loci associated with flower colour differences. The analysis reveals striking differences in demographic history between hybrid zones, highlighting the importance of demographic context for interpreting genomic patterns. We also provide a systematic comparison of different genealogical inference methods, offering practical guidance for researchers choosing among available approaches.

Chapter 6 integrates top-down and bottom-up approaches to study the genetic architecture of flower colour in the Antirrhinum hybrid zone. We conduct the first genome-wide association study for flower colour in this system and compare the results with population genomic scans and genealogical analyses. The integration of multiple approaches

reveals the challenges of applying GWAS in natural populations while confirming the role of known colour loci as barriers to gene flow. This chapter demonstrates both the potential and the limitations of combining different genomic approaches to study speciation.

Chapter 7 provides a preliminary exploration of the evolutionary history of flower colour loci using coalescence time estimates from ancestral recombination graphs. By comparing coalescence times around colour loci with background genomic regions, we explore the underlying evolutionary processes, and try to date the approximate time of the selective sweeps. This analysis illustrates how genealogical methods can provide temporal resolution that is difficult to achieve with traditional approaches, offering insights into the dynamics of barrier loci establishment during speciation.

Together, these chapters make several key contributions to speciation genomics. First, they discuss a possible framework for defining and analysing haplotype blocks based on genealogical relationships. Second, they demonstrate the practical application of genealogical methods to study speciation in two different natural systems, revealing insights that would be difficult to obtain using traditional approaches. Third, they develop and validate methodological approaches for generating high-quality haplotype data at population scale using linked-read sequencing. Finally, they illustrate both the potential and the limitations of integrating different genomic approaches to study the complex process of species formation.

Chapter 2
On the origin and structure of haplotype blocks†

Abstract

The term “haplotype block” is commonly used in the developing field of haplotype-based inference methods. We argue that the term should be defined based on the structure of the ancestral recombination graph (ARG), which contains complete information on the ancestry of a sample. We use simulated examples to demonstrate key features of the relation between haplotype blocks and ancestral structure, emphasizing the stochasticity of the processes that generate them. Even the simplest cases of neutrality or of a “hard” selective sweep produce a rich structure, often missed by commonly used statistics. We highlight a number of novel methods for inferring haplotype structure, based on the full ARG, or on a sequence of trees, and illustrate how they can be used to define haplotype blocks using an empirical dataset. While the advent of new, computational efficient methods makes it possible to apply these concepts broadly, they (and additional new methods) could benefit from adding features to explore haplotype blocks, as we define them. Understanding and applying the concept of the haplotype block will be essential to fully exploit long and linked-read sequencing technologies.

† This chapter is published and can be found online at: https://doi.org/10.1111/mec.16793

2.1 | Introduction

One of the breakthroughs of long and linked-read sequencing technologies is the emergence of new methods for obtaining reliable haplotype information for large data sets (Meier et al., 2021). Although most studies of genome-wide variation still focus on SNP data, we are approaching the stage where population-scale haplotype information will be widely available for organisms across the tree of life. In light of this shift from site-based to haplotype-based inference, this article considers one of the fundamental concepts for haplotype-based inference—the definition of the haplotype block.

"Haplotype" and "haplotype block" are widely used terms in evolutionary genetics, and have increased in importance across many disciplines (Delaneau et al., 2019; Leitwein et al., 2020; The International HapMap Consortium, 2007). An important, but often overlooked fact, is that populations evolve through changing frequencies of blocks of the genome, and not individual sites. Therefore, we should be most interested in understanding the trajectories of the underlying haplotypes, yet these are often obscured at the level of SNPs (Castro et al., 2019; Clark, 2004). Thus, disentangling the evolutionary history underlying genomic patterns can be challenging using solely site-based statistics. For example, while whole-genome scans for signatures of selection can reveal individual SNPs associated with fitness differences (Poelstra et al., 2014; Tavares et al., 2018), the actual causal loci/haplotypes can remain difficult to pinpoint (Burri, 2017; Grossman et al., 2010; Ravinet et al., 2017; Rockman, 2012; Stankowski et al., 2019; Tavares et al., 2018; Wolf and Ellegren, 2017). As another example, shifts in polygenic scores from genome-wide association studies (GWAS) can be misinterpreted as signals of selection, as opposed to artefacts of population structure (Berg et al., 2019; Novembre and Barton, 2018; Sella and Barton, 2019), which often leave clearer signatures in shared haplotype structure. Similarly, methods for estimating population density and gene flow struggle to distinguish among a virtually infinite number of possible population structures, made worse by assuming independence between SNPs, rather than haplotypes (Richardson et al., 2016; Sousa et al., 2011; Whitlock and McCauley, 1999).

By accounting for haplotype structure, it should be possible to make inferences more accurate and more efficient. Haplotypes carry information not only from *mutation* but also from *recombination*, which provides an additional 'clock' that can help to reveal past events (Ralph and Coop, 2013). Primarily for these reasons, there has been a steady increase in analytic methods that aim to infer haplotype structure from sequence data, or that exploit haplotype structure to make inferences about selection, gene flow, and population structure.

Although there has been significant progress toward the broader use of haplotype information in empirical studies (see overview in the Box 3), much of this lacks a unifying concept across fields spanning evolutionary and conservation genetics (Leitwein et al., 2020), human and medical genetics (Crawford and Nickerson, 2005), and animal and plant breeding (Bhat et al., 2021; Mészáros et al., 2021). There is thus often little consensus on how haplotype blocks are defined, which complicates comparison of results. Worse, it may

preclude insights that may otherwise arise from spotting commonalities that emerge under vastly different population parameters.

Motivated by the arrival of powerful new data sets and analysis methods, the main goal of this paper is to examine the fundamental definition of the haplotype block. We propose a definition of haplotype block based on the full genealogy, represented by the Ancestral Recombination Graph (ARG). Using simulations of simple but general scenarios, we explore how the characteristics of haplotype blocks relate to the origin of the samples and segregating SNP variation. We then discuss how the proposed definition relates to practical inference methods and their applications in large- scale population studies. We consider how different methods make use of haplotype information and infer haplotype blocks, their underlying assumptions and respective limitations.

2.2 | Defining haplotype blocks

A haplotype has a clear definition: it is simply a haploid genotype (for example, the genotype of the sperm or egg). In contrast, the term “haplotype block" is used widely, but in many different ways (Al Bkhetan et al., 2019; Clark, 2004; Schwartz et al., 2003; Taliun et al., 2014; Zhang et al., 2002). Since haplotype structure arises through segregation and recombination, our understanding of “haplotype blocks” must depend on the processes of coalescence and recombination that generate it in the first place. With this in mind, we contrast alternative definitions, and settle on one, which is based on branches in the underlying genealogy.

In sequence data, we usually observe the diploid genotypes; resolving them into the two haploid genotypes is termed "phasing". With n heterozygous sites, there are 2^n possible pairs of haplotypes—more than a million with just $n = 20$. However, in real populations there are usually far fewer haplotypes, due to linkage disequilibrium (LD) across polymorphic sites, which produces strong haplotype structure. This allows “statistical phasing”, through which one reconciles diploid genotypes into the underlying haplotype pair (Browning & Browning, 2011). Looking across individuals in larger genotype panels, the more frequent haplotypes often appear as stretches of shared, “banded” blocks of SNPs (Fig. 1A). This can be especially striking when different haplotypes become fixed across populations, which can produce block-like patterns in data even when individual haplotypes cannot be observed (Fig. 1B); in some cases, they have been referred to as ‘haploblocks’ (Todesco et al., 2020).

Whilst a blocklike structure may be apparent within empirical genetic data, we argue here that there should be a more fundamental definition of haplotype block, based on the true ancestry of the sequences, independent of the mutations that generated observable SNPs. Thus, we separate the *definition* of haplotype blocks from the *estimation* of these blocks from actual data.

There have been previous attempts at defining haplotype blocks via the classical concept of identity by descent (Carmi et al., 2013; Hartl and Clark, 1997; Thompson, 2013). Imagine an initial population, where each founder genome is labelled by a different colour. At some later time, each region of the genome must derive from one or other founder, and

so will appear as a mosaic of blocks of different colours, each corresponding to their ancestors. This naturally defines blocks that descend from a given set of founders (Fig. 2). Fisher showed that the junctions between IBD blocks segregate like Mendelian variants, and used this idea to understand the distribution of runs of homozygosity (Fisher, 1954).

In artificial populations, we can now sequence the founders, and thus directly observe blocks defined in this way (Lundberg et al., 2017; Otte and Schlötterer, 2021; Wallberg et al., 2017). Moreover, if we disregard new mutations, the evolutionary processes subsequent to the founding of the population are entirely described by the block structure. Identity-by-descent is usually defined with respect to a specific ancestral reference population (but note that coalescent definitions of IBD also exist (Wakeley, 2009). However, for natural populations, there is no obvious reference population, so the block structure will vary depending on our arbitrary choice of founders at an arbitrary time point (Fig. 2).

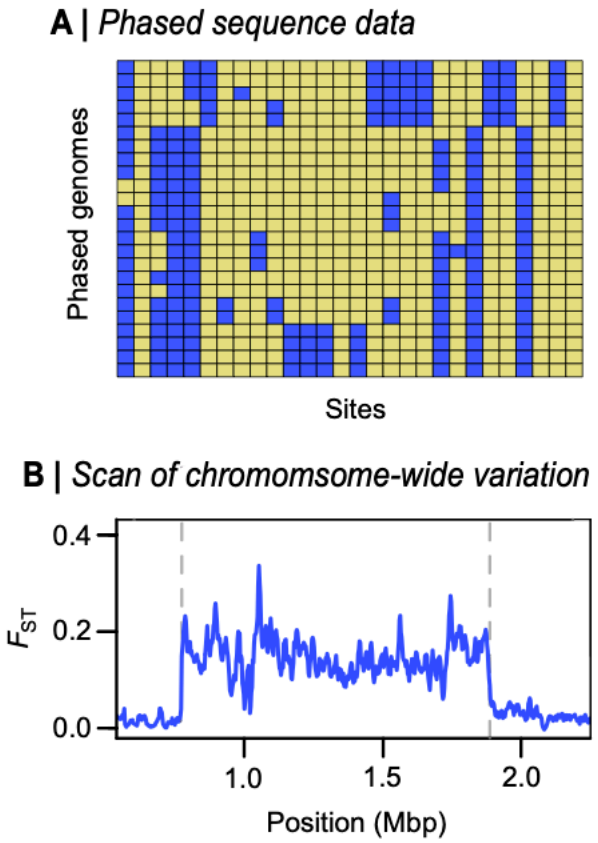

Figure 1. Block-like patterns in empirical data. (A) Block-like patterns in phased DNA sequences from *Mimulus aurianticus* within the gene MaMyb2 (Stankowski et al., 2015). Rows show 24 individual haplotypes. Each column is a site with yellow and blue squares representing ancestral and derived sites, respectively. (B) A F_{ST} scan across *Heliconius* chromosome 2 reveals a large plateau of differentiation on chromosome 2 between races of *H. erato* (Meier et al., 2021). This large block-like pattern coincides with a chromosomal inversion, the boundaries of which are illustrated by the dashed line.

To eliminate this subjectivity, we will base our definition of 'haplotype block' on the full ancestry of the sampled genomes, namely, on the ancestral recombination graph (ARG) (Hudson, 1983). The ARG consists of the segments of past genomes that are ancestral to our

sample; looking back in time, it is generated by a series of coalescence events that join lineages and of recombination events that split lineages (Box 1). We emphasise that these are real events: coalescence occurs when an actual individual leaves two or more offspring that are each ancestral to our sample, and recombination occurs between the two haploid parent genomes during meiosis in an ancestral individual. Together, these processes are embedded in the ARG (Fig. B1).

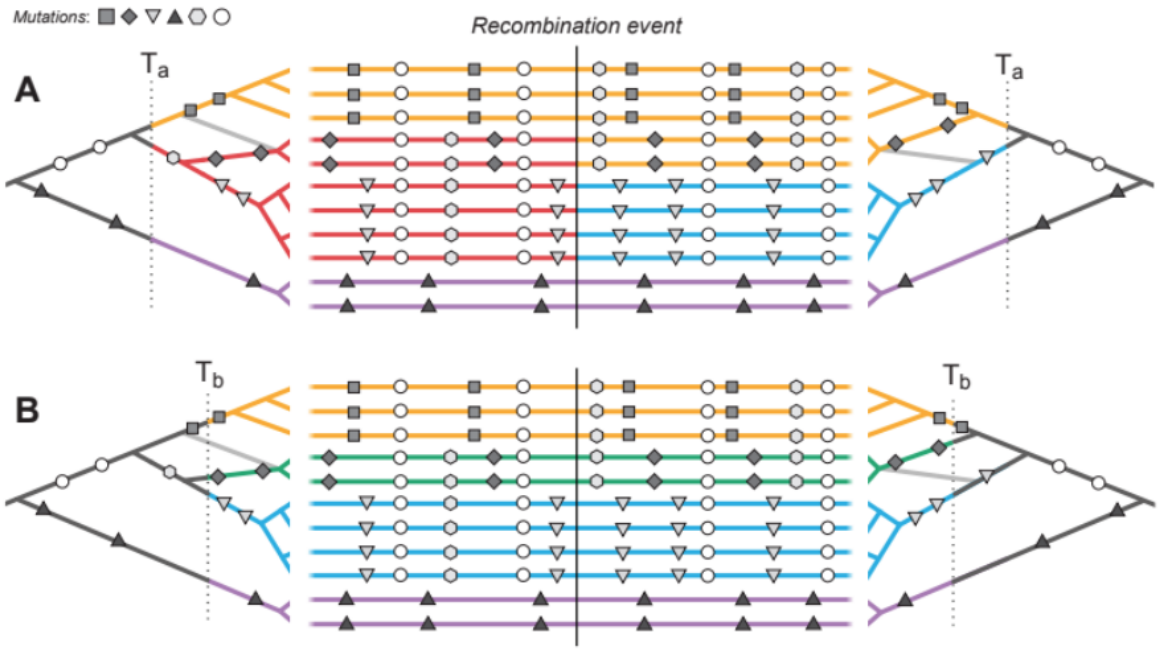

Figure 2. Haplotype blocks defined through identity by descent (IBD). Panels A and B show the same 11 hypothetical DNA sequences depicted as horizontal lines. The trees on the left and right sides show the genealogy for the set of sequences on either side of a recombination event (indicated by the vertical black line); the light grey branch in both trees shows how the effect of recombination changes the structure of thein the genealogy on either side. Mutations are shown as symbols that correspond to the branches upon which they arose. Under the IBD definition, haplotype blocks can be defined based on DNA segments that derive from a given set of ancestors, shown here by the coloured sections of branch and DNA sequence. The only difference between panels A and B is that these ancestors are defined at two different arbitrary time points, T_a and T_b, yielding different haplotype structure.

In large populations, and over long timescales, the ARG is approximated by the coalescent with recombination; in the simplest case, the rate of coalescence is the inverse of the effective (haploid) population size, and the rate of recombination is just the rate of crossover (Griffiths and Marjoram, 1997; Hudson, 1983). Importantly, the coalescent does not describe the entire genealogical relationship of the whole population, sampled or otherwise. Rather, it only summarises how the subset of sampled individuals are related to each other. Spatial and genetic structure can also be included: ancestral lineages carry a particular set of selected alleles (i.e., a particular genetic background), and are at a particular spatial location. Tracing back in time, lineages move between backgrounds by recombination, and between locations by migration.

Informed by the ARG, we could define a haplotype block as a contiguous region of the genome in which all sites share the same genealogy, i.e., a local gene tree. However, adjacent

genealogies differ by a single recombination event, and so blocks defined in this way will be vanishingly small (especially with large samples) and will usually differ trivially (see A in Fig. 3 and Fig. B1A). Moreover, as samples get larger, blocks defined this way will become so small as to be impractical.

Instead, we define a haplotype block as the set of genomic regions that descend from a particular edge in the ARG which is defined by a unique coalescence event, and by the set of descendant samples. Before elaborating on this definition, we clarify our terminology (see Table 1: Glossary). Consistent with the literature, we continue to use "haplotype block" to refer to a region of the genome with a shared pattern of ancestry, without forcing a precise definition. By "branch," we refer to a lineage on a genealogical tree that connects two coalescence events, or a sampled gene with a coalescence. By "edge," we refer to the extension of a branch along the genome. Thus, a branch is one- dimensional, with length measured in generations, whilst an edge is two- dimensional, with dimensions measured in generations as we trace back through time, and in Morgans as we trace along the genome. An edge is associated with a specific coalescence event, and also with a specific set of descendant samples.

This definition means that haplotype blocks exist completely independent of SNPs that may happen to arise on a given edge. However, if mutations have occurred, haplotype blocks will be associated with the set of derived SNP alleles that arise on the focal edge that just precedes the coalescence event. In other words, the set of haplotypes descending from this edge are distinct from all other sampled haplotypes in that they— and only they— share the set of SNPs occurring in the common stem lineage. If enough SNPs happen to arise on an edge, the haplotype block is revealed directly by these shared SNPs.

2.3 | Implications of the definition

We next elaborate on the definition and illustrate the relationships between genealogies, SNPs and haplotype blocks using example simulations (a neutral scenario and a selective sweep; Appendix S1 and accompanying GitHub repository: https://github.com/DaSh-bash/Suppl_Materials_On_the_origin_2022). The simulation uses the standard coalescent (Wakeley, 2009) to generate the ARG, thereby tracking the ancestors of a sample of genomes back through time, until all ancestral genomes are ancestors to the whole sample. It assumes a Wright– Fisher model with a constant population size 2N haploid genomes. A region of the genome of map length R is followed, with the selected locus at the leftmost point (i.e., at 0). For simplicity, we allow at most one crossover per generation, with probability R; we simulate $R << 1$, so this is close to the case with no interference between crossovers. The simulation can be conditioned on a selective sweep, which is defined by the numbers of copies of the favourable allele in the population. Once the ARG is constructed, genealogies along the genome can be followed, and edges can be identified. Neutral SNPs can be added, assuming

infinite- sites mutation; each SNP is associated with an edge in the ARG (more details on simulations in Appendix S1).

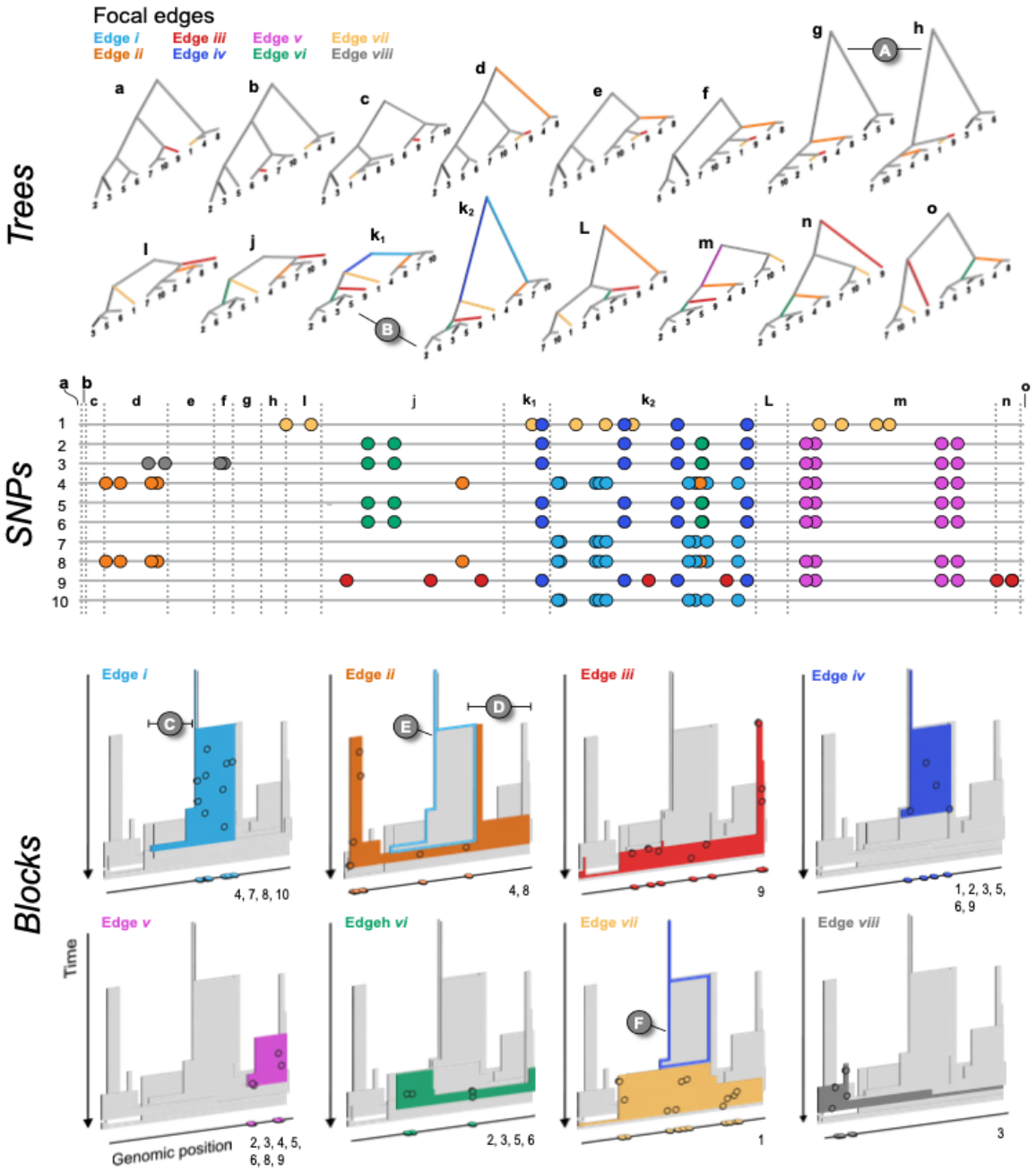

Figure 3. The relationship between **(a)** trees**, (b)** SNPs**,** and **(c)** haplotype blocks in the neutral simulation (see Main Text for simulation details). The ARG has been decomposed into marginal trees (*a*–*o*) to show all of the unique topologies that coincide with the genomic spans shown in the central panel (also labelled *a*–*o*). The branches for each tree are coloured according to the eight edges in the ARG that we chose to focus on (also labelled *i*–*vii*). A: two neighbouring topologies that differ only slightly due to recombination. B: an example of two trees (k1 and k2) that have the same topologies but different lengths. The central panel shows 10 haploid genomes (labelled 1– 10, top to bottom, coinciding with the tips of the trees). The SNPs that arose on the eight focal edges are indicated by the coloured circles. The lower panel (c) shows the haplotype blocks for each edge. The coloured block in each panel is the focal edge, with the other seven blocks shown in grey. The mutations shown in the central panel are projected onto each block (black circles) at the genomic location and time that they arose. Similarly, the numbers at the bottom right corner indicate which DNA sequences the mutations are associated with. C and D: Examples of regions of blocks that, by

chance, are not revealed by mutations arising on the corresponding edge. E and F: Examples of nested haplotype blocks, where the ancestral block is highlighted with a coloured outline

Figure 3 shows the relationship between trees, SNPs and haplotype blocks arising from the first simulation— a neutral example capturing the ancestry of 10 genomes, sampled from a population of 100 haploid individuals, across 10 centimorgans (cM) of the genetic map (Appendix S1). SNPs were generated by infinite- sites mutation with mutation at twice the rate of recombination. Despite the relatively short map and few individuals, this simulation is general because time and map distance both scale with population size (Hudson, 1990). Thus, the 268 generations taken for every part of the simulated genome to coalesce in a single common ancestor scales to 2.68 N, and the simulated map length scales to 10/N, where N is the effective size. Thus rescaled, this simulation shows a generic pattern, independent of population size.

The central panel of Figure 3 (middle panel, “SNPs”) shows the distribution of SNPs on the 10 sampled genomes, coloured according to the edge on which they arose (we illustrate eight edges with four or more SNPs each, out of 55 unique edges). Recombination events (24 in total) have divided the genome into 34 intervals due to nested recombination events (which split longer genealogies into nesting, inner intervals; Figure 3a). This illustrates how recombination modifies the coalescent (also see Figure A1 for a schematic representation of the process). The ARG can be decomposed into 24 unique marginal trees, some of which show an identical topology and differ only in timing (branch length); thus, 15 distinct topologies are shown in the Figure 3a (trees and corresponding regions on the genome labelled a–o; compare k1 and k2 for an example of genealogies that share topology but differ in depth, B in Figure 3).

The coloured blocks shown in the lower panel of Fig. 3 (‘Blocks’) illustrate the extent of each edge along the genome, and through time. The mutations arising on each branch are projected onto the block at the time and genomic position that they arise. The number of SNPs arising on each branch is Poisson distributed, with the expected number proportional to the area of the block; this area is the sum of the genomic lengths that each ancestor carries, and that is ancestral to the coalescence event that defines the branch. We emphasise that the visualised colour blocks represent *true* genealogies—and are independent from mutations. Because mutation, or SNP occurrence, is a random process, some regions may not carry any informative SNPs. For example, though branch *i* (light blue) is relatively well covered by 9 SNPs, none of them fall in the shallow region to the left (C in Fig. 3). Similarly, branch *ii* has only 6 SNPs, none of which happen to fall in the rightmost region (D). Ultimately, the distribution of SNPs sets a limit on what can be *inferred* from sequence data; branches without mutations will be invisible to us, and our ability to infer the length of a block depends entirely on where mutations happen to fall.

Each edge coincides with a specific coalescence event that brings together a specific set of lineages: in other words, edges/branches are defined by both the coalescence event *and* the set of lineages. A single coalescence, i.e., a single ancestor, may generate multiple edges: the two genomes that come together in that event may carry a mosaic of ancestral material, in several combinations. A single coalescence event may even generate an edge that

carries disjunct segments of the genome. This did not occur for any of the focal edges in the example of Fig. 3, but is not unlikely, especially in a selective sweep. Conversely, two different coalescence events may happen to bring together the same sets of lineages; their edges could only be distinguished through the different times of coalescence.

Because each edge is generated by a single coalescence, it begins at the same time across its whole extent (so, edges are bounded by a horizontal line at their base in the lower panel of Fig. 3). Recombination events split distal segments, thus limiting the span of the block along the map. Tracing back in time, edges must end in coalescence events that combine them with yet more descendants. These may occur at different times if there have been recombination events, so that the upper boundary is typically ragged.

Haplotype blocks overlap in their genomic extent, since multiple lineages exist at any time after the MRCA; this is shown by the overlapping 3-D blocks in Fig. 3 ('Blocks'). Haplotype blocks will also overlap in the genome when edges are nested in the genealogy, giving rise to nested haplotype blocks. For example, branch *ii* (orange), which is ancestral to genomes 4 and 8 descends in the middle part of the genome from branch *i* (blue), which is ancestral to genomes 4, 7, 8 and 10. Thus, haplotype block *i* is nested above block *ii* in Fig. 3 (see also F for another example of nested haplotype blocks).

If we start at a particular site on the chromosome, and work along the genome, at some point an edge will be split by a recombination event. If the recombination occurs on the edge itself, the edge will persist, but probably with a different depth. If the recombination event occurs out with the lineages that descend from the focal coalescence, but coalesces into those lineages, then the set of descendants will be augmented, and the edge will end. Conversely, if the recombination event occurs among the descendant lineages, then some descendants will be lost, and the edge will again end.

As we work out from a given locus, the incidence of recombination is proportional to the branch length, and so we expect that if a branch traces back deep into time, it will extend over a short region of the genome. Conversely, shallow branches will extend over a longer genomic span. This pattern is seen clearly in Figure 3c, where edges consist of segments that are either deep and narrow, or shallow and wide. However, this relationship is not precisely inverse; if it were, edges would tend to have the same area, whether they were deep or shallow, and hence would carry similar numbers of SNPs. In fact, the distribution of areas of blocks is highly skewed, and so most SNPs are on a few deep branches (see discussion on branch depth in Appendix S1).

Note that under the coalescent process, large numbers of sampled lineages rapidly coalesce down to a few, which are then likely to trace back deep into the genealogy. Thus, in a given region of the genome a substantial fraction of SNPs will fall on long, deep, branches, whereas the tips of the genealogy will be hard to resolve. Moreover, in a large sample, it is unlikely that different coalescence events will bring together exactly the same set of lineages by chance, so that we can usually identify unique coalescence events as corresponding to unique sets of lineages. This is one reason why haplotype-based analyses can be particularly useful in disentangling genetic structure.

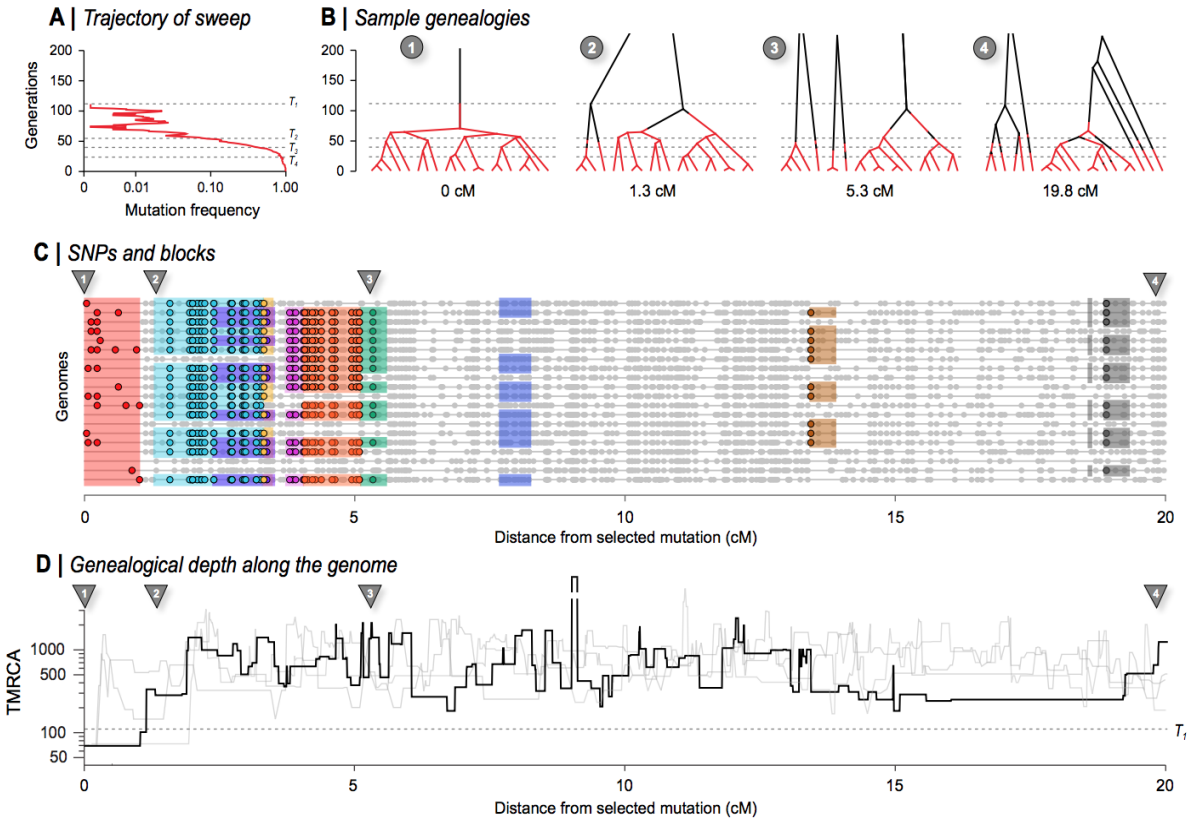

Figure 4. The effects of a recent selective sweep on linked genealogies. (A) A mutation with advantage 10% arose in a population of 400 haploid individuals, and swept to fixation in 110 generations, at which time 20 genomes were sampled; 20cM of the genome is followed back in time, with the selected locus at the left.; dashed lines (T_1 - T_4) show times when the favoured allele was in 1 copy, at 10%, at 50%, and at 90% (110, 53, 38, 22 generations back). (B) shows genealogies at positions 0, 1.3cM, 5.3cM, and 20 cM, branches are coloured in red when on the fitter background, and black when on the ancestral background. Thus, changes in colour show recombination events that change the genomic background. Note that such events are unlikely when the allele is near fixation (i.e., at the base of the tree, below the lower dashed line), and conversely, become common whilst the allele is rare, simply because it will almost always meet with the opposite background. Before the mutation occurs (i.e., above the upper dashed line) lineages must either trace back to that mutation (top left) or recombine out into the ancestral background; thus, all lineages must appear black above the upper dashed line (110 generations back). Note that the disjunct branches in trees 2 - 4 all coalesce further back in time, but only 200 generations are shown for visibility. (C) shows SNPs along the 20 sampled genomes. The 9 of the most substantial branches are shown. (These have more than 8 descendants, formed by coalescence more recently than the sweeping mutation, and have areas >0.5). The red block at the left shows the region linked to the selected locus, which coalesces in a single common ancestor 69 generations back, just after the sweeping mutation arose. Grey dots show those SNPs that are not on these 9 highlighted branches. (D) shows the time back to the most recent common ancestry (TMRCA) along the genome, on a log scale. The bold line shows the example simulated above, whilst the three grey lines show replicates, generated conditional on the same sweep; the break in the line shows an area where the TMRCA extends further back than the extent of the y-axis. The dashed line across the plot corresponds to T_1 in panel A.

Figure 3 illustrates the simplest case of the standard coalescent with recombination. In reality, population structure and selection complicate genealogies. For example, in the island model, lineages either coalesce quickly within a deme, or escape to coalesce much further back in time. This exaggerates the tendency for genealogies to be dominated by a few long branches (Wakeley, 2009). Selective sweeps have a somewhat similar effect (Maynard

Smith and Haigh, 1974). In the classic case, all lineages at the selected locus coalesce in the individual that carries the favoured mutation. Moving out from this locus, recombination frees lineages to coalesce much further back.

Figure 4 illustrates such a selective sweep (Supplement 2). The sweep greatly reduces diversity around the selected locus, because all lineages must trace back to the successful mutation (Fig. 4B). This region of complete coalescence is shown in red, but note that it contains some diversity, due to mutation subsequent to the sweep. As we move away from the selected locus, lineages recombine out onto the ancestral background, and coalesce with the rest of the genealogy much further back (Fig. 4B). This process can be seen in the time to the MRCA (Fig. 4D), which jumps from a low value at the selected locus, through successive recombination events, back to a time that fluctuates around $4N_e$=800 generations, under the standard coalescent. However, the replicates in the lower panel show that there is considerable variation in this process, which sets a fundamental limit on our power to detect a sweep and estimate its properties.

At the selected locus, all lineages coalesce in the favoured mutation. Successive recombination events each free one or a few lineages from the new background, so that the exceptionally large and recent cluster gradually diminishes in size, until the genealogies follow a close to neutral distribution. Thus, edges with large numbers of descendants are associated with the sweep, and can be distinguished by the characteristic sets of SNPs that they carry; nine such edges are illustrated in Figure 4c.

We close this section by commenting on possible connections between our description of the ARG and practical inference. Stern et al. (2019) proposed a method that infers the allele frequency trajectory from the genealogy at the selected locus, which is assumed to be known. The extent of the focal edge along the genetic map gives additional information, with a predicted constant rate of recombination out into the ancestral background, at a rate equal to the frequency of the ancestral allele. Additional edges give more information: in particular, several lineages may coalescence early in the sweep, but then recombine out (e.g., the second genealogy in Figure 4b). This generates multiple long branches, whose distribution depends on $4N_e$ value (Barton and Charlesworth, 1998). There is considerable scope for using the extent of edges along the genome, as well as the genealogy at specific loci.

Nevertheless, we make two cautionary comments. First, there is considerable variability between different realisations, given the same trajectory (e.g., Fig. 4D) (i.e., the selective sweep itself is a stochastic process). Moreover, if the locus is identified from a genome-wide scan, ascertainment bias will distort the ARG: indeed, sequence variation around a neutral locus that experiences a sweep *by chance* may be indistinguishable from a genuinely selected locus. This poses fundamental limits to our ability to estimate selection at a particular locus. Second, sophisticated methods based on simple scenarios will be confounded by deviations from the model. For example, the extent of reduced diversity along the genome is the inverse of the time taken to reach high frequency - but that may be greatly increased by population structure. The visualisations that we develop here may have the

greatest value in allowing us to check whether the fine structure of a candidate region is actually consistent with some simple model. It remains to be seen how far the rich information contained in the structure of such branches will help us improve our inferences.

2.4 | The definition in practice

Having defined haplotype blocks conceptually, we next consider the problem of inferring haplotype blocks from empirical datasets. Current sequencing and genotyping technologies make it straight-forward to identify SNPs or small insertions/deletions, but it remains non-trivial to connect these to the haplotypes in which they are embedded. For that reason, sophisticated algorithms have been developed for phasing, imputing genotypes and inferring genealogies (Browning and Browning, 2013, 2009; Davies et al., 2021, 2016; Howie et al., 2011; Marchini et al., 2007). These tasks all engage different facets of the same problem, and rely to various extent on the haplotype structure. However, these methods tend to focus on phasing and stop short of inferring underlying haplotype structure and in particular the ARG, and haplotype blocks as we define them. In this section, we wish to focus on how haplotype blocks can be defined and visualised in practice. Given our ARG-based definition, we used ARGweaver (Rasmussen et al., 2014) to analyse an empirical dataset. We discuss other methods that use the ARG or approximations to it. We discuss the underlying assumptions of these methods and highlight where they could be extended to capture further information in light of our proposed definition of haplotype blocks as edges. Separately, in Box 2, we outline classes of simpler methods that use fixed genomic windows or genomic segments as a proxy for the haplotype block.

The full ARG contains all the information needed to apply the haplotype block definition to empirical datasets. Therefore, one could start by inferring the ARG (or even a set of genealogies along the genome) from a sample of sequences, identifying important edges on that ARG, and consequently the haplotype blocks that descend from those edges. ARGweaver (Rasmussen et al., 2014) and its extension ARGweaver-D (Hubisz et al., 2020) are among the most powerful tools for direct inference of ARG. One practical speed-up employed by ARGweaver is to discretize time, effectively making the ARG space finite by limiting recombination and coalescence events to discrete time points. Further, ARGweaver uses an approximate model, sequential markovian coalescent (SMC) (McVean and Cardin, 2005), extended by Marjoram and Wall (2006) to sample from a distribution of ARG. While making inference more tractable, the SMC precludes the inference of disjunct blocks, because only one immediately prior state is considered as one moves along the genome. However, even with these key innovations, inference of the “full” ARG remains computationally expensive, making ARGweaver feasible for up to approximately 50 samples.

To illustrate our concept, we applied ARGweaver to infer the ARG from an empirical, phased data set from *Heliconius erato* butterflies. The data set was generated by haplotagging, a technique for producing linked-read sequence data (Meier et al., 2021). We focus on the genomic region containing the gene *optix*, where a selective sweep was

previously inferred using site-based statistics (Figure S3, Appendix S2). For comparison, we also sampled ARGs from a neutral background locus (Figure S3: Appendix S2). Figure 5 shows a focal region (~3 Mb long) located ~100 kbp upstream from optix that may correspond to a distal regulatory hub controlling the distinctive wing rays— "Ray" and "Dennis" elements (Wallbank et al., 2016), possibly corresponding to *obs132*, *LR1/2* and *obs214* (Lewis et al., 2019)—at which all lowland *H. e. lativitta* butterflies (red labels in Fig. 5a) share a haplotype (Meier et al., 2021). To run ARGweaver, we used $\mu = r = 2.9 \times 10^{-9}$ and $N_e = 1.94 \times 10^6$ (calculated by estimating π from the neutral region) and 30 discrete exponentially distributed time points (Fig S1). The key step for visualising haplotype blocks is identifying unique edges based on the sampled ARGs. We identified edges using custom scripts to parse marginal trees along the genome in order to identify branches that originate at a particular coalescent time-point (say, t_1), and that are ancestral to a fixed set of individuals (say, x_1, x_2, x_3) (see Table S1, S2). This set of branches represents an edge that is defined by $\{t_1\},\{x_1, x_2, x_3\}$ and the trees that contain the above branches (see branches in Fig 5A coloured according to haplotype blocks as edges in Fig. 5C). Here, we chose to focus on the 6 edges that are supported by 3 or more SNPs (Fig 5B: SNPs; 5C: haplotype blocks as edges) (see Table S3 for a list of all edges supported by SNPs). We then visualised haplotype blocks based on these edges, including both their genomic span (i.e., the tree-spans along the genome containing the edge) and temporal span (length of each tree branch that constitutes the edge) (Fig. 5C).

We find a rich structure of haplotype blocks in the focal region. First, in regions of shallow coalescence, we see structured uninterrupted haplotype blocks that carry SNPs fixed in all *H.e.lativitta* samples (Fig. 5). For example, the shallowest coalescence region constitutes the uninterrupted haplotype block defined by edge *iv*, and is supported by 9 SNPs. Adjacent to edge *iv*, we find the largest haplotype block defined by edge *iii* and supported by 22 SNPs, which coincides with a putative selective sweep that was previously identified using the omega statistic (Wallbank et al., 2016). Unlike edge *iv* (green), edge *iii* (red) exists as disjunct blocks separated primarily by recombination events that shifts the total coalescence within *H.e.lativitta* samples to occur slightly further back in time (see Trees 3,4,5,6 for reference). Moving along in both directions from the central shallow TMRCA region, we observe other haplotype blocks, exhibiting different histories (originating deeper in the past) spanning as disjunct blocks along the entire genomic region and supported by fixed SNPs (edge *i* and *ii)*. Although the SMC precludes ARGweaver from inferring disjunct coalescence events, disjunct blocks (especially, edge *i*) may be an artefact of the way we identify unique edges from the ARGweaver output, together with discretization of time points (in principle, they could be distinct coalescence events but are forced to coalesce at the same time). Moreover, these blocks can also stem from including the *H.e.notabilis* population in our analysis (Trees 1,2,5,6 has one or more *H.e.notabilis* individual clustering together within the *H.e.lativitta* population), and hence are present as nested blocks within edge *iii* and *iv* in the central regions of the shallowest coalescence. In future, it would be interesting to examine if additional evidence can suggest older, historical sweep events to explain the disjunction of these blocks, or whether they are simply remnant structures/signatures from other stochastic

events. It is important to note that 4 out of the 6 substantial edges drawn here are supported by SNPs fixed within the *H.e.lativitta* samples, and spans in disjunction (except edge *iv*) throughout the region. This is in agreement to theoretical expectations in a swept genomic region where uninterrupted edges are expected to exist with multiple SNPs supporting them (edge *iv* - 9 SNPs, *iii* - 22, *ii* - 16, *i* - 14). In addition to these 4 blocks, we also see more recent blocks with greater spans, explained by singletons (see Fig S7) and doubletons (*edge v* and *vi* in Fig 5). Edge *vi* specifically spans the furthest along the genome, since it is formed by a coalescent event between only two samples, suggesting that there has not been sufficient time for recombination to break it down.

Our ARGweaver analysis shows the complex relationships between SNPs, edges and haplotype blocks inferred from real data, but also demonstrates the possibility of visualising haplotype blocks as edges, and utilising statistics from these blocks as a signal to make further evolutionary inference. This could potentially produce alternative hypotheses, such as multiple selective sweeps, or differentiate between sweeps and random shallow coalescence events. However, we should point out several limitations to such analyses. First, since there is an infinite number of possible genealogical histories, inference of the full ARG comes with a degree of uncertainty. Specifically, ARGweaver estimates a coalescent model from the data (set of SNPs) and produces a distribution of trees that is consistent with the data, but only supported strongly at and around the SNPs. In other genomic regions, the model produces a random distribution of trees given the parameters that are still consistent with the data but not necessarily supported by any SNPs (see Fig S9 for comparison between haplotype block structures between MCMC iteration). This suggests that inferences should only be made only from edges robustly supported by the SNP configuration, rather than the full ARG. Due to its computational tradeoffs, ARGweaver inference can also be prone to differences in user choices such as discrete time-points, recombination/mutation rates, estimation of effective population size, MCMC parameters. Despite these limitations, features of haplotype block structures from our analysis can carry potentially important features in non-neutral regions of the genome. Here, we simply demonstrate with a small example one way to identify significant edges and haplotype blocks from empirical data. Although beyond the reach of this paper, we hope that the rich haplotype structure revealed here can spur development of new methods that take advantage of different layers of information.

The computational requirement and feasibility of ARGweaver were addressed by two other methods - *tsinfer* (Kelleher et al., 2019) and *Relate* (Speidel et al., 2019) that attempt to approximate the ARG in much larger populations with thousands of samples by focusing on topology (or 'succinct tree sequences'), rather than a full inference of the ARG. They do so by representing genomes as a series of tree topologies: Relate as distinct trees; tsinfer as 'tree sequences' connected via ancestral haplotypes. Both achieve this remarkable speed-up by relying on the Li and Stephens' hidden Markov model (Li and Stephens, 2003) see Box 2 for further details) to infer local pairwise distances (Relate) or ancestral haplotypes (tsinfer). As an added advantage, tsinfer doubles as an efficient, lossless compression algorithm by indexing population genomic variation as SNPs-on-trees as opposed to the traditional (and

highly redundant) SNP-by-individual matrix (implemented as a tskit library; Kelleher et al., 2019). Put another way, the tree sequence encoding can fully capture the variation data in entire populations, for a fraction of the storage space. Such a representation also effectively encapsulates a number of population genetics summary statistics (Kelleher et al., 2019; Ralph et al., 2020). These developments may prove essential, as sequencing of entire national populations increasingly becomes routine.

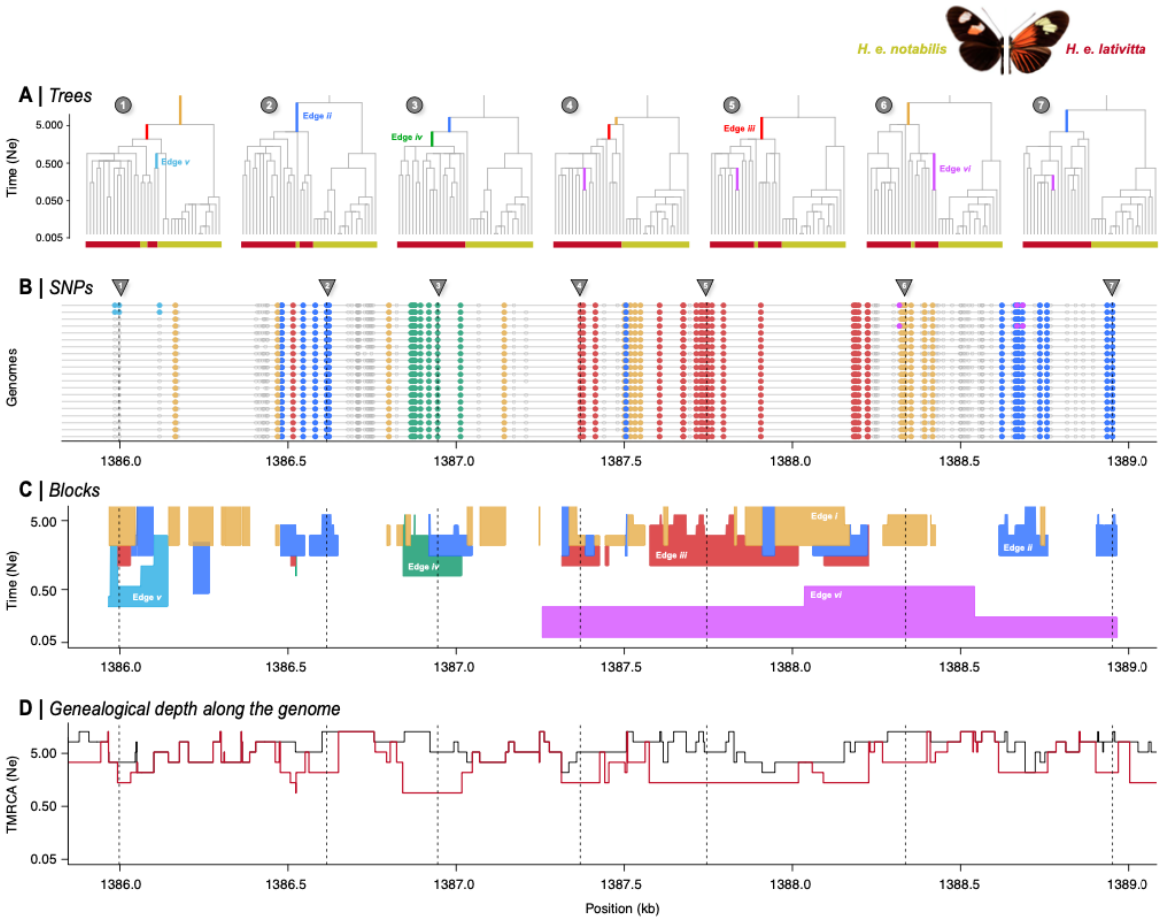

Figure 5. Visualisation of haplotype blocks based on application of ARGweaver to the *optix* region of *Helconius erato* butterflies (*2n* = 20). ***A:*** Genealogical trees at each genomic point, marked by arrows and corresponding numbers. *Red* labels - *H.e.lativitta* samples; *Yellow* labels: *H.e.notabilis* samples. Branches on trees are coloured according to the corresponding edges in *C*. Both highland *H. e. notabilis* population (2*n* = 20) and lowland H.e.lativitta (2n = 20) are included in the analysis so that we can estimate the length of branches that are ancestral to all *H.e.lativitta* samples (2*n* = 20), however *B,C,D* only exhibits features related to H. e. lativitta population. ***B:*** Genomic location of SNPs for all 20 *H.e.lativitta* haploid samples. Out of a total of 137 SNPs, only those that explain the 6 substantial edges are coloured accordingly. Edges are defined as substantial if 3 or more SNPs occur on them. 2 out of 6 edges (*v* and *vi*) are explained by doubleton SNPs, whereas the rest are fixed within the *H.e.lativitta* samples. SNPs that do not appear on any significant edge are coloured grey if they have higher allele frequency within the samples, white otherwise. ***C:*** Visualisation of haplotype blocks as edges similar to Figure 3 - plotting blocks along the genomic (*x-axis*) and temporal span (*y-axis*). Since edges always originate at a fixed coalescent time point, the bottom line of the block is always smooth. The ragged tops of the blocks denote the length of the edges interrupted by recombination events. Note that edges can also be disjunct due to one or more samples and recombining out and back into the same lineage. ***D:*** 50-percentile TMRCA estimates along the genome: black: total TMRCA to the all 40 samples, red: TMRCA to only *H.e.lativitta samples*.

Among practical methods, tsinfer and Relate are the state-of-the-art in representing large populations. All three approaches, including ARGweaver, approximate some aspects of the ARG well, and give an accurate distribution of coalescence time under simulation of the standard coalescent (Brandt et al., 2022). For our purposes, they are also useful approximations to the ARG that highlight some of the key advantages we wish to emphasise in our haplotype block definition. For example, Relate presents a suite of statistics that goes beyond SNP information. One advantage of Relate is that branches are dated, as opposed to a strict encoding of topology alone in tsinfer. Having dated branches allows, among other things, the possibility of estimating temporal changes in mutation rates. Another useful feature, in our view, is tsinfer's placement of SNPs onto branches, which is the essential feature that distinguishes haplotype blocks from each other under our definition, even though our definition is independent of the SNPs themselves.

We note that efforts are already underway to bridge across methods and address their limitations. For instance, tsdate now adds coalescence times estimates and branch lengths from tsinfer's output (Wohns et al., 2022). In the context of our exploration of haplotype blocks and their overlapping structure (Fig. 3C, D), we have noted that they may not be accurately captured under the Li–Stephens models in tsinfer and Relate, in a way that may bias the inferred ARG. However, this is an open question, so more work is needed to understand how different methods perform across a range of parameters relevant to non-model organisms.

In summary, there has been a recent spurt in innovation in genealogy/ARG-based methods. Among these, ARGweaver arguably comes closest to inferring the full ARG, but at considerable computational cost. Both tsinfer and Relate are robust and scalable to thousands of samples with minimal, reasonable tradeoffs, but infer haplotype blocks only as an incidental output. Ultimately, we hope our discussion here will encourage development of new methods to infer haplotype blocks as we define them, and to use these for further explanation and inference.

Assuming that a method becomes available for inferring blocks as we have defined them, there are still practical considerations that we will need to face. For example, we see from Figs. 3 & 4 that haplotype blocks, defined via branches in the genealogy, have a complex structure, tracing back in time for a number of generations that varies along their span (e.g., blocks ii and iii). This makes it (for example) hard to define the extent of haplotype blocks in any simple way, especially since they may be disjunct. Should this be their maximum length, or should it rather be weighted by the depth? It is not clear which description would be better for inference and this may even depend upon the specific process that we wish to infer. These kinds of issues could be investigated by estimating parameters under a variety of specific models in which case we can evaluate the strength and weaknesses of different descriptions of haplotype structure in characterizing different processes.

2.5 | Conclusions and future directions

In this article, we have outlined a definition of the haplotype block, explored the implications of the definition with simple simulations, and considered how current methods can infer such blocks from empirical data. In our view, haplotypes and haplotype blocks should be the core concepts through which we understand population genetic processes. Under this view, it follows that ideally, genomic datasets should come directly as resolved haplotypes, rather than diploid genotypes that require phasing and further processing. We therefore welcome new developments in linked- and long-read sequencing techniques, analysis software, and visualization tools that are designed with sequencing and population datasets in mind (Davies et al., 2016; Meier et al., 2021).

Our simulations and empirical example show that haplotype blocks contain rich information about the demographic and selective history of the locus. Making the most of this information will require a fundamental rethink of our linear, reference-based genome assemblies, and a move towards a graph, or tree-based assembly standard to take advantage of their capability to natively encode variation (Eggertsson et al., 2017; Hickey et al., 2020). We will also need new concepts and vocabulary to describe features in these graphs (e.g., super-graphs and “bubbles”; (Cheng et al., 2021; Turner et al., 2018; Weisenfeld et al., 2017)) informed by a robust understanding of the generative process discussed above, and we need to align our mental models with inference schemes and their encoding (as in, e.g., tsinfer). For that reason, we hope our discussion here can focus our effort towards this new standard, as haplotype-resolved sequencing becomes routine.

2.6 | Supplementary Information

Appendix S1 contains summary of the simulations, and can be found online with the original publication at: https://doi.org/10.1111/mec.16793. Appendix S2 describes the empirical implementation of haplotype blocks with ARGweaver and is included in the thesis in Appendix A.

Box 1. Ancestral Recombination Graph (ARG)

The ARG describes the complete ancestry of a sample of genomes through a series of real coalescence and recombination events (Griffiths and Marjoram, 1997; Hudson, 1983). At any given site on the genome, the relationship can be described through a genealogy (Kingman, 1982); all contemporary samples coalesce and eventually trace back to one single ancestor. Moving along the genome, the relationship inevitably changes due to recombination. This leads to a series of observable genealogies along the genome (Fig B1A), which are embedded in a single structure—the ARG (Fig B1B).

The full ARG (Fig B1B) is a graph structure that depicts individuals (both ancestral and extant), lineage relationships in time. Each node in the ARG represents a real coalescence or recombination event, whilst edges represent the ancestry of a particular genomic segment, along a genetic lineage (depicted by coloured/grey segment for inherited/non-inherited genetic material in Fig B1B). Altogether, an ARG describes the entire ancestral history - each recombination and each coalescence event, which imply the genealogy for each non-recombined genomic block. Crucially, the ARG describes ancestry but not allelic state, so is independent of all the mutations that lead to the observed polymorphism in the present sample.

It is important to note that the full ARG (Fig B1B) contains more information than the series of tree sequences along the genome (Fig B1A). First, a series of tree sequences lack information on the timing of recombination events, unless these are separately stored. Second, while some recombination events lead to observable changes in genealogical trees, others might not. Figure B1A depicts such cases - some recombination events might not change the tree topologies at all (trees *ii* and *iv* are exactly the same), whereas others might only lead to temporal changes in coalescence nodes (tree *i* differs from trees *ii* and *iv* by 1 node position, but all have the same topology). Therefore, while there are 4 non-recombining genomic regions, there are only 2 unique tree topologies (trees *i*, *ii* and *iv* have the same topology) and 3 distinct trees (trees *ii* and *iv* are exactly the same). Some coalescence events can also be entirely invisible and not be represented in any of the individual trees – coalescence at t_2 in Fig B1B is not represented in the series of trees in Fig B1A. Furthermore, two disjunct blocks of the genome can be inherited from the same ancestor, so that a unique coalescence event (e.g., marked by * in Fig B1A) can generate disjunct blocks of ancestry. It should also be noted that although Fig B1 shows the inevitable coalescence of the whole genome into a single common ancestor, this typically takes an astronomically long time: each non-recombining region of the genome coalesces at various time points, and the single lineages ancestral to each region then take an extremely long time to coalesce in one common ancestor, in a process which is in principle unobservable.

Since the ARG contains full information about the genealogy of the sample, it is in theory sufficient to infer any evolutionary process: the ARG necessarily gives more

information than commonly used statistics like SFS, F_{ST}, EHH, which are low-dimensional summaries of the ARG (Ralph et al., 2020). Therefore, the ARG should serve as the foundation for developing new methodologies. However, we note that whilst the ARG is a sufficient statistic, it remains an open question how much the extra information it gives can improve inference: the intrinsic variability of the evolutionary process sets a bound on the accuracy of our inferences.

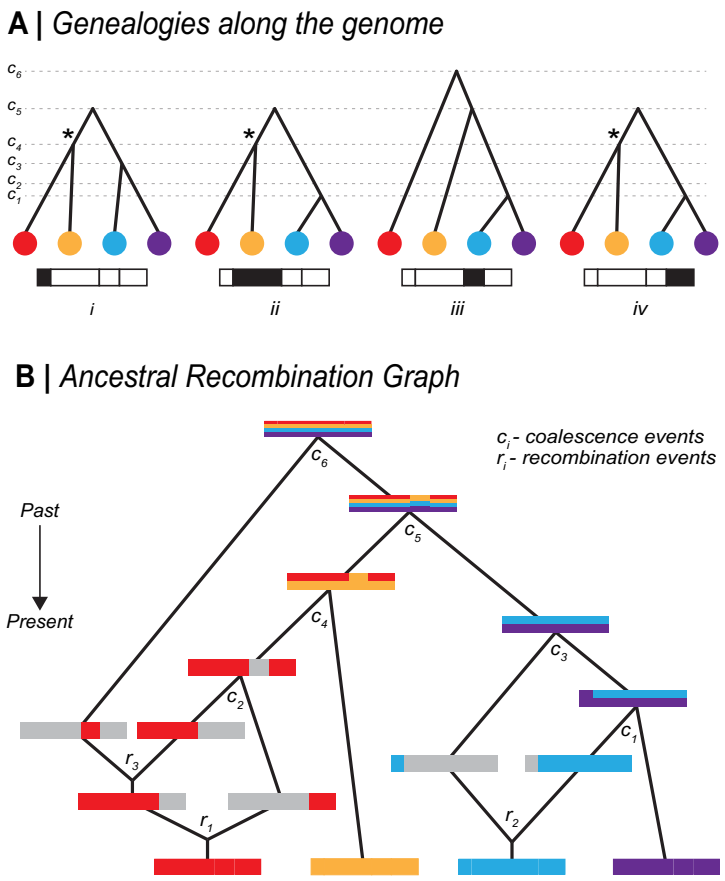

Figure B1. Relationship between Genealogies and the ARG. (A) Genealogical trees along the genome, corresponding to the ARG - each tree describes the ancestral relationship for each of the 4 non-recombined regions. c_1, c_2, ..., c_6 denote time points for each coalescence event. Trees can either change, have the same topology, or marginally differ by only temporal positions of coalescence nodes. Asterisk (*) denotes a unique coalescence event that is ancestral to disjunct genomic regions. (B) Full representation of Ancestral Recombination Graph (ARG) - Tracing back ancestry of four genomes, there is either recombination splitting lineages or coalescence merging lineages. Inherited ancestral genomic regions are coloured corresponding to the contemporary genomes. Recombination is represented by splitting the genome into two; where grey denotes non-ancestral genomic region. Coalescence is represented by two genomes merging, with inherited genomic regions denoted by mixed colours. There are 3 recombination and 6 coalescence events in the full ancestral history of the four genomes. c_1, c_2, ..., c_6 denotes time points for each coalescence event. r_1, r_2, r_3 denotes time points for each recombination event.

Box 2. Population genetic methods that make use of haplotype information

Many methods for inferring evolutionary processes make use of haplotype structure. These can be roughly grouped into three types based on their underlying paradigm: window-based methods, segment-based methods and tree-based methods. These methods vary in complexity from simple heuristics to full statistical treatments. Here we discuss window-based and segment-based methods, but we reserve our discussion of tree-based methods to the main text.

Of the three classes, window-based methods tend to be the simplest, and primarily operate *across* sets of individuals. In the simplest form, haplotypes are operationally defined as the set of alleles observed at the segregating sites within a predefined window of an arbitrary length, say, 50 SNPs or 100 kilobase. Ideally, window sizes should be short enough to minimize spanning recombination breakpoints. One example is H_{12}, which detects selective sweeps (Garud et al., 2015). In this test, for any given window, haplotypes are rank-ordered by their frequencies; in the case of a selective sweep at a given locus, we expect the two most common haplotypes (H_1 and H_2) to dominate the population. The H_{12} test features enhanced power to detect selection, especially under competing sweeps between recurring mutations. However, the test does not attempt to capture the real haplotype block length and is rather heuristic. Other fixed window-based applications include ones exploiting local genomic structures, especially ones showing geographical structure or associated with local adaptation (data-driven clustering/DDC in (Jones et al., 2012), see also (Li and Ralph, 2019; Todesco et al., 2020). While window-based methods do not explicitly infer or use information of haplotype block length, they sometimes do take the genealogical structure into account, e.g., *Twisst* (Lohse et al., 2016; Martin and Van Belleghem, 2017). Often, the simplicity of window-based methods is also their main appeal in the era of SNP genotyping.

Segment-based methods are more sophisticated. They operate primarily on individual sequences, with the aim to represent haplotypes as a mosaic of segments from a haplotype panel, often under some version of Li and Stephens algorithm (Box 2). These segments offer a more realistic model of recombination breakpoints and confer superior power to capture signatures due to linkage. Extended haplotypes homozygosity (EHH) (Sabeti et al., 2002) is an excellent example of such segment-based statistics for inferring selection. Along with its derivatives, such as integrated haplotype score (iHS) (Szpiech and Hernandez, 2014) and cross-population EHH (XP-EHH) (The International HapMap Consortium et al., 2007), they have been widely used to detect selection in many systems (Cao et al., 2011; The International HapMap Consortium, 2007). These methods typically seek to capture the decay of a signal, say, in the extent of haplotype sharing, from an *a priori* defined core SNP. More sophisticated methods based on hidden Markov models to

infer the haplotype structure are especially helpful in uncovering admixture and introgression (e.g., fineSTRUCTURE, Lawson et al., 2012). This allows for the visualization of the haplotype-specific ancestry and improved fine-scale analysis of population structure that is not obvious from unlinked markers.

Box 3. Applications and limits of the Li and Stephens model

Li and Stephens (2003) (LS) proposed a hidden Markov model (HMM) framework that underpins a large number of existing inference methods. Originally developed to model patterns of linkage disequilibrium, it has since been widely applied to develop analytical tools and address empirical problems, such as, phasing and imputation of genomic data (Browning and Browning, 2011; Howie et al., 2011; Li et al., 2010; Marchini et al., 2007; Stephens and Scheet, 2005), inference of population structure and demographic history (Hellenthal et al., 2014; Lawson et al., 2012; Steinrücken et al., 2019, 2018), characterisation of local admixture (Price et al., 2009; Sundquist et al., 2008), inference of local genealogies (Kelleher et al., 2019; Rasmussen et al., 2014; Speidel et al., 2019), and many more. The LS HMM framework is highly tractable and efficient. However, underlying assumptions make it incompatible with the haplotype definition we propose.

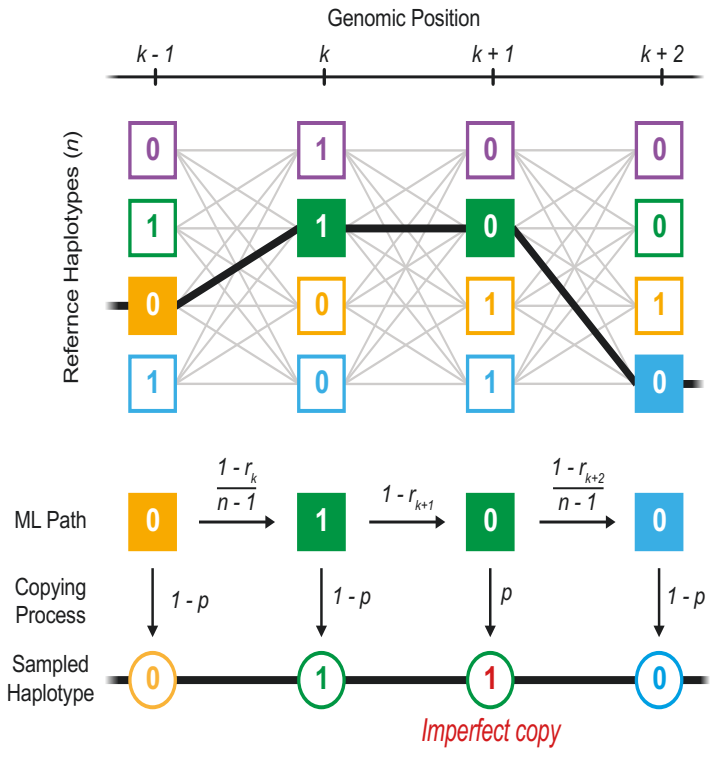

Figure B2. Schematic representation of Li and Stephens hidden Markov model. A new haplotype can be sampled as an imperfect copy of *n* reference haplotypes (hidden states). To find the most likely path taken through the hidden states, the LS model works along the genome (*k-1*, *k*, *k+1*, …), calculating the probabilities of changes in the attributed haplotype. The transition probability to continue or switch the attributed haplotype is a function of the recombination rate *(r)* between adjacent sites, whilst the emission probability to copy the attributed allele with or without error is a function of the mutation rate (*p*). Moving along the genome, the LS model compares the probability of every possible copying path and infers the most likely one.

The LS algorithm requires a reference sample of haplotypes, or if presented in a sequence, previously observed haplotypes. It gives a framework to decide whether some focal haplotype represents a) an entirely new haplotype or b) a mosaic of previously encountered haplotypes, and determines the breakpoints and transitions in this mosaic. Whilst the LS model captures genetic relatedness among chromosomes through recombination, it assumes that the reference haplotypes are known. This would be valid in a selection experiment, if we know the founder genomes; in this case, blocks are defined by IBD to this reference population. However, if we only have contemporary genomes, the reference panel is an approximation. Secondly, the model assumes that genomic states depend solely on the immediately preceding site. This is also an approximation, since in the true ARG, recombinant lineages can coalesce back to any lineage that existed in the preceding genome, which yields disjunct haplotype blocks.

Table 1. A glossary of key terms

Term	Definition
Ancestral recombination Graph	A graphical representation of the complete ancestry of a sample of genomes through a series of coalescence and recombination events. The ARG can be decomposed into a series of marginal trees that give the relationships between samples within each nonrecombining region
Branch	A part of a genealogical tree at a single locus, which connects two coalescence events
Coalescence	The merging of lineages in a common ancestor, as one traces lineages backward in time
Edge	A set of genomic regions that are the immediate ancestors of a specific coalescence event, and that are ancestral to a specific set of sampled genomes. An edge has two dimensions (generations × map length). Any SNP that falls on an edge will be shared by the set of descendant genomes, and only by those genomes
Haplotype	A haploid genotype. A diploid genotype consists of a pair of haplotypes
Haplotype block	The set of genomic regions that descend from a particular edge in the ARG, which is defined by a unique coalescence event, and by the set of descendant samples
Identity by descent (IBD)	Segments of the genome are identical by descent if they descend from the same common ancestor
Lineage	A chain of genes that descends from parent to offspring, or (tracing backwards) from offspring to parent
Linkage disequilibrium	Non-random association of alleles at different loci
Phasing	The process of assigning alleles to the maternal and paternal chromosomes in a diploid individual
Time to the most recent common ancestor (TMRCA)	The time of the most recent coalescence event from which a focal set of samples descends

Chapter 3
The genetic basis of a recent transition to live-bearing in marine snails†

Abstract

Key innovations are fundamental to biological diversification, but their genetic basis is poorly understood. A recent transition from egg-laying to live-bearing in marine snails (*Littorina*) provides the opportunity to study the genetic architecture of an innovation that has evolved repeatedly across animals. Individuals do not cluster by reproductive mode in a genome-wide phylogeny, but local genealogical analysis revealed numerous small genomic regions where all live-bearers carry the same core haplotype. Candidate regions show evidence for live-bearer-specific positive selection and are enriched for genes that are differentially expressed between egg-laying and live-bearing reproductive systems. Ages of selective sweeps suggest live-bearer-specific alleles accumulated over ~200k generations. Our results suggest that novel functions evolve through the recruitment of many alleles, rather than in a single evolutionary step.

† This chapter is published and can be found online at: https://doi.org/10.1126/science.adi2982

3.1 | Introduction

Evolution is a gradual process, but occasionally results in sudden changes in form and function that allow organisms to exploit new ecological opportunities (Miller et al., 2023; Wagner, 2011). These game-changing traits—including flight, vision, and the bearing of live offspring—are known as 'key innovations' (Baum and Larson, 1991; De Queiroz, 2002; Miller et al., 2023). Key innovations are all around us, and have catalyzed the diversification of many groups (Wagner, 2011). Despite their significance, we know surprisingly little about the origins and genetic basis of innovations (Wagner, 2011). This is because most of them originated deep in the past, making it difficult to disentangle causal loci from the countless other genetic changes that accumulated up to the present.

A recent transition in female reproductive mode offers a rare opportunity to study the genetic basis of an innovation that has evolved many times across the animal kingdom (Whittington et al., 2022). We focus on a clade of intertidal gastropods (genus *Littorina*) where the ancestral state is to lay a large egg-mass, but one species gives birth to live young (Fig. 1A, fig. S1) (Reid, 1996; Reid et al., 2012). Egg-layers have a gland that embeds fertilized eggs into a protective jelly. In the live-bearer*, L. saxatilis*, this structure has evolved into a brood pouch where embryos develop inside the mother. Live-bearing is the only taxonomic character that is diagnostic of *L. saxatilis*, as no other known trait differs consistently between the live-bearing and egg-laying individuals (Reid, 1996; Reid et al., 2012). In fact, sympatric populations are so similar that the difference in mode was long thought to reflect within-species polymorphism (Reid, 1996; Seshappa, 1947) and molecular markers are needed to identify males and juveniles where live-bearing and egg-laying species coexist (Stankowski et al., 2020).

Live-bearing is thought to be an adaptation allowing snails to reproduce in areas where eggs would be exposed to harsh conditions (Reid, 1996). This is reflected in the much broader ecological and geographic distribution of *L. saxatilis* compared with the two closely-related egg-laying species, *L. arcana* and *L. compressa* (Reid, 1996) (Fig. 1B and 1C, fig. S2). Egg-laying and live-bearing species have adapted in parallel to contrasting environments across the intertidal zone (Johannesson, 2003; Reid, 1996), largely decoupling reproductive mode from other axes of phenotypic divergence (Fig. 1B). There is also evidence for rare gene flow between egg-layers and live-bearers (Stankowski et al., 2020). These features provide an opportunity to identify and study the genetic changes underlying the live-bearing innovation.

3.2 | Results and Discussion

3.2.1 | Live-bearing snails do not form a monophyletic group

We used whole-genome sequences from 108 individuals to test the existing hypothesis of a single origin of live-bearing, inferred by parsimony analysis in earlier phylogenetic studies (Fig. 1D, figs. S3 & S4, tables S1 & S2) (Reid, 1996; Reid et al., 2012). Rather than forming a single clade, live-bearers formed two separate, well supported clades

in a genome-wide phylogenetic tree (Fig. 1E): one containing all *L. saxatilis* from Iberia (hereafter 'Iberian *saxatilis*'), and another including all other *L. saxatilis* ('Northern *saxatilis*') that was sister to egg-laying *L. arcana* (same pattern observed in an ML tree, phylogenetic network and PCAs; figs. S5 & S6).

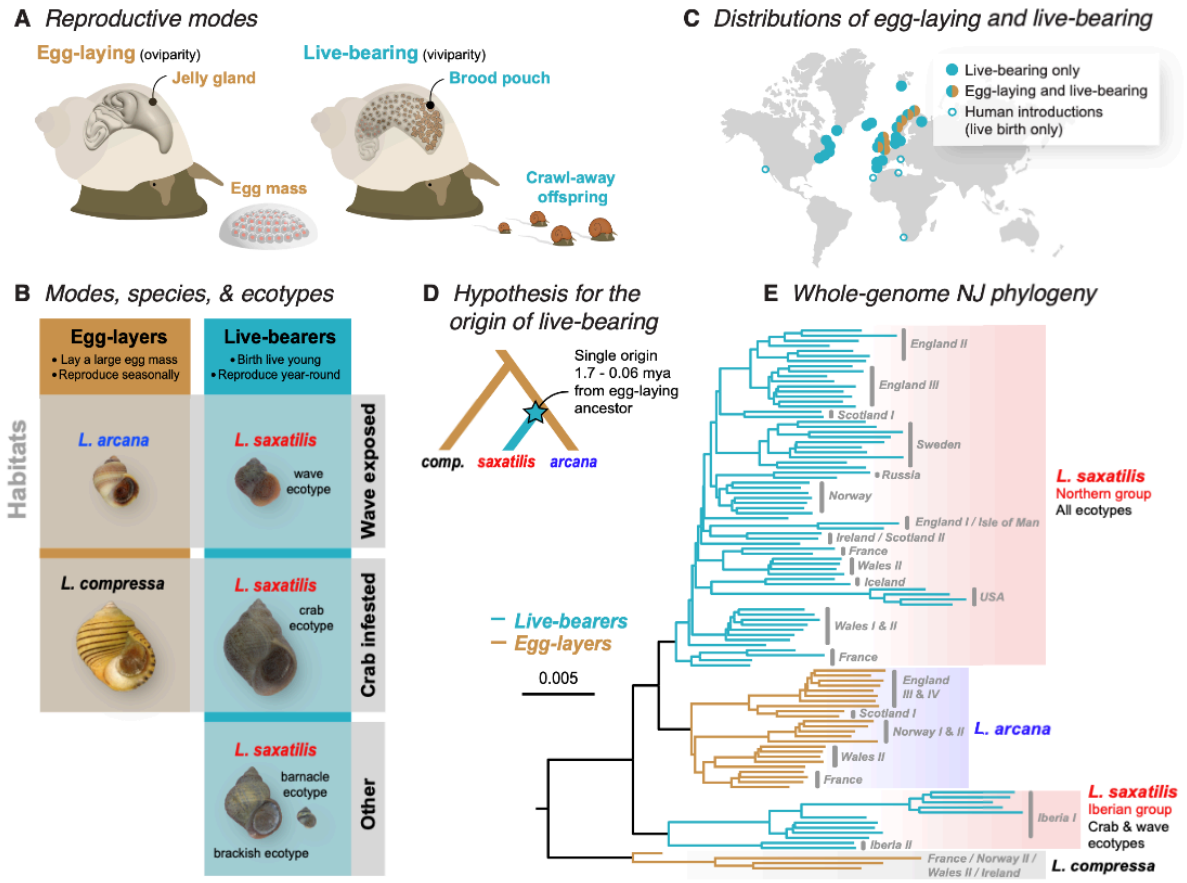

Figure 1. Variation in reproductive mode in *Littorina*. (A) Anatomical differences between modes. (B) Egg-layers reproduce during a limited breeding season, while live-bearers release offspring year-round. The two egg-layers share their habitats with ecotypes of the live-bearer, *L. saxatilis*. (C) Approximate distributions of the modes, highlighting the broader distribution of live-bearing. (D) Existing hypothesis for the origin of live-bearing inferred by past phylogenetic studies (Reid, 1996). (E) Neighbour-joining phylogenetic tree based on whole-genome sequences (108 individuals and 18.5 million variable sites). All nodes have 100% bootstrap support.

The discordance between evolutionary relationships and reproductive mode has several possible explanations. One interpretation of the genome-wide tree is that there has been more than one transition between egg-laying and live-bearing. However, because the tree represents the aggregate signal from all loci in the genome, it does not necessarily reflect the evolutionary history of these groups, or of any single locus or trait (Hahn and Nakhleh, 2016). This means it is also possible that live-bearing may have evolved once, and that causal alleles became associated with two different lineages via the interaction of gene flow and selection (Hahn and Nakhleh, 2016). If this were the case, we would expect genealogies for

loci that cause live-bearing to be strongly discordant from the genome-wide tree, with samples grouping by reproductive mode.

3.2.2 | Topology weighting reveals rampant genealogical discordance and loci associated with reproductive mode

With this expectation in mind, we used topology weighting (Fig. 2A) to identify genomic regions associated with reproductive mode. For each genomic window, topology weighting calculates the degree of monophyly toward three possible taxon subtrees (Fig. 2B & 2C, fig. S7): (*i*) the background topology, Tb, observed in our genome-wide analysis, (ii) the reproduction topology, Tr, where samples cluster by reproductive mode, and (iii) the control topology, Tc, which is of no specific interest but provides a control for distinguishing incomplete lineage sorting from other processes that cause genealogical discordance (e.g., gene flow). We used non-overlapping 100-SNP windows (mean size 5.8 kb, fig. S8), and calculated topology weights (Martin and Van Belleghem, 2017) for each window by sampling 10,000 subtrees (Fig. 2A).

We analysed the joint distribution of topology weights in a ternary framework. This approach exploits the geometric properties of the ternary plot, allowing us to visualize and quantify various properties of the genome-wide distribution of discordance (Fig. 2A). We used simulations to illustrate how different factors, including the timing of population splits and gene flow between non-sister lineages, shape the ternary distribution of topology weights for large numbers of loci (Fig. 2B; Supplementary text, figs. S9—S19; tables S3 & S4).

We expected the empirical distribution of weights to be biased toward Tb, because this was the topology observed in the genome-wide analysis. However, the observed bias was only slight (Tb = 0.380, Tc = 0.310, Tr = 0.308), with just 62 of ~155,000 genomic regions perfectly fitting Tb (*i.e.*, Tb = 1) (Fig. 2C). Instead, the bulk of the distribution fell close to the center of the triangle. This indicates that sequence variation is broadly shared between groups, resulting from extensive incomplete lineage sorting due to rapid diversification relative to the effective population size, or widespread gene flow during divergence, or both (Hudson, 1990; Maddison, 1997). Thus, although well-supported statistically, the genome-wide tree is a very poor predictor of evolutionary relationships at any given genomic region.

We found substantial left-right asymmetry in the distribution of topology weights (Fig. 2D). Such a bias is not expected to arise from incomplete lineage sorting, because there is an equal chance that a given gene tree will more-closely resemble either alternative topology (Fig. 2B, Supplementary Materials) (Maddison, 1997). We detected asymmetry using a new statistic, D_{LR} (Fig. 2D, fig. S19). A genome-wide test, performed by calculating D_{LR} between the two halves of the triangle, revealed a 3.4% excess of windows shifted toward the control topology (D_{LR} = 0.034, permutation test p = 1e-5). D_{LR} calculated between analogous left- and right-side sub-triangles, revealed that this asymmetry was driven by an excess of trees with a small bias toward Tc (Fig. 2D, table S5). Further exploration showed that this bias is due to 10 previously-identified chromosomal inversions (Reeve et al., 2024), none of which is associated

with reproductive mode (figs. S20—S23, table S6, Supplementary Materials). For each inversion, one arrangement is more common in Spanish *L. saxatilis* and *L. arcana*, and the other is more common in *L. compressa* and Northern *L. saxatilis.* When the chromosomal inversions are removed, we find no significant left-right asymmetry at the genome-wide level (D_{LR} for regions outside inversions = -0.007, p = 0.074, fig. S20).

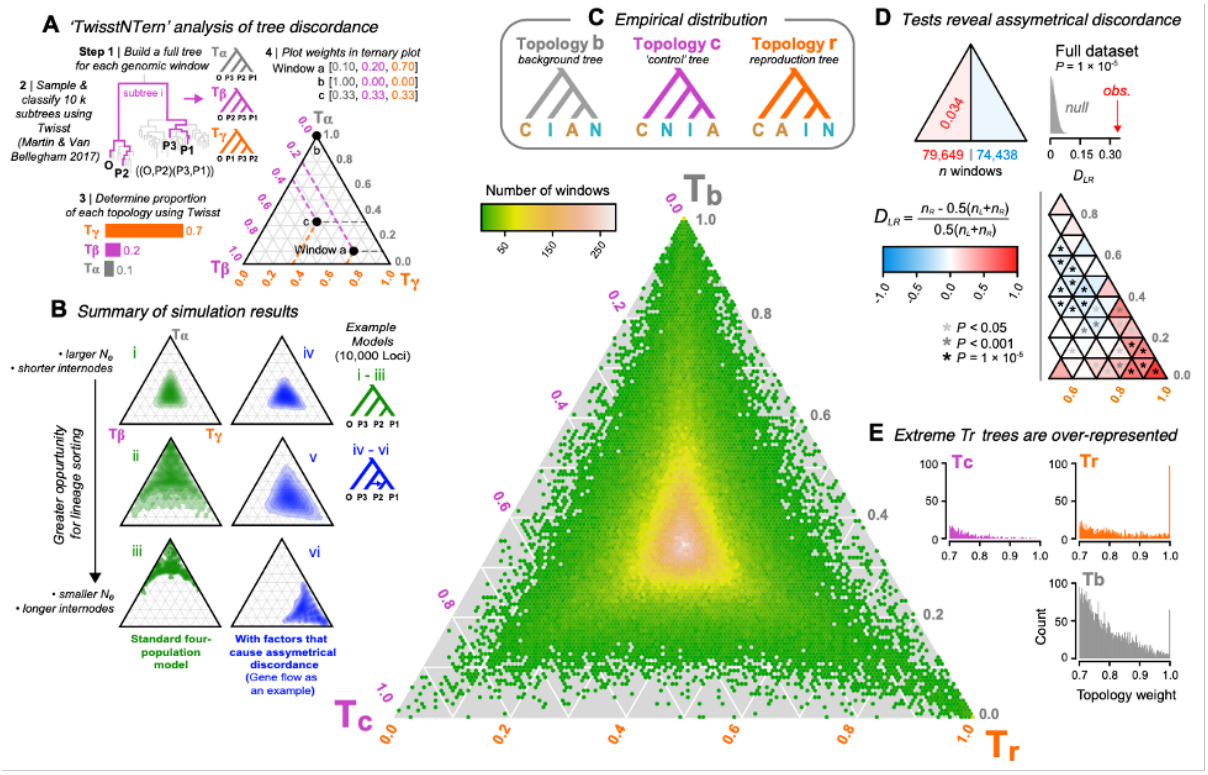

Figure 2. Topology weighting reveals genomic regions associated with reproductive mode. (A) For each genomic window, we inferred a tree for all haplotypes, and then classified 10k 'subtrees' by randomly picking one haplotype per group. Topology weights are the proportions of each topology among all subtrees. Windows were plotted in a ternary plot based on the weights. (B) Simulated distributions of weights. A greater opportunity for lineage sorting (i - iii) biases the distribution toward the topology that matches the demographic history. Incomplete lineage sorting yields genealogies that are a better fit to one of the discordant trees, but the distribution is always symmetrical between the left and right half triangles. Additional factors, including gene flow, create a bias toward one discordant genealogy (panels iv - vi). (C) Possible topologies and the empirical distribution of weights for the 154,971, 100-SNP windows; C: *compressa*, A: *arcana*, I: Iberian *saxatilis*, N: Northern *saxatilis*. Hexagonal bins are coloured by window count. (D) Counts of windows in the left and right half-triangles, with asymmetry quantified using D_{LR}. Further division into sub-triangles reveals left-right asymmetry throughout the distribution. Asterisks indicate significant asymmetry between corresponding left- and right-sided sub-triangles. (E) Distributions of weights > 0.7.

Much stronger asymmetry was observed between the far left and right sub-triangles, corresponding to windows that more strongly fit one of the alternative topologies (Fig. 2D, fig. S24). However, the asymmetry was in the opposite direction to the genome-wide pattern, with a large excess of windows strongly biased toward the reproduction tree compared with the control tree (Tr > 0.7 = 1151 windows vs. 461 for Tc; D_{LR} = -0.43, p = 1e-5). A total of 88

windows perfectly fit the reproduction topology (*i.e.*, Tr = 1, table. S7), compared with no windows that perfectly fit the control topology (D_{LR} = 1.00, p = 1e-5; Fig. 2E, fig. S25).

3.2.3 | Evidence for live-bearer specific positive selection

Although neutral gene flow can generate strong asymmetry under some circumstances, we are unable to explain the observed Tr bias without invoking natural selection (Supplementary Materials, table S8). We found strong additional evidence for live-bearer-specific positive selection in regions associated with reproductive mode. First, window-based estimates of nucleotide diversity (π) in live-bearers decreased substantially with increasing Tr weight (Fig. 3A), but we found no such relationship in egg-layers. Eighty-four (95%) of the 88 perfectly associated regions showed reduced π in live-bearers (mean $\pi_{live\text{-}bearer}$ = 0.0029 vs $\pi_{egg\text{-}layer}$ = 0.0065; paired Wilcoxon test, p = 1.313e-15, Fig. 3A, fig. S26). These results are consistent with selection having purged diversity from haplotypes associated with live-bearing (Maynard Smith and Haigh, 1974). Although this result could in principle result from a live-bearer-specific demographic bottleneck, we can rule this out because live-bearers and egg-layers have similar levels of genome-wide diversity (mean $\pi_{live\text{-}bearer}$ = 0.0065 vs. $\pi_{egg\text{-}layer}$ = 0.0062; Fig. 3A, fig. S27). Further, relationships between π and the other weights (Tb and Tc) were weak, and similar for both groups, confirming that reduced π in live-bearers is specific to Tr rather than being a general feature of windows with extreme weights (fig. S28). The site-frequency spectra (SFS) and sample-size-corrected estimates of private alleles for perfectly associated regions provide further evidence for selection (Fig. 3B—3D; figs. S29—S31; tables S9—S11): the live-bearer SFS was strongly skewed toward rare variants (Tajima's D = -1.89, 95% CIs -1.77 – -2.01; fig. S29), the majority of which (80%) were private to live-bearers. Both results are expected during the phase when diversity is recovered by mutation after a selective sweep (Braverman et al., 1995). In contrast, the SFS for egg-layers was much closer to the neutral expectation (Tajima's D = -0.24, 95% CIs -0.037 – -0.437), with polymorphic sites being 2.14 times more abundant in egg-layers after accounting for the difference in sample size.

We next characterized footprints of selection within contigs to estimate the number and size of candidate regions more accurately (Fig. 3E). The 88 perfectly associated windows mapped to 50 contigs in our genome assembly (mean 1.7 ± sd 1.5 windows per contig; table S9). Associated regions were narrow, mostly spanning less than 20 kb (mean 12 kb ± sd 14.4). Sliding-window analysis of each contig generally revealed clear peaks of allele frequency differentiation (F_{ST}) and sequence divergence (d_{xy}) between the egg-layers and live-bearers, as well as valleys of nucleotide diversity (π) in live-bearers (Fig. 3E). We also inferred ancestral recombination graphs (ARGs) for selected contigs to refine candidate regions (Fig. 3E). Unlike trees inferred from genomic windows with arbitrary start and end positions, each tree in an ARG corresponds to an inferred non-recombining segment of the genome (Shipilina et al., 2023). Thus, by applying topology weighting to the sequence of marginal trees, we were able to identify more precisely the segment of genome retained by all live-bearing samples

following the selective sweep. In both cases, the core live-bearing haplotype spanned less than 2 kb. Live-bearers showed much shallower coalescence in these regions than egg-layers, as expected following a sweep (Fig. 3E).

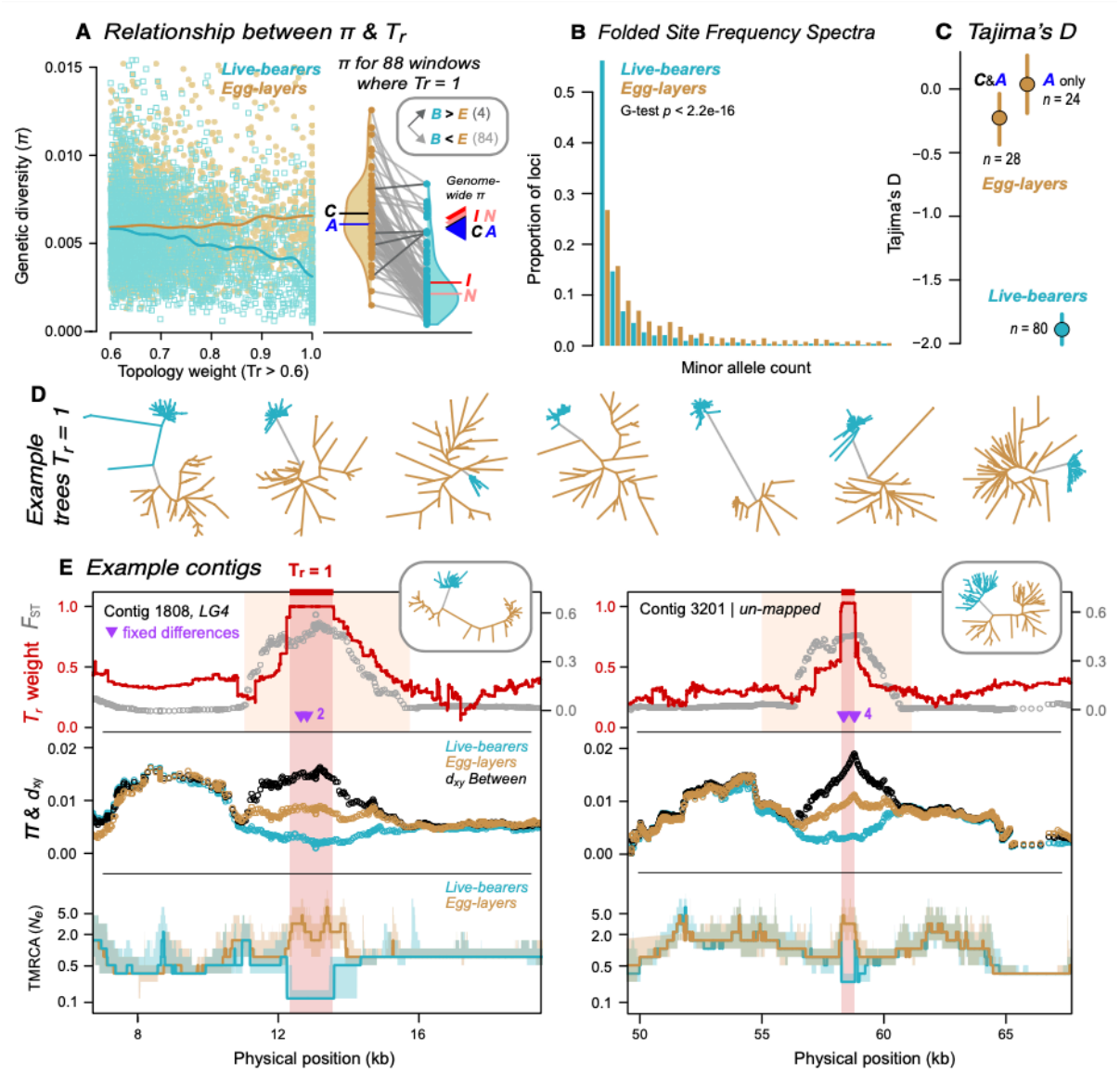

Figure 3. Evidence for positive selection on haplotypes associated with live birth. (A) Relationship between π and Tr for both reproductive modes. Triangles on right: genome-wide π. (C: *compressa*, A: *arcana*, I: Iberian *saxatilis*, N: Northern *saxatilis*). Violin plots: distributions of π for windows where Tr = 1. (B) Folded SFS for each mode in perfectly associated regions, projected at the same sample size for comparison. (C) Estimates of Tajima's D with 95% CIs for perfectly associated regions. (D) Examples of trees for windows where Tr = 1. (E) Variation across two example contigs that contain a window where Tr = 1 (span of the orange box). The tree associated with each region is shown. Top panel: F_{ST} between egg-layers and live-bearers in 3kb sliding windows (30 bp step). Tr shows the results of topology weighting applied to marginal trees obtained from inferred ancestral recombination graphs (ARGs). Purple arrows show fixed differences between modes. Middle panel: π and d_{xy} in sliding windows. Bottom panel: traces of time to the most recent common ancestor (TMRCA) obtained from ARGs. Bold lines: median estimates; Envelopes: 95% CIs. The red box shows the inferred length of the core haplotype block associated with live birth.

Approximate estimates of the timing of each selective sweep at the mode-associated loci, estimated from the accumulation of private mutations ($T = \pi_{private}/2\mu$), span a broad range from ~20 k to 200 k generations before present, with a median of 70 k generations before present (fig. S32). Assuming 2 generations per year, this equates to 100 k to 10 k years before present, with a median time of 35 k years.

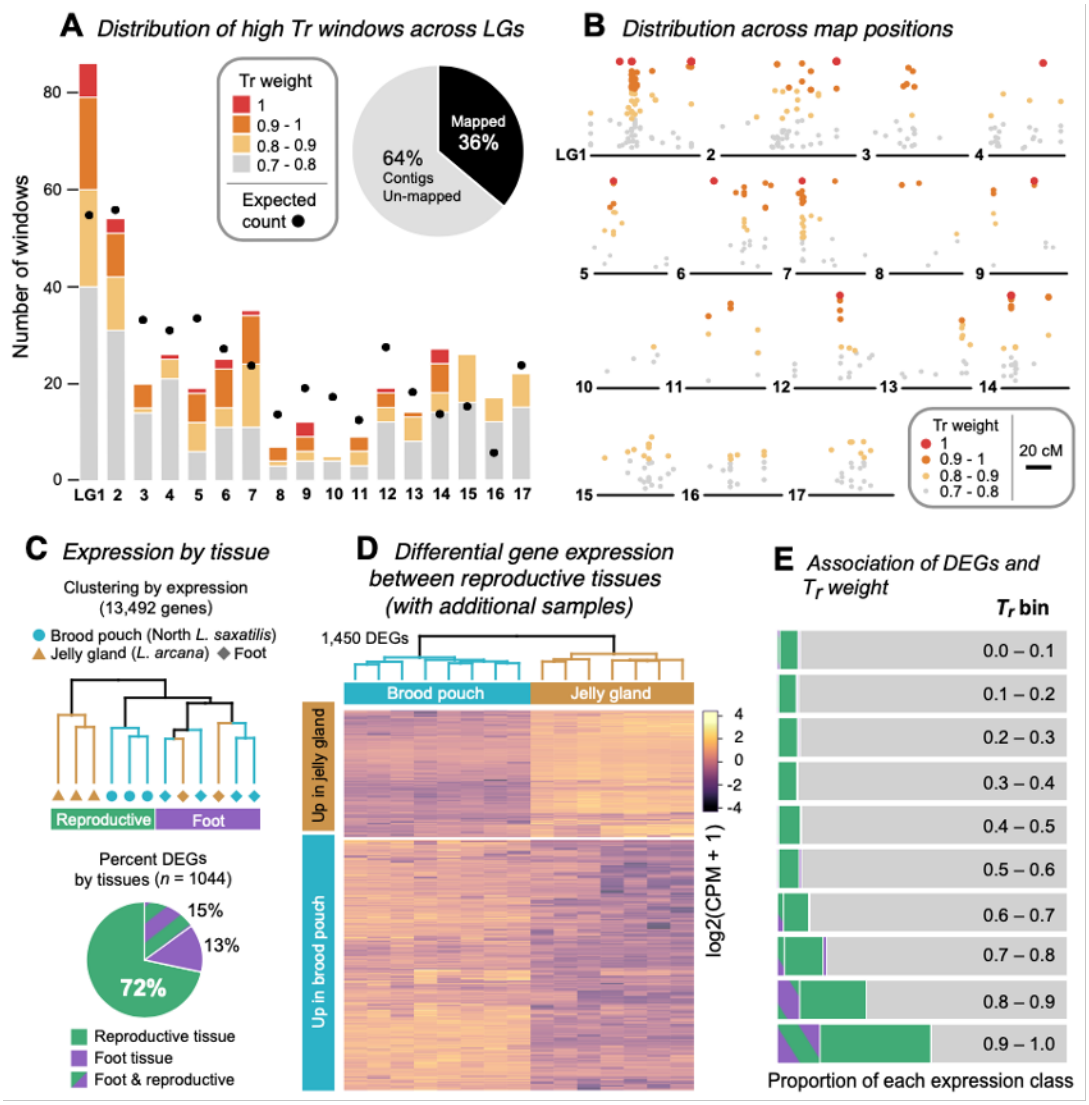

Figure 4. Candidate regions are widespread across the genome and enriched for genes that are differentially expressed (DEGs) between reproductive systems. (A) The number of high Tr windows (Tr > 0.7) assigned to each of the 17 *L. saxatilis* linkage groups (LGs). Dots show the expected number of windows given the total assigned to each LG. (B) Distribution of high Tr windows across LGs. (C) Clustering of tissues by expression and the number of differentially expressed genes (DEGs) in each expression class. (D) Clustering of reproductive tissues based on patterns of expression. (E) The proportion of genes in each differential expression class after binning genes according to the Tr weight.

3.2.4 | Mode-associated regions are widespread and enriched for genes that are differentially expressed between reproductive systems

The assignment of contigs to a genetic map revealed that reproductive-mode-associated windows are widespread across the genome, rather than co-localizing to one or a few genomic regions (Fig. 4A). As expected for a polygenic trait, the number of mode-associated windows on each linkage group (LG) was strongly predicted by LG size (Tr > 0.7, r = 0.79, p < 0.0001; Tr > 0.9, r = 0.71, p < 0.005). Associated regions were also widespread within linkage groups, in some cases with strong associations near opposite ends of the same LG (Fig. 4B).

Candidate regions also showed strong enrichment of genes that are differentially expressed between female live-bearing and egg-laying reproductive tissues. To identify differentially expressed genes (DEGs), we collected reproductively mature female *L. arcana* and Northern *L. saxatilis* at peak breeding season from a single location (to control for environmental effects) where sympatric egg-layers and live-bearers are morphologically cryptic aside from their reproductive anatomies. We first compared transcriptomes from pools of reproductive systems (brood pouch vs. jelly gland) and foot tissue paired from the same individuals. Clustering analysis based on patterns of gene expression (13,492 genes) revealed that pools of reproductive tissue grouped by system type, but egg-laying and live-bearing species did not group based on expression in foot tissue. Differential expression analysis revealed 1,044 DEGs, and showed much higher rates of tissue-specific differential expression between reproductive systems (Fig. 4C, fig. S33). To increase power to detect DEGs between the reproductive systems, we sequenced additional pools of reproductive tissue (Fig. 4D, fig. S34). This analysis detected 1450 DEGs, 66.1% (858) of which showed higher expression in the brood pouch of live-bearers. To test for the enrichment of DEGs in regions associated with reproductive mode, we binned each DEG according to the Tr score of its associated genomic region (Fig. 4E, fig. S35, table S12). We found that the proportion of reproductive mode DEGs strongly increased with increasing Tr weight (Spearman's rho = 0.903, p = 9e-04) (table S13). No correlation was observed between Tr weight and foot-tissue only DEGs (Spearman's rho = -0.410, p = 0.217) (table S13).

Gene ontology (GO) analysis and functional annotation suggest that the transition to live-birth involved genes with diverse functions. Separate GO analyses conducted on a sequence-based gene set (574 genes in regions where Tr > 0.7) and expression-based gene set (1,450 reproductive mode DEGs) yielded 37 enriched GO terms, including transmembrane transport, calcium-ion binding, and ion channel activity (Fig. S36). We examined the putative functions of the 27 genes found in both sets in more detail (table S14). These included genes putatively associated with antibacterial activity (lectin L6-like protein; higher expression in brood pouch), the synthesis of mucin-type oligosaccharides (GALNT10-like; higher expression in brood pouch), the formation of structural tissue (IFB-like and CMP-like, both higher expression in brood pouch), and two secretory genes involved in egg-mass production in another marine snail (both with lower expression in brood pouch).

3.3 | Conclusions

Our analyses show that live-bearing in *Littorina* is associated with selection on many loci, as in the only comparable analysis in *Zootoca* lizards (Recknagel et al., 2021). Although our genome-wide analysis suggested two independent origins of live-bearing, the high sequence similarity of live-bearer-specific alleles indicates that they had a single origin. Given the number of associated loci, the history of their origin and spread may be highly complex, potentially varying among loci. One possible scenario is that all live-bearer-specific alleles originated in one location (e.g., in the ancestor to the Iberian clade), after which the range of live-bearers expanded until they encountered egg-layers. Hybridization may have then allowed beneficial live-bearing alleles to introgress onto the local egg-laying genetic background, or may have eroded genome-wide differentiation between egg-layers and live-bearers while selection maintained alternative sets of alleles at mode-associated loci. It is also possible that alleles arose in numerous locations, and that associations built up between them at different times, perhaps as live-bearing spread. Regardless of the precise history, which we cannot resolve at present, the interaction between gene flow and selection has allowed us to identify loci associated with reproductive mode.

Live-bearing is the only known trait that consistently distinguishes *L. saxatilis* from the egg-laying species, making associated loci good candidates for causing the difference in reproductive mode. We found supporting evidence in our expression analysis, as associated regions are strongly enriched for differentially expressed genes between the reproductive systems. This suggests that selection has acted on differences in gene expression, driving the evolution of live-bearing, including development of the brood pouch. Because reproductive mode is a complex trait, associated loci may also underpin a diverse range of biological functions, including the difference in the synchronization of egg-production (Reid, 1996) differences in embryo retention time (Shine, 1983), and variation in immune function and metabolism (Shine, 1983). It is, however, important to emphasize that some loci may not be causally associated with reproductive mode, and may instead underlie other less conspicuous traits that are functionally linked or associated with the live-bearing or egg-laying lifestyle.

Polymorphic inversions often underpin local adaptions (Jones et al., 2012; Lowry and Willis, 2010), and are thought to maintain beneficial sets of alleles by suppressing recombination (Wellenreuther and Bernatchez, 2018). This makes a role in the evolution of a key innovation seem likely but we found that known chromosomal inversions in *Littorina* are not associated with the difference in reproductive mode. Many large inversions are shared among these *Littorina* species (Reeve et al., 2024), with evidence that they play a key role in ecotype formation and reproductive isolation (Faria et al., 2019; Koch et al., 2021; Morales et al., 2019). For example, in *L. saxatilis*, repeated adaptation to contrasting crab-infested and wave-swept environments usually involves 8 to 12 major inversions, where the alternative arrangements contribute to differences in morphology and behavior (Koch et al., 2021; Morales et al., 2019). However, egg-laying species coexist with *L. saxatilis*, and likely use the same inversions to adapt to local selection pressures (Reeve et al., 2024). Thus, the

independence of genetic architectures for ecotype formation and reproductive mode may be a major factor permitting the local coexistence of egg-laying and live-bearing species.

Our estimates of the timing or sweeps suggest that alleles associated with live-bearing were recruited gradually over the last 200k generations (~100k years). This finding is relevant to long-standing debate about the genetic basis of evolutionary novelty. Because key innovations are not visible to selection before they arise, models of saltational evolution invoke large-effect macromutations to explain their evolution (Theißen, 2009). We do not know which mutation caused the threshold from egg-laying to live-bearing to be crossed: some potentiating mutations may have preceded live-bearing but been critical to its origin, and others may have refined live-bearing after it arose. Nevertheless, our results suggest that novel functions evolve gradually through the recruitment of alleles at many loci, rather than arising in a single evolutionary step (Blount et al., 2012; Meyer et al., 2012; Recknagel et al., 2021).

3.4 | Methods Summary

Detailed methods are available at: https://doi.org/10.1126/science.adi2982
Along with the core theme of this thesis, methodological description of *TwisstNTern* and genealogical analysis (contributed by AP) is detailed here.

3.4.1 | Local tree building and topology weighting

We used local tree building and topology weighting to characterize patterns of genealogical variation across the genome, following a custom pipeline established by Martin and Van Belleghem. The following pipeline was run twice independently from the phasing all the way through to the topology weighting, and produced highly similar results.

We first phased our genotype data using Beagle 5.3 (Browning and Browning, 2013), with the settings recommended in the manual for a large outbred population. Each assembly contig was then divided into non-overlapping genomic windows that each contained 100 SNPs. This yielded 154,971 genomic windows, with a mean physical length of 5.8 kb ± s.d. 5.3 kb (fig. S8). A neighbor joining tree was then constructed for each genomic region using the program *Phyml* (Guindon et al., 2010). We assumed a GTR model and inferred trees using the settings suggested by Martin and Van Belleghem (2017). Selected trees were illustrated using the *R* package *ggtree* (Yu et al., 2017) using the unrooted layout.

Topology weighting was then performed on each of the 154,971 trees using the program *Twisst* (Martin and Van Belleghem, 2017).. Topology weights are values that describe the degree of monophyly in a large genealogy in light of the possible monophyletic taxon-level relationships that could be observed. Consider a large genealogy for a single non-recombining locus with 100 sampled individuals, but only four taxa: 1, 2, 3, 4 (Fig. 2A, fig. S7). Although there is an inordinate number of possible unrooted subtrees that one can observe given the large number of tips, there are only 3 possible unrooted topologies that can be observed if we

randomly sample a subtree from the full tree that includes only one tip of each taxon: ((1,2)(3,4)), ((1,3)(2,4)), ((1,4)(2,3)). If the topology of many random four-taxon subtrees (say 10,000) is determined by iterative sampling of the full tree, then the fraction of each subtree gives a quantitative description of the degree of monophylly toward each of the three possible taxon subtrees.

Topology weighting was performed using the Newick tree files as the input, with the following taxon-level topologies specified for calculating the weights (Fig. 2C). Given that we have four taxa in our genome-wide phylogeny (*L. compressa* = C; Iberian *L. saxatilis* = S, *L. arcana* = A, and Northern *L. saxatilis* = N), the three possible unrooted taxon topologies are: Tb = ((C,I)(A,N)), Tr = ((C,A)(I,N)) and Tc = ((C,N)(I,A)) (Fig. 2A). Tb—the background topology—is consistent with the taxon-level relationships that we observed in our genome-wide tree (but note that the full topology will not necessarily be exactly the same as the background tree because relationships may differ within each of the four groups). whereas Tr and Tc represent possible alternative relationships. In Tr—the reproduction tree—the taxa group according to their reproductive mode, whereas in Tc—the 'control' tree—the taxa do not cluster as expected based on their geographic relationships or any known aspect of their biology. We refer to Tc as the control topology, *sensu* Martin et al. (Martin et al., 2013), because it provides a useful control for distinguishing incomplete lineage sorting from other processes that can generate discordant gene trees.

Because each of the 154,971 trees is large, consisting of 2*n* = 216 tips, we calculated the weights based on a fixed number of subtrees rather than performing an exhaustive search of all possible quartets. We sampled 10,000 subtrees, which has been shown to give very narrow limits on the true topology weights (Martin and Van Belleghem, 2017).

3.4.2 | Plotting of topology weights in a ternary plot

Rather than plotting the topology weights across the genome (Martin and Van Belleghem, 2017), we took the novel approach of analyzing the joint distribution of topology weights within a ternary plot (Fig. 2A). The ternary plot has been used to study concordance in phylogenetic studies (Allman et al., 2022), and is also a natural framework for analyzing the distribution of weights in a tree with four taxa, as it makes possible a graphical representation of each genomic window as a single point in an equilateral triangle based on the three topology weights.

The three corners of the ternary plot—[1,0,0], [0,1,0], [0,0,1]—correspond with genomic windows that show taxon-level relationships that are consistent with one of the three possible subtrees; that is, 100% of the sampled subtrees (in our case 10,000) perfectly match one of the three alternative trees, implying that samples from each of the four groups are monophyletic (Fig. 2A). In contrast, the very center of the ternary plot—[0.33,0.33,0,33]—corresponds with a genomic window where all three of the possible subtrees were sampled at equal frequency. Any other location in the ternary plot indicates a bias toward one of the subtrees, but with some resemblance to at least one of the other alternative topologies.

3.4.3 | Exploring the ternary framework with simulation

To understand how different evolutionary processes shape the distribution of topology weights in the ternary framework, we simulated genealogies for non-recombining haplotypes in a four-taxon framework using MSprime 1.2.0 (Baumdicker et al., 2022). We specified a demography where four descendent populations—O, P3, P2, P1—are produced by population splits at three points going backwards in time (table S3): T1, which splits population P12 to give rise to descendent populations P1 and P2; T2, which gives rise to the populations P12 and P3; and T3, where the common ancestor of all the descendent population splits, giving rise to lineages O and P123. The size of each population (*X*) is set to N_eX haploid sequences, and uni- or bi-directional migration can occur between a single pair of ingroup taxa (*i.e.*, between P2 and P3). Simulated genealogies were visualized in demesdraw 0.3.2 (Gower et al., 2022).

We varied these 12 parameters to produce 101 unique models that fall into the 9 scenarios outlined below (*a* – *h*, fig. S9; table S4). For each model, we simulated 10,000 coalescent trees, which were then output in Newick format using tskit 0.5.4 (Kelleher et al., 2018), and passed to the *Twisst* algorithm for topology weighting as described above. The topology weights were plotted in a ternary plot using the R library *Ternary*. The results of the simulations are shown in figs. S10-S18, and described in detail in the supplementary text.

a. *Uniform N_e and even time between splits:* We conducted 12 simulations where we varied the number of generations between the three splits, while keeping populations identical in size. These simulations range from very short durations between splits, mimicking a scenario with almost no demographic separation (T1 = 1, T2 = 2, T3 = 3), through to very long durations between splits relative to the N_e (T_1 = 7290, T_2 = 14580, T_3 = 21870). N_e was set to 500 for all populations and no migration was allowed between them

b. *Varying but equal time between T2 and T3, with T1 set to 5k generations:* We conducted 8 simulations where we varied the durations between splits T2 and T3, but fixed T1 at 5000 generations to allow populations to diverge after all 3 splits had occurred. N_e was set to 500 for all populations across all simulations. No migration was allowed between populations.

c. *Uneven time between splits:* We conducted 10 simulations where the number of generations between population splits varied in evenness. This was done by fixing the values of T1 and T3, and varying the time of T2. We used 2 combinations of T1 and T3 (90 and 270 and 405 and 1215) to see how the unevenness of splits interacted with variation in the duration of divergence. N_e was set to 500 for all populations across all simulations with no migration between populations.

d. *Uniform variation in N_e:* We performed 10 simulations using a model with equal time between splits times but with different values of N_e. The split times used for all populations were T_1 = 90, T_2 = 180, T_3 = 270. We varied the N_e from 5 to 5000, keeping it uniform for all populations. No migration was allowed between populations.

e. *Varying N_e in one population:* We performed 16 simulations using a model with equal time between splits, but with the N_e of one population (always P3) set to a value that

was larger or smaller than all other populations where N_e was always 500. This was done for 4 different sets of split times, ranging from short durations between splits (T_1 = 10, T_2 = 20, T_3 = 30), to very long durations (T_1 = 7290, T_2 = 14580, T_3 = 21870). Four different values of N_e were tested for each scenario. No migration was allowed between populations.

f. Unidirectional migration between P2 and P3: We conducted 16 simulations with varying rates of migration between populations P2 and P3. We defined four histories with the same N_e for all populations (N_e = 500) and equal time between splits, but with the duration between splits varying from very short (T_1 = 10, T_2 = 20 T_3 = 30) to very long (T_1 = 7290, T_2 = 14580, T_3 = 21870). For each of the four histories, we modelled four different rates of migration: m = 0, 0.001, 0.010 or 0.100.

g. Bidirectional migration between P2 and P3: We conducted 16 simulations as described above for the unidirectional migration scenario, but with migration in both directions (P3 to P2 and P2 to P3).

h. Unidirectional migration for 10% of the genome: We conducted 12 simulations where migration only occurs for a fraction of the genome. We defined two scenarios with an N_e of 500 for all populations and equal time between splits, but with the duration between splits varying between the two scenarios (T_1 = 100, T_2 = 200 T_3 = 300; T_1 = 405, T_2 = 945, T_3 = 1215). For each scenario, we modelled heterogenous migration by combining two simulations together. 90% of genealogies were simulated under a 'background' demography without gene flow, and the remaining 10% were simulated under the same model, but one of six rates of migration from P3 to P2: m = 0, 0.001, 0.005, 0.010, 0.05, 0.01, or 0.05.

i. Ancestral structure: We modeled ancestral structure by combining simulations of two scenarios together, following Martin et al (2015). Ninety percent of the genome was simulated under the 'background' demography (O(P3(P2,P1))), and the remaining 10% under the 'alternate' demography (O(P2(P3,P1))). Split T2 was set to be more ancient in the background to mimic a region of the genome that is polymorphic at particular loci. N_e was set to 500 for all populations for all simulations. No migration was allowed between populations.

3.4.4 | Quantifying asymmetry with the D_{LR} statistic

The results of simulations outlined above show that divergence under an idealized four-population model without migration produces a symmetrical distribution of topology weights between the left and right halves of the ternary plot (figs. S10 & 14; Supplementary text). This is because there is an equal chance that a given gene tree will more closely resemble either alternative topology under incomplete lineage sorting (ILS) (Guerrero and Hahn, 2018; Maddison, 1997). However, deviations from a simple four population model (*e.g.*, gene flow) can lead to a bias in the probability toward one alternative topology, leading to an asymmetrical distribution of topology weights (figs. S15 & S19; Supplementary text).

To measure genealogical bias, we developed a statistic that can be used to detect and quantify asymmetry in the ternary framework. D_{LR}, which is similar to Patterson's D statistic (Green et al., 2006), can range from negative 1 to positive 1, gives and an indication of the

strength of the bias toward one alternative topology relative to the expectation of equality (D_{LR} = 0). A genome-wide estimate of D_{LR} can be calculated directly from the topology weights as:

$$D_{LR} = \frac{(n_{T\gamma > T\beta}) - \left(0.5\left((n_{T\beta > T\gamma}) + (n_{T\beta < T\gamma})\right)\right)}{0.5\left((n_{T\beta > T\gamma}) + (n_{T\beta < T\gamma})\right)} \quad \text{(eq. 1)}$$

Where Tβ and Tγ are the two alternative topologies (see the notation for topologies in Fig. 2A and Guerrero and Hahn (2018)), $n_{T\beta} > n_{T\gamma}$ is the number of windows where Tβ is greater than Tγ, and $n_{T\beta} < n_{T\gamma}$ is the number of windows where Tγ is greater than Tβ. Because these quantities indicate the number of windows in the left and right halves of the triangle, equation 1 can be expressed more intuitively as:

$$D_{LR} = \frac{n_R - (0.5n_T)}{(0.5n_T)} \quad \text{(eq. 2)}$$

Where n_R is the number of trees on the right side of the plot and n_T is the sum of the windows on the left and right sides. Under ILS the expected number of windows in each half of the plot is simply half of the total number of windows (*i.e.*, $0.5n_T$). Note that n_T can be a slightly lower number than the total number of windows analyzed because some may fall precisely on the line that bisects the base of the ternary plot, such that they are neither left- nor right-sided (*i.e.*, when Tβ = Tγ).

In addition to a genome-wide estimate (*i.e.,* comparing the full left- and right-half triangles), symmetry can also be quantified for any arbitrary area of the of the distribution, provided that analogous sections from both sides of the ternary plot are compared. Detailed examples of genome-wide and partial estimates of D_{LR} are shown in fig. S19.

3.4.5 | Advantages over existing site-based statistics

Why use topology-based estimates of genealogical asymmetry when site-based methods already exist? The framework presented here has many advantages over existing methods.

i. *No need for a defined outgroup*—Site-based statistics like Patterson's D require one or more appropriate outgroup taxa that are used to define alleles at each SNP as ancestral or derived. This is challenging (usually requires multiple outgroups for high confidence assignment), and assumes that there is no allele sharing due to gene flow or ancestral structure between the ingroup and outgroup taxa. This is not required in the framework presented here, as estimates of symmetry are inferred from how samples cluster in a tree topology.

ii. *Site-based statistics are low dimensional summaries of topologies*—Given that the underlying genealogies are what we care about, it makes more sense to study them directly, rather than variation at individual sites.

iii. *There is a movement away from site-based analyses to those based on genealogies*—Inference tools like Tsinfer (Kelleher et al., 2019), Relate (Speidel et al., 2019), and ARGweaver (Rasmussen et al., 2014) are making it possible to infer full genomic sequences of genealogical

trees from genome-wide data. Unlike the arbitrary windows that are used in most genomic studies, the genomic spans of the trees inferred with these methods have biological meaning, as their boundaries are defined by recombination events that occurred in the past (*i.e*., the windows are not arbitrary). Topology weighting can be performed directly on these marginal trees (as described in the section *'*Inference of ancestral recombination graphs*'*, pp. 26-27, and as show in Fig. 3E), and the D_{LR} statistic can be calculated directly form the distribution of topology weights.

3.4.6 | Plotting and symmetry analysis of the empirical ternary distribution

Because of the large number of genomic windows analysed, we plotted the empirical distribution of weights using a ternary histogram, where the triangle is divided up into equal-sized hexagons that are coloured to indicate the number of windows that fall into that area of the ternary distribution. This was constructed in the R package ggtern (Hamilton and Ferry, 2018) using the *geom_hex_tern* function.

We tested for asymmetry of the distribution between the left and right side of the diagrams by computing D_{LR} as described above (eq. 1). We used a permutation test (99,999 permutations) to calculate the probability of obtaining the empirical estimate of D_{LR} by chance, assuming that there is an equal probability of a window falling on either side. (Main text, Fig. 2D). To understand how symmetry varies more locally across the ternary distribution, we subdivided each side of the ternary plot into 45 sub-triangles (fig. S24), calculated D_{LR} separately for each one, and used permutation tests to determine the significance of each estimate. Permutation tests were conducted in R using custom scripts.

3.4.7 | Inference of ancestral recombination graphs

The genomic windows used in analysis up to this point (both in this paper and in evolutionary genomic studies in general) are essentially arbitrary, chosen to contain the same number of positions. Ideally, we would partition the genome in a way that coincides with the historical coalescence and recombination events that generated our observed set of genomes.

Here we do this by inferring Ancestral Recombination Graphs (ARGs) from our genomic dataset. The ARG is a graph structure that depicts the entire ancestral history of a set of genomes—including each recombination and coalescence event—that contains the genealogies for each non-recombined genomic block. Thus, by inferring the ARG, we can obtain tree topologies for each inferred non-recombining region, and more precisely infer the genomic span of haplotypes that are perfectly associated with live-birth.

We inferred ARGs using the program ARGweaver (Rasmussen et al., 2014). We focused on two 30kb genomic regions (Contig1808:1-30000 and Contig3201:45000-75000) that spanned areas perfectly associated with reproductive mode. We focused only on these regions because the method is computationally expensive, and these regions are very well sequenced and assembled which is necessary for meaningful ARG inference. Due to

computational limitations, we also restricted the analysis to a subset of 50 individuals (25 egg-layers and 25 live-bearers; *i.e*., 100 haploid samples), as suggested by the authors of ARGweaver. These individuals were chosen to give equal and broad representation across the four clades.

For both contigs, we converted the SNP information from the VCF format to *sites* format, which is compatible with ARGweaver. The *sites* format only includes information on genomic positions that vary within the 100 samples. Positions that were absent in the original VCF were masked from being incorrectly treated as invariant sites while inferring ARGs.

ARGweaver was executed on both contigs using the following parameters: N_e = 20,000, $\mu = r = 1.5 \times 10^{-8}$. ARGweaver simplifies ARG inference by forcing coalescence and recombination events to occur at discrete timepoints. Here we used the 30 specified time points, with the maximum possible TMRCA set to 20*Ne*. ARGweaver was run for 10,000 iterations, out of which the first 5000 iterations were considered as burn-in after visual examination of the MCMC traces and therefore, excluded from subsequent inference. From each of the 5000 remaining iterations, we estimated TMRCA for all the samples, as well as within live-bearers and within egg-layers. These estimates were used to determine the median and confidence limits for the TMRCA shown in main text Fig. 3E.

We also performed topology weighting on the marginal trees from the ARG. This allowed a more precise identification of regions of the genome that were perfectly associated with live-birth (*i.e.*, trees where Tr = 1). To do this, we extracted all marginal trees from the last MCMC iteration (*i.e*., iteration 10,000) and performed topology weighting as described in an earlier section. We used the combined information from the topology weights and TMRCA to identify core haplotype blocks associated with live-birth, as described in Shipilina et al. (2021). Specifically, we used the topology weights and estimates of TMRCA to identify unique edges in the ARG that were both ancestral only to all live-bearing samples (Tr = 1), and that were defined by a single coalescence time. Block inference was performed with a custom R script (*36*).

3.4.8 | Estimates of the ages of live-bearing alleles

We used theory outlined in Tavares et al. (2018) to roughly estimate the age of sweeps at live bearing alleles. During a sweep, diversity is eliminated close to the selected site. In the case of a hard sweep, all ancestral diversity is eliminated from the swept region, but new mutations in the swept segment at rate μ will generate $\pi_w = 2\mu T$ at T generations after the sweep. Thus, the time of the onset of the sweep can be estimated as $T = \pi_w/2\mu$. Diversity arising in other ways complicates the estimation, causing T to be overestimated. For example, some ancestral variation may be retained if the beneficial allele was segregating on multiple haplotypes before the sweep (soft sweep). This means that π_w would reflect mutations that arose both before and after the sweep. Similarly, gene flow and recombination between populations may introduce additional variation into a swept region that did not arise though

new mutation. Finally, it is difficult to define precise sweep boundaries, and error in doing so would tend to inflate π_w.

To correct for these confounding sources of diversity, we estimated the age of each sweep (*i.e.*, the time in generations that each sweep began) by calculating π_w from private alleles. This assumes that alleles private to live-bearers arose in the swept segment after the sweep, and that alleles shared between egg-layers and live-bearers reflect some confounding factor. However, this approach would over-estimate the age of a soft sweep if the private alleles were already private to haplotypes carrying the beneficial allele when the sweep began.

For each of the 50 regions associated with reproductive mode (table S9), we identified and then removed SNPs that were polymorphic in both egg-layers and live-bearers (see section 'Analysis of private alleles', p. 23). We then calculated π_w for live-bearers as described above (see section 'Estimating diversity, divergence and differentiation in 100 SNP windows', p. 21). We corrected the denominator in the π_w calculation for the removal of the shared sites.

For calculating *T*, we assumed a mutation rate of $\mu = 1.5\times10^{-8}$. We converted time in generations into years using two estimates of generation time for *Littorina saxatilis*: 2 generations per year and 1 generation per year (Westram et al., 2018). Estimates of time in generations varied roughly ten-fold, ranging from 23,266 to 227,094 generations before present, with a median of 69,790 generations before present (Fig. S32). In years, this equates to a minimum of 11,633 – 23,266 years before present, and maximum of 113,547 – 227,094 years before present, with a median of 34,895 – 69,790 years before present. The distribution of ages is strongly negatively skewed, with the majority of sweeps being initiated in the last 100,000 generations.

3.5 | Supplementary Information

Detailed supplementary information is available at: https://doi.org/10.1126/science.adi2982
Along with the core theme of this thesis, supplementary information related to *TwisstNTern* and genealogical analysis (contributed by AP) is in Appendix B.

Chapter 4
Haplotagging pipeline for 1,084 whole-genome sequences in an *Antirrhinum* hybrid zone

Abstract

Recent advances in linked-read sequencing, particularly *haplotagging*, have made it feasible to generate population-scale, phased whole-genome datasets in non-model organisms. Here, we present a comprehensive pipeline for sequencing 1,084 individuals from an *Antirrhinum majus* hybrid zone. By benchmarking imputation against highly accurate KASP genotypes, we have found >92% genotype calls at ~20 million variable sites. Molecular phasing of high-coverage samples (≥5x) provided a robust reference panel for statistical phasing of all individuals, achieving switch error rates in order of ~5%. Allele polarity was determined using the closely related *A. molle*, with robust classification of ~86% of variant sites into ancestral or derived. The resulting genomic resource—the largest of its kind in this system—offers a valuable dataset to study the process of speciation in a hybrid zone and a transferrable methodological framework to other non-model species.

4.1 | Introduction

Recent advances in sequencing and computational methods have revolutionized evolutionary genomics, enabling detailed investigations of the genetic basis of adaptation, speciation, and population history (Lou et al., 2021; Meier et al., 2021; Nielsen et al., 2024; Pokrovac and Pezer, 2022). This progress allows genome-wide scans for selection, demographic reconstruction, trait mapping, and fine-scale population structure analysis and more, including in non-model species (Cheng and Steinrucken, 2024; Liu et al., 2014; Salojarvi et al., 2024; Suda et al., 2025; Zhang et al., 2024). Consequently, the field has transitioned decisively into the genomic era, focusing on adaptive architecture and genomic consequences of gene flow, selection, and recombination (Whiting et al., 2024; Todesco et al., 2020; Shi et al., 2023).

Despite this progress, most population genomic studies still rely on short read technologies like Illumina due to their cost-effectiveness and high throughput, enabling genotyping of thousands of individuals at relatively low cost. Since genetic material is passed through generations in sets of linked variants rather than single sites, access to long haplotype information (i.e., specific allelic combinations inherited on the same chromosome) is critical for inferring evolutionary processes, including recombination, gene flow, and adaptation, that are otherwise difficult to study without accurately phased data (Garud et al., 2015; Leitwein et al., 2020; Sedghifar et al., 2016; Shipilina et al., 2023; Tewhey et al., 2011; Todesco et al., 2020). Short reads inherently limit haplotype reconstruction because their lengths (50-300bp) rarely span multiple variants, complicating phasing of distant alleles and assembly through repetitive or structurally complex genomic regions (Ebert et al., 2021). In contrast, long-read sequencing platforms (PacBio, Oxford Nanopore) offer greater read lengths that could significantly improve imputation, phasing, and structural variant detection, etc, but remain prohibitively expensive and lower throughput for large populations (Jain et al., 2018; Logsdon et al., 2020). Long read sequencing also typically requires high-quality, high molecular weight DNA, that can be challenging for non-model organisms. Therefore, although short reads dominate due to affordability and scalability, their limitations motivate integrating long-read data where feasible to capture more complete genetic variation and haplotype information.

To overcome the limitations of short-read sequencing in resolving long-range haplotypes, a number of linked-read technologies have been developed. Linked-read sequencing techniques tag long DNA molecules before fragmenting and sequencing them, thus preserving long-range haplotype information while still using standard short-read sequencing platforms. Notably, haplotagging, a linked read method introduced by Meier et al. (2021), provides a cost-effective, scalable solution by utilizing a transposase-based protocol to uniquely barcode long DNA fragments before sequencing. This tagging enables the association of short reads back to their original long molecules, thus allowing molecular haplotypes to be reconstructed across 10–100Kb. Unlike some commercial linked-read options such as 10x Genomics, haplotagging does not require specialized instruments or proprietary reagents and is readily compatible with standard Illumina platforms, greatly

lowering barriers to widespread adoption. This makes haplotagging particularly suitable for studies demanding large sample sizes, diverse species, and population-level resequencing at variable coverage—challenges often encountered in ecological and evolutionary genomics, especially for non-model organisms. Beyond phasing distant variants, barcoded molecules can improve alignment accuracy and reduce mapping bias in repetitive or highly polymorphic regions as well as help identify structural variants, further enhancing the utility of the data. By combining the affordability, accuracy, and throughput of short-read sequencing with long-range linkage information, haplotagging offers a pragmatic path forward for generating comprehensive haplotype data at population scale, especially in non-model organisms that often lack extensive genomic resources.

The snapdragon study system where two varieties of *Antirrhinum majus* subspecies *majus*—var. *pseudomajus* and var. *striatum*—hybridise to form a narrow hybrid zone in the Spanish Pyrenees, presents a great opportunity to use haplotagging to sequence an entire population from a hybrid zone transect (~1000 individuals). This hybrid zone has been studied since 2008, with ~50,000 individuals sampled over a decade, but has only been sequenced at ~100 SNP markers or with PoolSeq to answer a range of questions from understanding the dynamics of how the hybrid zone is maintained to analysing cline patterns at divergent SNP markers to estimate strengths of selection (Ringbauer et al., 2018; Surendranadh et al., 2025, 2022; Tavares et al., 2018; Whibley et al., 2006). Haplotagging, now makes it possible to transition into whole genome sequencing along the hybrid zone, to study the genomic landscape of speciation divergence with bottom-up approaches like F_{ST}, genetic basis of traits under selection with top-down methods like GWAS, and reconstruct a genealogical history of the regions that act as barriers to gene flow.

Since haplotagging is a rather new method, there is no best-practice downstream pipeline available for processing the raw sequences. In this chapter, I describe the analytical pipeline we devised to generate a high-quality haplotype-resolved dataset using haplotagging in *Antirrhinum*. In addition to testing the accuracy of this method in our system, we aim to provide a methodological framework for others seeking to apply haplotagging to their system. Our approach not only benchmarks key tools and parameters such as imputation from low-coverage sequencing, but also illustrates the trade-offs between accuracy, coverage, and missingness that are inherent to population-scale sequencing projects. The lessons from this study are broadly applicable to other non-model systems. More generally, we show that haplotagging, when coupled with careful data processing and imputation, can yield high-resolution haplotypes across hundreds of individuals, opening new avenues for evolutionary genomic analyses in non-model organisms.

4.2 | Study system and sampling

We collected and sequenced 1084 individuals of *Antirrhinum majus* (Supplementary Data 1). Most of the samples were collected near Planoles in the Spanish Pyrenees. The previously studied area includes a hybrid zone between two varieties—*Antirrhinum majus pseudomajus*

and *A. m. striatum* (Field et al., 2025; Ringbauer et al., 2018; Surendranadh et al., 2025, 2022; Tavares et al., 2018; Whibley et al., 2006). These varieties are distinct in their flower colour, with *A. m. pseudomajus* to the east having magenta flowers and *A. m. striatum* to the west having yellow flowers. The difference in flower colour is largely due to three major-effect loci *Rosea* (*ROS*), *Eluta (EL)* and *Sulfurea* (*SULF*) (Bradley et al., 2017; Schwinn et al., 2006; Tavares et al., 2018). *ROS* and *EL* are physically linked on chr 6 (Schwinn et al., 2006; Tavares et al., 2018), while *SULF* is located on chr 4 (Bradley et al., 2017). Within the narrow hybrid zone (~1Km), hybridization between the subspecies has caused alternative alleles at these loci to segregate and recombine, resulting in diverse phenotypes.

The majority of sampled plants (*n* = 1015) were distributed along a ~3 km stretches of road (GIV-4015, colloquially hereafter referred to as *Lower Road*) that traverses this *Antirrhinum* hybrid zone (Fig 1). Most plants (960) were sampled between May and August of 2021, while the rest were sampled in either 2017 or 2019. This road has been previously divided into three areas largely according to the difference in flower colour: the core of the hybrid zone (the central 1 km segment) and the yellow and magenta flanks on the western and eastern sides of the core. In the core area, we sampled every individual that we could safely access (*n* = 763). In addition, we also collected samples from three discrete sites in each flank (22–29 samples per site), positioned at varying distances from the core area (magenta flanks 1, 2, 3 and yellow flanks 1, 2, 3).

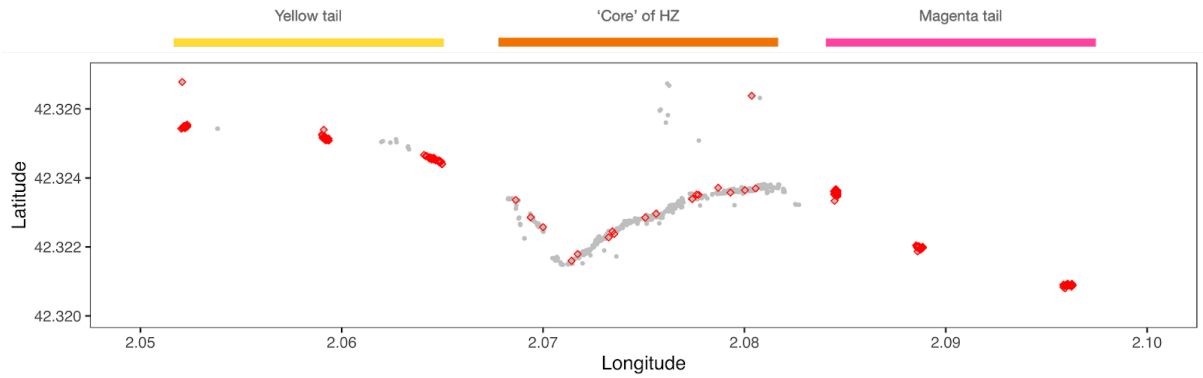

Figure 1. All samples (n = 1015) along the Planoles hybrid zone. Red dots indicate samples sequenced at ≥5x coverage (n = 180). The rest of the samples are from Avellanet hybrid zone (Pal et al., 2025).

We also included samples from nearby locations and species as outgroups. This included 19 samples of each *A. majus* subspecies (*n* = 38) from a second hybrid zone located near the town of Avellanet, located roughly 80 km west of Planoles. Informed by a phylogenetic study on the *Antirrhinum* genus we also included several other species to serve as outgroups (Durán-Castillo et al., 2022). This included 4 individuals of *A. molle*, a sister species to *A. majus*; and 1 individual from other more distantly related species—*A. sempervirens*, *A. latifgolium*, *A. siculum*, *Misopates orontium* and *Lantana salzmannii.*

We also sequenced 9 parent-offspring trios and 4 full-sibs without parents in order to evaluate the performance of imputation and phasing. Four of the trios sequenced are made up of 12 plants from our field site at Planoles and were identified through the construction of

a large, multi-generation pedigree (Surendranadh et al., 2022). The remaining 6 trios and full sib offspring were generated in the greenhouse at the Institute of Science and Technology Austria (ISTA). The trios were produced by raising 3 plants from field collected seed and crossing them in all possible combinations to produce 6 full and half-sib offspring (Table S1).

4.3 | DNA extraction

For all samples of *A. majus* and *A. molle*, several leaves were collected from each individual in the field or greenhouse and refrigerated at 4ºC overnight before processing. DNA was preserved by placing leaf tissue in a paper envelope, and then placing envelopes into an air-tight plastic bag with silica gel. Leaf samples of the other five outgroup species were frozen in liquid nitrogen and kept at -80ºC prior to DNA extraction.

DNA was extracted from a 1 cm by 0.5 cm piece of dry leaf using a custom protocol optimised for isolating high molecular weight DNA. (i) Tissue was first crushed to a fine powder in a 1.5 ml Eppendorf tube using a micropestle. (ii) 0.5 ml of lysis buffer was added, containing 400 µl PureLink Genomic Digestion Buffer (Thermofisher: K182301), 40 µl Proteinase K (20mg/ml), 60 µl of 18% Polyvinylpyrrolidone (40 kdaG), and 2% Beta-mercaptoethanol. (iii) Samples were incubated for 15-20 mins at 60ºC with mixing at 950 rpm with occasional inverting of the tube, followed by 30 secs on ice. (iv) 10 µl of 5 mg/ml RNaseA was added and mixed by inverting the tube several times, and incubated for 5 mins at room temperature. (v) 155 µl 5M of potassium acetate was added and immediately mixed by inverting 10-15 times, followed by incubation for 2 mins on ice. (vi) The samples were centrifuged at 13k g for 10 mins at 14ºC and 400 µl of lysate was transferred to a new tube using a wide-orifice tip. (vii) 200 µl magnetic beads were added and immediately mixed by inverting 10-15 times and incubated for 10 mins at room temperature. (viii) Samples were pulse-spun for 2 seconds, and put on a magnet stand for 5 mins and the lysate was discarded. (ix) The beads were washed twice for 1 min with 80% EtOH; All EtOH was carefully removed after the second wash and the sample left open to evaporate for 2-4 mins at room temperature. (x) Samples were removed from the magnet and 100 µl of 10 mM Tris pH=8, 0.2mM EDTA was added to the tube and incubated for 10 mins at 42ºC to elute the DNA. (xi) Samples were mixed gently by inverting to re-suspend the beads and allowed to incubate for 10 mins at 42ºC. Eluted DNA was stored with beads at 4ºC.

The concentration of each DNA sample was determined using a Qubit fluorometer or an M200 Pro Tecan plate reader. A rough estimate of size and quality was determined using agarose gel electrophoresis. Samples were diluted to 5 ng/µl using Tris, pH=8, 0.2mM EDTA and stored at 4ºC for future use.

4.4 | Library preparation and sequencing

We used *haplotagging* to sequence the 1074 samples. Unlike standard Illumina short read sequencing, *haplotagging* enables the molecular barcoding of each DNA molecule so that

sequencing reads can be associated with their parent DNA strand. Sequencing libraries were constructed by mixing genomic DNA with a pool of *haplotagging* beads and incubating for 10 mins at 55°C followed by Tn5 stripping (0.3% sodium dodecyl sulphate, 10 min at 55°C) and PCR amplification essentially as described in Meier et al. (2021). Amplified libraries were cleaned up and size-selected using Ampure magnetic beads (Beckman Coulter), Qubit quantified, and adjusted with 10 mM Tris, pH 8, 0.1 mM EDTA to 2.5 nM concentration for sequencing.

Libraries were sequenced with Illumina paired-end sequencing (2x 150 bp) across 5 lanes of Novaseq 6000 S4 by Azenta Life Sciences (Leipzig, Germany), yielding a total of 16.3 billion reads. For one lane, the BX tag (*i.e.,* the molecular barcode used to identify each molecule) was not sequenced correctly, effectively making 83.79% of all reads in that sequencing lane equivalent to standard Illumina short-read data. Thus, the data for most individuals includes a mix of BX-tagged and non-BX tagged reads. The specific details for the construction and sequencing of each library can be found in Table S2.

We used a tiered approach, sequencing sets of individuals to three different levels of coverage. This design allows for improved haplotype reconstruction in high coverage individuals without compromising population level coverage. About 15% of samples (*n*=173) were sequenced to an expected depth greater than 10x coverage. These were mainly from the 6 discrete sample sites in the flanking regions of the hybrid zone, with a small number (*n*=12) from the core area. Samples from the parent offspring trios were sequenced to roughly 5x coverage. The remaining samples from Planoles and Avellanet were sequenced to a mean depth of 2x coverage. Specific details of sequencing depth of each sample is provided in Supplementary Data 1.

4.5 | Processing of raw haplotag reads, variant discovery and filtering

We mapped raw reads to the *Antirrhinum majus* reference genome v. 3.5 (Li et al., 2019) using *EMA v0.7.0* (Shajii et al., 2018a), a BX-tag-aware variant of *BWA* (Li, 2013), thus favouring alignments that are consistent with having come from a single molecule. Prior to mapping, we translated haplotag barcodes with BX tags into 16-bp barcodes using the *16BaseBCGen* tool (https://tinyurl.com/16baseBCgen). In a pre-processing step, *EMA* first counts the number of barcodes, distinguishes between correct and faulty BX-tags, and places the correct barcodes in 500 bins to facilitate parallel mapping. Reads with faulty BX-tags, which made up less than 5% of the reads from each correctly sequenced lane are grouped separately (Table S2). The reads in each barcode bin were then aligned to the reference genome; the reads with faulty BX-tags were aligned to the genome using *BWA v0.7.17*. More than 96% of reads successfully mapped for all 5 lanes (Table S2).

After alignment, BAM files from all bins were merged and indexed with *sambamba v0.8.2* (Tarasov et al., 2015) to produce a single BAM file for each individual. 16bp barcodes were then reverted to haplotag BX-tags using *samtools v1.18* (Danecek et al., 2021) and a

custom *awk* script. For some individuals, raw sequence data were obtained from more than one sequencing run. If this was the case, read mapping was performed as above for the output of each sequencing run separately, and the resulting BAMs were then merged together and indexed with *sambamba* to create a single indexed BAM for each individual. Each merged BAM file was then checked for quality using the *multi-bamqc* command in *qualimap v2.2.1* (Okonechnikov et al., 2016). We also used the program to compute a number of summary statistics for each sample BAM (e.g., coverage, mean mapping quality, and median insert size; Supplementary Data 1). Prior to variant calling, we marked and removed PCR and optical duplicates from subsequent analyses using the *markdup* tool in *sambamba*.

Table 1. Summary of filtering steps and the change in sites at each filtering step in *bcftools*. Total sites includes all sites that are invariant and variant including SNPs, INDELs, and other complex variants. Total variant sites include true variant sites and invariant sites that are homozygous for the alternative (ALT) allele (following standard naming conventions of *bcftools*). The final count after all filtering steps consists of only bi-allelic variant sites.

Dataset	Total Sites	Total invariant sites	Total variant sites	Total multi-allelic variant sites
Full unfiltered dataset	429,764,911	383,131,005	52,439,232	3,369,036
After removing homozygous REF/REF sites	54,976,284	0	52,439,232	3,369,036
After removing INDELs, variant sites within 5bps of INDELs	48,089,635	0	48,089,635	3,073,072
After removing homozygous ALT/ALT sites	46,774,818	0	46,774,818	3,073,072
After removing multi-allelic variant sites	43,701,746	0	43,701,746	0
After dropping sites based on DEPTH, QUAL and MQ	28,956,743	0	28,956,743	0

We next used the *mpileup* and call commands in *bcftools* (Danecek et al., 2021) to identify a set of candidate SNPs that were subsequently used for final genotype calling and imputation using the program *STITCH* (Davies et al., 2016). This was performed with the multi-allelic calling program, *-m* and *--annotate AD,ADF,ADR,DP,QS,SP*. This initial variant calling discovered 43,791,746 candidate SNPs across the 1074 samples.

We next filtered these variants to remove sites that are likely to be errors. To inform this filtering, we used *bcftools stats* to produce distributions of per-site depth, mapping quality and genotype quality across all samples using the following options: *--af-bins <(seq 0*

0.1 1) --depth 0,25000,25. After visual inspection of these results, we filtering the variants using *bcftools* as follows: we removed all monomorphic REF sites (*bcftools view -m 2*) (NB: REF refers to the alleles present in the reference genome, not ancestral allele), all insertions and deletions (bcftools view -V indels), all SNPs within 5 base pairs of indels (bcftools filter –SnpGap 5), all monomorphic ALT sites (bcftools view -e "AC==AN || AC==0") (NB: ALT refers to alleles alternate to the reference genome, not derived allele), and all sites with more than 2 alleles (bcftools view -M2). We then filtered the remaining biallelic SNPs, removing all sites with a total depth of <500 or greater than >7732 reads (i.e., ≥2.5 times the mean coverage). We also dropped sites with a genotype quality score of <20 and a mapping quality score of <30, as defined in bcftools (bcftools filter -e “INFO/DP<500 | INFO/DP>7732 | QUAL<20 | MQ<30”). The number of sites removed by each filtering step can be found in Table 1. After these steps, we retained 28,956,743 putative variant positions that were used to inform subsequent calling in STITCH (Table 1).

4.6 | Optimisation of SNP calling and imputation

4.6.1 | Generation strategy for optimisation

We used the program *STITCH* (Davies et al., 2016) to call variants at the putative variant positions identified using our bcftools pipeline. *STITCH*, models each chromosome in the population as a mosaic of *K* unknown founders or ancestral haplotypes using both the underlying sequence reads and the linked-read information encoded in the BX-tag. Unlike traditional callers, *STITCH* calls (imputes) genotypes in the presence of missing data based information available from haplotype information across all sequenced individuals.

STITCH can be optimised by varying a number of model parameters – including the number of founding haplotypes (*--K*), the estimated number of generations since founding (*--nGen*), recombination rate (*--expRate*) and tuning parameters, such as the number of MCMC iterations (*--niterations*) and read coverage downsampling (*--downsampleToCov*). The quality of *STITCH* calls is usually evaluated using the *INFO_SCORE* for each site. The *INFO_SCORE* represents the amount of information gained by imputing the missing genotype and indicates its confidence (expected imputation accuracy). In short, it is derived from the posterior genotype probability distribution and is calculated as the ratio of the difference between the probability of imputed genotype and the most likely observed genotype to the maximum possible difference between the two probabilities. In addition to the *INFO_SCORE*, we also evaluated and optimised the accuracy of calling by comparing the *STITCH* genotypes against high-accuracy genotypes at 30 SNP markers obtained through *KASP* genotyping (He et al., 2014). These markers were previously developed for *A. majus* (Ringbauer et al., 2018) and have a near-zero error rate, determined as the number of mismatches in the repeat genotyping of ~3,000 individuals. Thus, genotypes obtained by *KASP* genotyping our new individuals are assumed hereafter as true genotypes when subsequently evaluating the accuracy of *STITCH* calls.

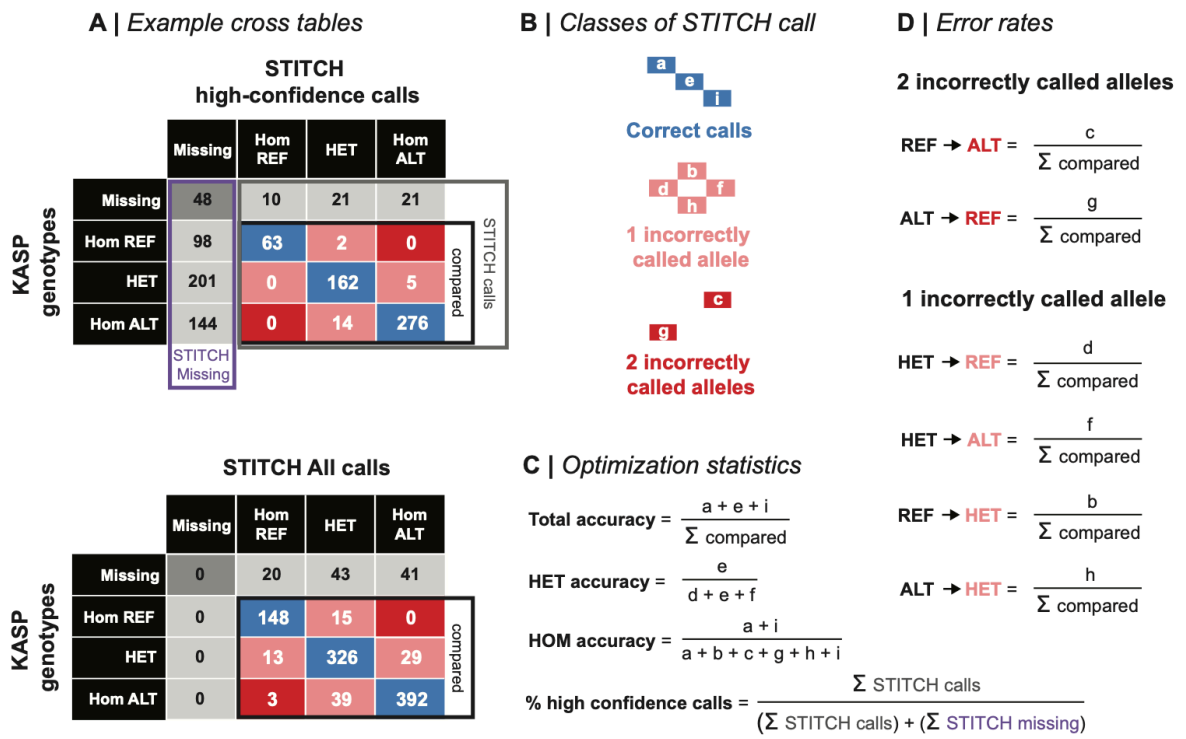

Figure 2. The scheme and metrics used to optimise STITCH calling. (A) Example cross tables for one marker showing the agreement between the STITCH calls and KASP genotypes. The top table shows the cross table for high-confidence STITCH calls while the bottom cross table is for all calls. (B) images showing the locations of the correct and incorrect STITCH calls in the cross tables. The letters in each box are used in the formula in *C* and *D*. (C) Statistics used to optimise STITCH calling, including the total accuracy (i.e., total proportion of correct calls), HET and HOM accuracy (i.e., the proportion of correct heterozygous and homozygous calls, respectively), and the proportion of high-confidence calls. (D) Estimates of the rates for the 6 possible types of genotyping error.

We focused our *STITCH* optimization on 100Kb genomic spans centred around the 30 *KASP* markers. We only included samples of *A. majus* in *STITCH* calling, with the 1074 BAM files and the list of putative variant positions and identity of REF and ALT alleles (identified based on the reference genome) provided as input. To evaluate each run, we used the following metrics (Fig 2) — (*i*) the mean INFO score across all variant sites within the 100Kb region, (*ii*) the accuracy, calculated as the fraction of perfect matches between *STITCH* genotypes and the *KASP* genotypes, (*iii*) the proportion of correctly called heterozygotes (i.e., the fraction of heterozygous *STITCH* genotypes where the *KASP* genotype is also heterozygous), and (*iv*) the fraction of correctly called homozygotes (i.e., the fraction of homozygous *STITCH* genotypes where the *KASP* genotype is also homozygous).

By default, *STITCH* outputs only genotypes that were imputed at high-confidence (i.e., sites with a posterior genotype probability >0.9 for the most probable genotype), hereafter, called high-confidence calls. However, we also obtained the genotype with the highest-probability for all positions including those that did not meet the cut-off for high confidence (hereafter, called all calls). We computed the above statistics separately for high-confidence calls and for all calls. Moreover, initial *STITCH* runs indicated that the accuracy of calls

improved drastically at a threshold of around 5x coverage, so we also assessed accuracy separately for samples greater than 5x coverage. (>5x coverage, *n*=180).

4.6.2 | Effect of parameters on imputation accuracy

We started our exploration of the STITCH parameter space by varying the number of founding haplotypes, *K*, since preliminary runs showed that *K* had the largest effect on quality. We performed runs with 22 different values of *K*, ranging from 5 to 160, while keeping the values of other parameters constant based on values that performed well in preliminary runs (*--nGen=100, --niterations=40, --downsampleToCov=20, --expRate=5*).

For all calls, we found that accuracy tended to be highest and fairly similar between *K*=40 and *K*=80, with poorer performance at lower and higher values. Accuracy for the high-confidence calls showed a different pattern, with the best performance observed at lower values of *K* (5 to 45) and gradually decreasing with increasing *K*. Accuracy tended to be higher for samples with more than 5x coverage compared with those with less than 5x coverage for both high-confidence calls and all-calls, though the effect of coverage was much greater for all-calls (Table 2).

Although the accuracy was generally higher for high-confidence calls, runs were always associated with a large amount of missing data when only high-confidence calls were considered (mean of 38% to 57% across the range of *K* values). We therefore chose a value of *K*=75, which maximised the accuracy for all calls while also showing good performance for high-confidence calls (i.e., total accuracy > 95%). For individuals with greater than 5x coverage, the total accuracy was 94.2% with 93.6% of heterozygotes called correctly and 94.4% of homozygotes called correctly. For all samples, the average accuracy was 89.2%, with 88.2% of heterozygotes and 89.8% of homozygotes called correctly. After optimising *K*, we tested the effect of the other model and tuning parameters including the number of generations since the population was founded (*--nGen*), the number of MCMC iterations (*--niterations*), the recombination rate (*--expRate*) and the maximum coverage included for each individual (*--downsampleToCov*). For these runs we kept *K*=75 constant and tested a range of values (Table 3). Only the maximum per-sample coverage had a discernible effect on accuracy, probably owing to the large variance in coverage across all samples. The final set of parameters chosen for genome-wide *STITCH* imputation were: *--K=75 –-ngen=100 --niter=40 --downsampleToCov=20 --expRate=0.5cM*.

Focusing only on the the results for this set of parameters, we characterised error rates in more detail by focusing on the six possible errors that could occur at the genotype level. We found that majority of the errors (11.2% for all samples, 5.9% for samples with >5x coverage) involved only one incorrectly imputed allele, i.e., inferring a heterozygous genotype as homozygous, or vice-versa, with both error types being roughly equally common. In contrast, we observed very few positions where both alleles were incorrectly inferred, resulting in switches between alternative homozygotes (similar rates in both directions). The

average rate for this error class was less than 0.19% for all samples, and less than 0.07% for the samples with >5x coverage (Table 4).

4.6.3 | Comparison of accuracy between SNP-calling by STITCH and hard calling by bcftools

To put the observed error rates into perspective, we compared the accuracy of imputed genotypes to those obtained via a widely used variant calling and filtering pipeline. Specifically, we extracted the calls from our bcftools pipeline after quality filtering (described above) and obtained accuracy estimates using the 30 KASP markers as described above. Since the accuracy of hard-calling is highly dependent on individual coverage (i.e., at least 2 reads are needed to infer a heterozygous genotype, while many reads needed to be statistically confident), we estimated the error for the hard-called genotypes using three different per-genotype coverage cutoffs (i.e., the minimum number of reads needed to include a genotype in the error estimation): (*i*) no cutoff, where genotypes supported by any number of reads were included, (*ii*) only genotypes supported by at least 6 reads (*-e 'FMT/DP<6'*) and, (*iii*) genotypes supported by at least 15 reads (*-e 'FMT/DP<15'*). These three accuracy estimates were calculated for three sets of samples grouped by their mean per-site coverage: (*i*) all samples, (*ii*) samples with >5x mean coverage, and (*iii*) samples with >15x mean coverage. This provided a total of 9 accuracy estimates that could be compared to the *STITCH* accuracy estimates. We also estimated the fraction of missing data for each of the 9 categories to compare to the post-imputation missing fraction, which is always 0.

The results of this comparison are clear, and nicely highlight the strengths and weaknesses of imputation and hard-calling as the per-genotype and mean sample coverage vary. In general, the accuracy of the hard-calling decreases as the per-genotype coverage increases (Table 5). The mean accuracy across the 30 *KASP* markers, lowest at 74% with 63% missing data across all samples (*n*=1070) when no per-genotype coverage cutoff was applied. The accuracy of *STITCH* was 15% higher, at 89% with no missing data. *STITCH* also out-performed hard-calling when samples with >5x coverage (*n*=180) were analysed with no genotype-depth cutoff, with an accuracy of 94%, compared to 88% accuracy and 3% missing data. When genotypes were supported by at least 6 reads, the hard-call accuracy increased to 95.6%, but with 64% missing data. The highest hard-call accuracy of 97% was observed for the set of samples with >15x coverage (*n*=15) when genotypes supported by at least 10 reads, but still 37% of the data were missing because of this strict filtering (Table 5).

By comparing the hard calls and *STITCH* calls, we are able to conclude that, although the error rates for *STITCH* are appreciable, *STITCH* calling results in considerably more data per sample at the cost of a relatively modest increase in mean error compared with the hard-calling of genotypes from relatively high coverage data. For example, the stitch accuracy for our 180 samples with >5x coverage (94%) is only about 3% lower than the accuracy that we would expect if all samples were sequenced at 15x at every variable site. Overall, *STITCH*

		Accuracy of all samples (*n=1070*)			Accuracy of samples with >5x coverage (*n=180*)		
		Total	Het	Hom	Total	Het	Hom
K	***INFO***	*all STITCH calls*					
5	0.320	79.90%	86.68%	73.86%	88.56%	90.77%	86.54%
10	0.389	82.96%	84.50%	81.01%	88.44%	89.67%	86.91%
15	0.432	83.85%	85.00%	82.50%	88.44%	89.70%	87.55%
20	0.465	86.06%	86.71%	85.20%	90.99%	91.20%	90.63%
25	0.494	86.88%	87.12%	86.39%	91.28%	91.11%	91.32%
30	0.517	87.53%	87.29%	87.45%	92.84%	92.36%	93.10%
35	0.535	87.94%	87.63%	87.89%	92.88%	91.62%	93.55%
40	0.551	88.21%	87.40%	88.52%	92.88%	92.22%	93.34%
45	0.566	88.57%	88.03%	88.73%	93.03%	92.59%	93.21%
50	0.577	88.85%	88.09%	89.21%	93.61%	93.31%	93.73%
55	0.589	88.86%	88.15%	89.23%	93.57%	93.22%	93.77%
60	0.599	88.88%	88.09%	89.24%	93.88%	93.70%	94.00%
65	0.607	88.84%	87.56%	89.66%	93.79%	93.66%	93.98%
70	0.615	88.58%	87.60%	89.24%	93.71%	93.20%	94.11%
75	0.622	89.21%	88.22%	89.79%	94.16%	93.55%	94.38%
80	0.629	89.03%	87.04%	90.34%	93.73%	91.67%	94.84%
85	0.637	88.22%	86.56%	89.30%	93.60%	91.96%	94.74%
90	0.642	88.31%	86.73%	89.41%	94.12%	92.89%	95.02%
95	0.648	88.08%	86.15%	89.41%	93.41%	91.94%	94.40%
100	0.653	87.77%	86.36%	88.66%	93.41%	92.42%	94.01%
120	0.673	87.50%	85.35%	88.98%	93.36%	91.72%	94.21%
160	0.704	86.21%	83.65%	88.14%	92.93%	90.77%	94.64%
K	***% calls***	*high-confidence STITCH calls*					
5	43.45%	97.35%	98.14%	96.64%	98.43%	99.66%	97.48%
10	47.64%	98.19%	97.54%	98.34%	98.87%	97.88%	98.93%
15	44.99%	97.56%	96.65%	97.90%	98.17%	97.12%	98.73%
20	48.23%	97.64%	97.48%	97.65%	98.47%	97.71%	98.68%
25	48.21%	97.69%	96.74%	98.10%	99.33%	98.87%	99.45%
30	49.01%	97.16%	96.03%	97.58%	98.44%	98.42%	98.71%
35	51.43%	97.30%	96.59%	97.63%	98.42%	97.31%	98.85%
40	47.48%	96.75%	95.14%	97.36%	97.85%	96.98%	98.37%
45	50.50%	96.87%	95.67%	97.30%	98.37%	97.75%	98.57%
50	50.67%	96.78%	95.95%	97.20%	98.09%	98.21%	98.06%
55	51.17%	96.24%	94.68%	96.95%	97.39%	96.97%	97.99%
60	48.40%	95.51%	93.94%	96.17%	97.76%	96.63%	98.08%
65	52.70%	96.28%	95.11%	96.89%	97.77%	97.87%	97.69%
70	51.23%	95.22%	93.11%	96.26%	97.56%	96.94%	97.79%
75	53.18%	95.45%	94.42%	96.01%	97.13%	97.13%	96.92%
80	53.31%	95.52%	93.48%	96.59%	97.51%	95.76%	98.05%
85	53.64%	94.49%	91.76%	96.04%	97.11%	95.75%	98.06%
90	53.70%	94.50%	92.02%	95.97%	96.71%	95.40%	97.36%
95	55.11%	93.98%	91.37%	95.49%	96.47%	95.28%	97.07%
100	56.58%	94.16%	91.85%	95.39%	96.77%	95.63%	97.53%
120	56.96%	92.50%	89.29%	94.31%	97.07%	95.80%	97.51%
160	61.94%	91.71%	88.73%	93.56%	95.43%	93.97%	96.54%

Table 2. Effect of the number of ancestral haplotypes (*K*) on STITCH accuracy. Each row shows the results for a run with a different value of *K*. The INFO score and accuracy of STITCH calls (defined as the percentage of matching STITCH and KASP genotypes) is provided for each run. Bluer values have a higher accuracy, redder values a lower accuracy. Accuracy was calculated separately all genotypes, and heterozygous and homozygous genotypes and for all samples and for samples with >5x coverage. Calls are further divided between high-confidence calls (posterior probability of most likely calls > 0.9), and all calls. The highlighted box shows the chosen *K*=75 for final genome-wide *STITCH* calling.

Table 3. Effect of the number of generations since founding *(ngen)*, number of MCMC iterations *(niter)*, number of down-sampled reads for *STITCH* calling *(cov)*, and expected recombination rate *(r)* on *STITCH* accuracy. Each row shows the results for a run with a different value of the above parameters. The INFO score and accuracy of *STITCH* calls (defined as the percentage of matching *STITCH* calls and KASP genotypes) is provided for each run. Bluer values have a higher accuracy, redder values a lower accuracy. Accuracy was calculated separately for all genotypes, and heterozygous and homozygous genotypes and for all samples and for samples with >5x coverage. Highlighted boxes for each parameter (*ngen*=100, *niter*=40, *coverage*=20, *r*=0.5) shows the chosen values for final genome-wide *STITCH* calling.

		Accuracy of all samples (*n=1070*)			Accuracy of samples with >5x coverage (*n=180*)		
	INFO	Total	Het	Hom	Total	Het	Hom
ngen	*(no. of generations since founding)*						
100	0.622	89.21%	88.22%	89.79%	94.16%	93.55%	94.38%
200	0.614	88.84%	87.33%	89.85%	93.67%	92.96%	94.21%
400	0.609	88.51%	87.01%	89.57%	93.73%	92.96%	94.32%
800	0.601	89.04%	87.82%	89.88%	94.22%	93.60%	94.78%
niter	*(no. of MCMC iterations)*						
40	0.622	89.21%	88.22%	89.79%	94.16%	93.55%	94.38%
60	0.628	88.29%	86.72%	89.35%	93.10%	91.58%	94.04%
80	0.630	88.55%	87.18%	89.41%	93.33%	92.52%	93.91%
100	0.630	88.28%	86.72%	89.31%	93.19%	91.94%	94.06%
coverage	*(downsampled read coverage in BAM files)*						
5	0.703	85.91%	84.58%	86.73%	91.23%	89.31%	92.23%
10	0.622	89.21%	88.22%	89.79%	95.45%	94.42%	96.01%
15	0.574	88.76%	87.71%	89.41%	95.66%	93.62%	96.82%
20	0.567	89.75%	88.65%	90.42%	96.85%	94.41%	97.78%
25	0.562	89.28%	88.26%	89.92%	96.88%	96.01%	97.38%
30	0.558	89.19%	88.11%	89.84%	97.61%	96.63%	97.86%
35	0.555	89.22%	87.97%	89.99%	97.05%	95.74%	97.62%
40	0.553	89.38%	88.03%	90.21%	97.27%	95.91%	97.77%
r	*(expected recombination rate; cM/MB)*						
0.1	0.622	88.76%	87.80%	89.36%	93.75%	93.61%	93.78%
0.5	0.622	89.21%	88.22%	89.79%	94.16%	93.55%	94.38%
1	0.625	88.29%	87.18%	89.13%	93.49%	92.44%	94.27%
5	0.623	88.14%	86.45%	89.27%	93.41%	92.20%	94.22%
10	0.623	88.97%	87.80%	89.71%	94.00%	93.33%	94.37%

Table 4. Estimates of *STITCH* error rates for the 30 sites with corresponding KASP genotypes. The error rate for each site was calculated for all samples, and for samples with greater >5x coverage. Estimates are given for the 6 possible classes of error between homozygous and heterozygous genotypes. REF and ALT are homozygous genotypes for the reference and alternative alleles, respectively. The linkage group, physical position in bp, total error for each site, and mean error across all sites are also provided.

Linkage Group & Position		Error rate for all samples (*n=1070*) (%)							Error rate for samples with >5x coverage (*n=180*) (%)						
		REF ↓ ALT	ALT ↓ REF	HET ↓ REF	HET ↓ ALT	REF ↓ HET	ALT ↓ HET	**Total**	REF ↓ ALT	ALT ↓ REF	HET ↓ REF	HET ↓ ALT	REF ↓ HET	ALT ↓ HET	**Total**
1	955185	0	0.21	1.14	2.8	2.28	4.24	**10.66**	0	0	0	0	0	0	**0**
1	45535107	0	0.1	1	1.3	1.8	2.7	**6.9**	0	0	0	0	0.58	0	**0.58**
1	47381180	0	0.1	8.06	0.91	2.32	4.93	**16.31**	0	0	3.64	1.21	0.61	1.82	**7.27**
1	51252783	0	0	0.6	0.4	0.8	1.99	**3.78**	0	0	0.58	0	0.58	0.58	**1.73**
1	60537705	0.2	0	3.6	2.2	6.29	2	**14.29**	0	0	2.91	0.58	4.07	2.91	**10.47**
1	71126303	0.1	0.2	2.11	1.31	0.7	3.32	**7.75**	0	0	0	0	0	2.92	**2.92**
2	55473945	0.1	0	1.92	1.52	1.21	0.71	**5.45**	0	0	1.18	0	0.59	0	**1.78**
2	58751341	0	0.1	13.8	0.51	3.94	9.49	**27.88**	0	0	14.79	0	2.96	8.28	**26.04**
3	1055959	0.9	0.3	5.29	7.49	9.48	13.17	**36.63**	1.12	0.58	2.89	7.51	6.94	6.36	**25.43**
3	3488331	0.1	0	1.61	5.12	4.02	3.31	**14.16**	0	0	3.55	0.59	1.18	2.37	**7.69**
3	44197469	0	0	2.01	1.71	2.01	1.81	**7.55**	0	0	0	0.59	0.59	0	**1.18**
4	7115521	0	0	1.01	1.61	1.51	1.31	**5.43**	0	0	0	0	0	0.6	**0.6**
4	44460308	0.1	0	1.29	2.18	2.98	0.89	**7.45**	0	0	0	0.59	1.18	0	**1.76**
4	50972302	0	0	1.8	2.1	2.9	3.8	**10.59**	0	0	1.17	0.58	0.58	0.58	**2.92**
6	15125731	0	0	0.31	0.31	1.33	0.92	**2.87**	0	0	0	0	0	0.59	**0.59**
6	53090212	0.1	0.2	5.01	1.5	3.01	7.11	**16.93**	0	0.58	4.09	0.58	1.17	4.09	**10.53**
6	53094790	0	0	1	0.3	1	2.49	**4.79**	0	0	0.58	0	0	0	**0.58**
6	53104387	0	0.21	2.41	1.05	2.31	3.56	**9.54**	0	0	1.2	0.6	2.41	0.6	**4.82**
6	53195127	0.7	0	0.8	10.54	4.52	6.12	**22.69**	0	0	0.6	4.22	4.22	4.82	**13.86**
6	53944185	0	0	1.39	0.3	1.1	0.3	**3.09**	0	0	0	0	0	0	**0**
6	32503688	0	0.1	0.6	0.8	1.1	0.8	**3.41**	0	0	0	0	1.18	0	**1.18**
6	52922047	0.4	0.2	5.89	4.29	5.89	5.89	**22.55**	0	0	1.75	2.34	5.26	1.17	**10.53**
6	52966811	0.1	0.1	3.12	0.91	3.93	4.33	**12.49**	0	0	2.4	0	2.4	4.79	**9.58**
6	52999112	0	0	0.79	0.69	0.99	0.89	**3.37**	0	0	0.58	0.58	0	0.58	**1.74**
6	53016551	0	0	1.83	0.51	0.91	1.83	**5.07**	0	0	1.18	0.59	2.94	1.18	**5.88**
6	53061080	0.2	0.1	4.88	1.99	5.07	2.99	**15.22**	0	0	0.58	1.17	1.75	1.17	**4.68**
6	53083229	0	0	4.07	0.41	1.83	2.44	**8.74**	0	0	1.22	0.61	2.44	3.66	**7.93**
7	49828363	0	0.5	6.7	0.5	1.4	3.4	**12.5**	0	0	2.38	0	1.19	1.19	**4.76**
8	27807304	0	0	2.58	0.69	1.19	1.68	**6.14**	0	0	0	0.58	0	1.16	**1.74**
8	38142268	0.1	0.5	6.65	1.59	2.68	6.55	**18.07**	0	0	4.62	0.58	0.58	5.2	**10.98**
Mean		0.1	0.09	3.11	1.91	2.68	3.5	11.41	0.04	0.03	1.73	0.78	1.51	1.88	5.99
SD		0.21	0.14	2.93	2.26	2	2.84	8.05	0.21	0.14	2.83	1.53	1.75	2.2	6.69

Table 5. Comparison of STITCH performance compared with hard-calling with bcftools. Accuracy, defined at the percentage of matching genotypes, was assessed for both methods by comparing genotype calls to corresponding KASP genotypes. We calculated accuracy for datasets that differ in their level of sequencing coverage, including all samples, those with >5x coverage, and those with greater than 5x coverage. We also estimated the error for the hard-called genotypes using three different per-genotype coverage cutoffs (i.e., the minimum number of reads needed to include a genotype in the error estimation): (i) no cutoff, where genotypes supported by any number of reads were included, (ii) only genotypes supported by at least 6 reads, and (iii) genotypes supported by at least 15 reads. The amount of missing data associated with each dataset is also provided; note that for the *STITCH* calls, there is no missing data.

	All samples *(n=1070)*				**Samples with >5x coverage *(n=180)***				**Samples with >15x coverage *(n=15)***			
	STITCH	***bcftools***			***STITCH***	***bcftools***			***STITCH***	***bcftools***		
per-genotype Depth	all	all	≥ 6	≥ 10	all	all	≥ 6	≥ 10	all	all	≥ 6	≥ 10
mean Accuracy (%)	88.59	73.59	95.44	96.94	94.01	88.29	95.58	97.06	95.93	95.92	96.78	97.03
mean missing (%)	0.00	62.69	93.37	98.45	0.00	2.74	63.63	90.80	0.00	4.08	11.33	36.67

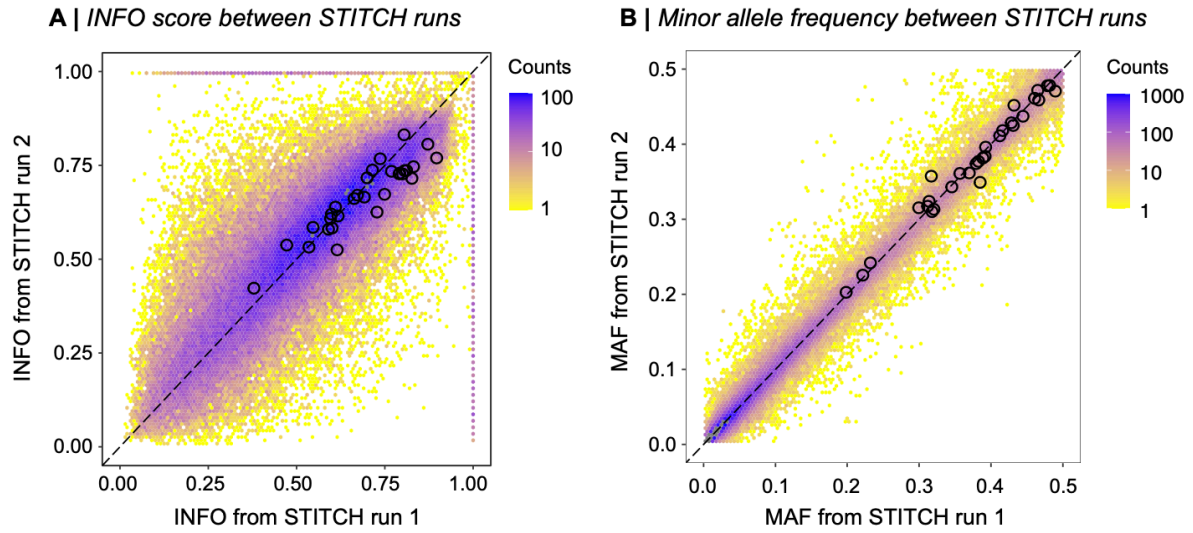

Figure 3. Consistency of site-level statistics between independent *STITCH* runs. **(A)** Joint distribution of the *INFO_SCORE* between runs 1 and 2. **(B)** Joint distribution of the minor allele frequency (MAF) between runs 1 and 2. Each point is coloured according to the density of sites (counts) that are present in that area of the plot.

calling has an accuracy 15% greater than hard-calling across all samples, and while providing more than twice as much data.

4.6.4 | Consistency of site-level statistics between independent STITCH runs

Satisfied that STITCH provides results that are reasonably accurate for our dataset, we conducted a second independent run of STITCH with the same set of parameters to quantify

the consistency between independent STITCH runs. Specifically, we repeated STITCH runs for the 30, 100Kb regions, allowing us to not only compare the accuracy for the 30 sites with corresponding KASP genotypes, but also compare the other metrics for the other 142,836 variable sites within these regions.

We first examined the consistency of site-level statistics returned by STITCH. We found a moderate relationship between the INFO obtained between the two runs (Spearman's *rho* = 0.54) indicating that INFO_SCORE is similar, but that the expected imputation accuracy differs somewhat between independent runs (Fig 3A). We found a much stronger relationship for the minor allele frequency (MAF) (Spearman's rho = 0.99), with the inferred MAF being highly consistent between the two STITCH runs (Fig 3B).

4.6.5 | Consistency of STITCH accuracy and genotype calls between independent STITCH runs

We next compared the consistency of the *STITCH* calls, by both comparing them to the imputed genotypes for the KASP markers, but also by comparing the genotypes from first *STITCH* run (Run 1) to the second *STITCH* run (Run 2). Focusing first on the comparisons with the *KASP* genotypes, we found that the estimates of accuracy were highly similar between the runs (Spearman's *rho* = 0.81 for all samples and 0.82 for samples with >5x coverage), such that a site with higher error in run1 also tended to have a higher error in run 2 (Fig 4A).

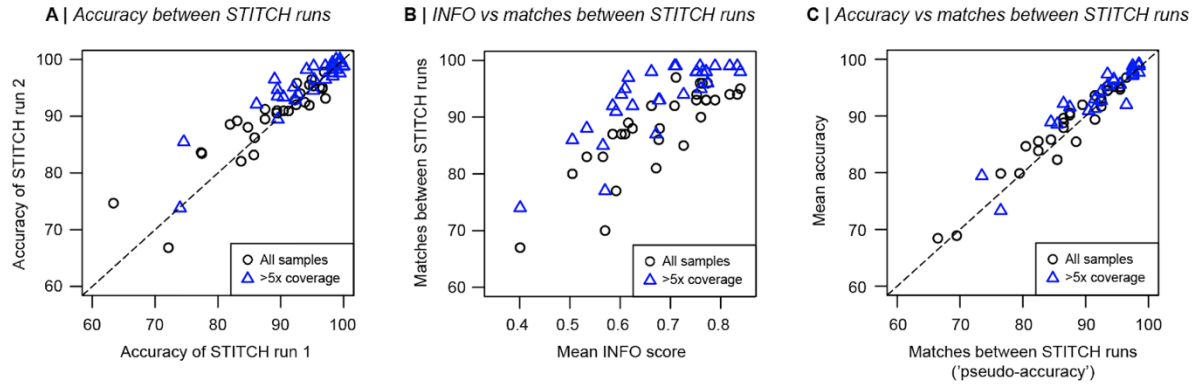

Figure 4. Consistency of genotype calls between independent *STITCH* runs. (A) The relationship between the accuracy calculated for the two *STITCH* runs. (**B)** The relationship between the percentage of genotype matches between the two runs and the mean *INFO_SCORE*. **(C)** The relationship between the mean accuracy across *STITCH* runs 1 and 2 and the percentage of matches between the two *STITCH* runs (subsequently referred to as 'pseudo-accuracy'). The dashed line in plots ***A*** and ***C*** is the 1:1 line.

Although the accuracy of the two runs was highly similar, the comparison of the genotypes from the two runs shows that the errors made by *STITCH* are not consistent between the independent runs. Specifically, we found that the percentage of genotype matches varied widely across the markers, from 63% to 97% (mean 11.4%) for all samples and 74% to 100% (mean 6.3%) for samples with >5x coverage. We found that the percentage of matches between the runs tended to increase as the mean INFO score between runs 1 and 2

increased (Spearman's *rho* = 0.63 for all samples, 0.60 for samples with >5x coverage) (Fig 4B). This suggests that while the error rate for each marker is fairly consistent between independent runs, the actual errors made by *STITCH* are somewhat stochastic and can be predicted partly based on the *INFO_SCORE*.

We next wanted to understand whether the percentage of matches between the *STITCH* runs could be used to predict our measure of accuracy, defined as the percentage of matches between the *STITCH* and *KASP* genotypes. Indeed, we found a very strong positive relationship between accuracy and the percentage of genotype matches for the 30 markers (Spearman's *rho* calculated between the % *STITCH* matches and mean accuracy between *STITCH* runs 1 and 2: 0.94 for all samples, 0.88 for samples with >5x coverage) (Fig 4C). This suggests that the percentage of matches between independent *STITCH* runs is a better predictor of accuracy than the *INFO_SCORE*.

4.6.6 | Estimating genome-wide distributions of accuracy

We next used the percentage of matches between the independent *STITCH* runs—hereafter referred to as pseudo-accuracy—to contain a more general picture of how accuracy varies across the genome. Unlike our measure of accuracy, which could only be calculated for the 30 markers with high-accuracy *KASP* genotypes, pseudo-accuracy can be estimated for all of the variable sites that were included in both *STITCH* runs.

Figure 5 shows the relationships between pseudo-accuracy, the *INFO_SCORE*, and the minor allele frequency for the 142,836 variable sites. This clearly shows that sites with the lowest pseudo-accuracy tend to be those with a higher minor allele frequency and an intermediate *INFO_SCORE*. There are two fairly distinct areas of high accuracy within the 3-D space. First, sites with a lower MAF tend to have a higher accuracy regardless of the *INFO_SCORE*. This makes intuitive sense, because accuracy is a site-level measure of the percentage of non-matching genotypes at a site. Thus, we expect the lower limit of the accuracy to be a function of the frequency of the minor allele (*i.e,* the maximum error attainable under a given MAF should be 2*n* where *n* is the number of copies of the minor allele). More specifically, if the minor allele is very rare, the error rate will be very low for a site, even if all genotypes that include the minor allele have been miscalled. The other area of high accuracy is associated with a high MAF and a high INFO score. Indeed, there is a clear positive relationship between the INFO score and accuracy when the minor allele frequency is large.

In summary, we find that the genome-wide accuracy of the *STITCH* calls is likely quite high. For example, the mean site pseudo-accuracy is 95% for all samples (95.9% for samples with >5x coverage), with more than 83.9% of sites having an accuracy greater than 90%. However, our analysis shows that the factors driving site-level accuracy vary across loci. For sites where the minor allele is very rare, we expect the accuracy to be high, because most individuals will be correctly called as homozygous for the common allele even if all genotypes involving the minor allele are miss-called. When the minor allele is common, there is more

potential for genotypes to be miscalled, in which case, the *INFO_SCORE* predicts accuracy at the site level.

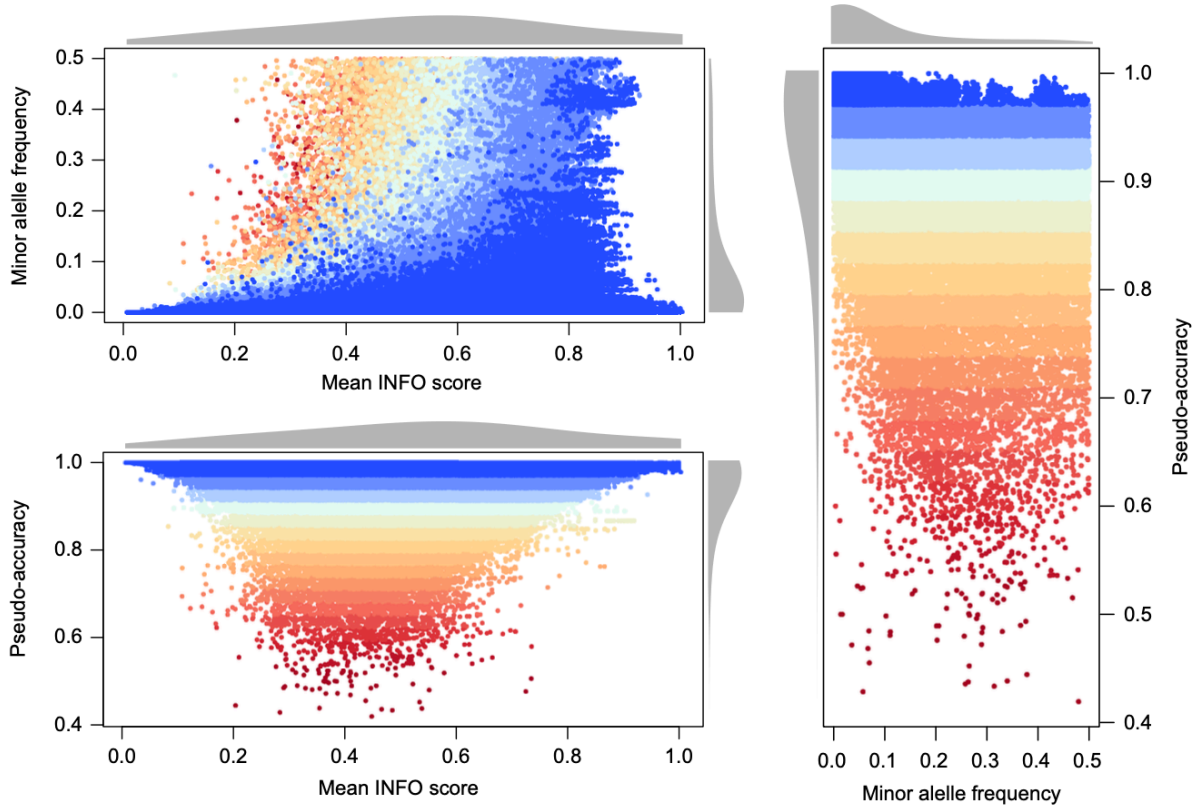

Figure 5. The relationships between the minor allele frequency, mean *INFO_SCORE* and pseudo-accuracy 142,836 variable sites. The three plots show the joint distributions for each pair of variables. The grey curves on each axis show the 1-D density for individual variables. Points are coloured by accuracy. Bluer sites have a higher pseudo-accuracy and redder sites have a lower pseud-accuracy.

4.6.7 | Final STITCH run

To obtain calls for the whole genome, we used the final set of parameter values (*--K=75, –-ngen=100, --niter=40, --downsampleToCov=20, --expRate=0.5cM*) to perform *STITCH* imputation across all 28,956,743 putative variant positions. For parallelisation, STITCH was performed on 1Mb genomic regions with an overlap of 100kb on each end. Out of all preliminary putative variant positions, 6,861,000 (23.69%) were inferred to be invariant (*i.e.,* the initial calls by *bcftools* and *STITCH* calls disagree) and were, therefore, removed from further analysis (Table 6). Furthermore, we removed sites with *INFO_SCORE* < 0.2 and retained a final 20,043,334 bi-alleleic SNPs.

4.7 | Estimation of recombination rates

We next used *LDhat v2.2* (Auton and McVean, 2007) to estimate the population-scaled recombination rate (ρ) between each SNP. For each discrete site on the magenta (MF1/2/3) and yellow flank (YF1/2/3) (Fig 1), we first used the *lkgen* function in *LDhat* to to generate a

log-likelihood lookup table, tailored to the number of haplotypes in each deme and with ϑ=0.009 (calculated from the average genome-wide nucleotide diversity in 10Kb windows, based on Field et al., 2025). Variable recombination rates were then inferred using the *interval* function with the parameters: *--its 10000000 --samp 5000 --bpen 5* (i.e., 10 million MCMC iterations sampled every 5000 iterations and a block penalty parameter of 5 as recommended by the authors of *LDhat* in case of no obvious/known recombination hotspots). Results from the MCMC iterations were summarised with the *stat* function, discarding a burn-in of the first 1000 iterations, to estimate the mean ρ between each SNP pair. To ensure computation efficiency, *LDhat* analyses were performed on sliding windows of 2000 SNPs with an 100 SNPs at each window boundary, subsequently combined to produce continuous recombination rate estimates for each chromosome.

Table 6: Number and characteristics of variants after final *STITCH* calling. We provided *STITCH* with a list of ~29 million putative variant sites, which determined that 6.8M were most likely to be invariant. Thus, our final dataset consisted of 22M variant positions that were used in subsequent analyses. The number and percentage of variants associated with different INFO score classes are shown.

Dataset		#	%
Putative Variant Sites		28,956,743	–
Invariant Sites after imputation		6,861,000	23.69
Final Variant Sites	*All Linkage Groups*	22,095,743	–
	info = 1	930,718	4.21
	info ≥ 0.8	3,110,828	14.08
	info ≥ 0.6	8,440,477	38.20
	info ≥ 0.4	14,647,305	66.29
	info ≥ 0.2	20,043,334	90.71

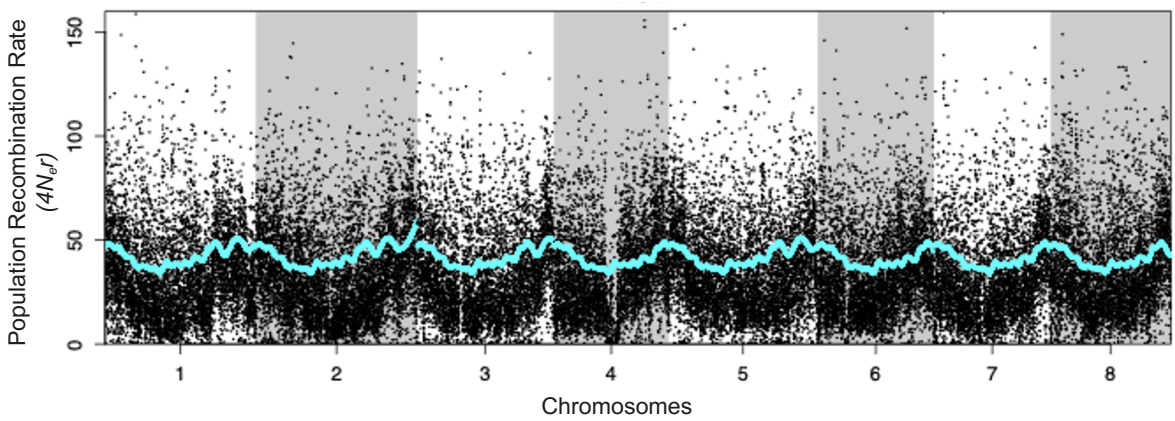

Figure 6. Population-scaled recombination rate ($4N_er$) inferred by *LDhat* along the genome. The recombination landscape is qualitatively similar to other estimates derived in Pal et al., 2025.

Population recombination rate (ρ) estimates for each deme were highly correlated (Spearman's *rho* = 0.8 – 0.83), and therefore averaged to produce a single estimate between each SNP, used in all downstream analyses. We converted ρ as $\rho = 4N_e r$, where N_e = 406694 (calculated as $\pi = 4N_e\mu$ where μ = 5.7x10-9 which is the estimated mutation rate for *Antirrhinum hispanicum*, Zhu et al., 2023; and π = 0.009 ,Field et al., 2025), and *r* is the number of crossovers per basepair per generation, that was subsequently used to represent the position of each SNP in centimorgans (cM) to produce the final recombination map. The final map length of the whole genome was estimated to be 299 cM (Fig 6).

4.8 | Hybrid strategy to phase variants using molecular and statistical phasing

4.8.1 | Molecular Phasing

We performed molecular phasing on the 20 million imputed SNPs with *HAPCUT2* (Edge et al., 2017). In contrast to statistical phasing, which infers haplotype phase from population-level genotype data, molecular phasing directly determines phase at each heterozygous site by leveraging sequencing reads that originate from the same DNA molecule, as identified by shared molecular barcodes.

For each sample, we first extracted all reads covering heterozygous sites from the BAM files using the *extractHAIRS* script provided by *HAPCUT2*. Subsequently, we grouped reads sharing the same BX-tag barcode, allowing a maximum gap of 50Kb between reads, using the *LinkFragments.py* script. Phasing was then performed in the 10X/linked-read aware mode of *HAPCUT2*, with a phase-block PHRED-scaled threshold of 10, and the call-homozygous option enabled to improve confidence in heterozygous variant calls. The resulting linked-read fragments defined phase blocks, each containing a set of heterozygous sites. Only a subset of these sites were phase-informative (i.e., covered by one or more reads) and thus could be phased by *HAPCUT2* (Fig. 7A).

Sequencing coverage had a pronounced effect on the number and length of phase blocks, as well as the fraction of SNPs that could be phased. Samples with higher coverage had fewer phaseblocks, but a higher fraction of SNPs that were phased (Fig 7A). In samples with ≥5x coverage (hereafter, high-coverage), we observed an average of 4,110 phase blocks per sample, with 91.2% of SNPs phased. In contrast, samples with <5x coverage (hereafter, low-coverage) exhibited an average of 19,316 phase blocks per sample, with only 48.2% of SNPs phased. More importantly, in high-coverage samples, 68.9% of phased SNPs were, on average, located within the 10 largest phase blocks (largest by number of phase-informative SNPs), whereas in low-coverage samples, it was only 4.3% (Fig 7A). For example, the sample with the highest coverage (111.4x) had 98.4% of SNPs phased, distributed across 2,268 phase blocks, with the 10 largest blocks spanning 50.5 Mb of chromosome 6 and containing 93.8% of phased SNPs (Sample A, Fig 7B). In contrast, a sample with 2.5x coverage had 18,923 phase

blocks, with only 57.4% of SNPs phased; the 10 largest blocks spanned just 2.7 Mb and contained a meagre 7.8% of phased SNPs (Sample B, Fig 7B).

While SNPs associated with 1 phaseblock are, in theory, expected to have the most accurate phase information informed directly by the molecular barcodes, the higher order phase between two phase blocks is often unknown and cannot be deduced directly from reads or molecules. Our findings indicate that most phase-informative SNPs in high-coverage samples are concentrated in a small number of long phase blocks, providing more contiguous and accurate phase information over long distances. In contrast, low-coverage samples yield shorter, more fragmented phase blocks, increasing the potential for phase switch errors. For example, 93.8% of phased SNPs in the sample with 111.4x coverage, are distributed within 10 phase blocks (Sample A, Fig 7), with only 9 possible phase switch errors between the phase blocks. In contrast, for the sample with 2.5x coverage, the same fraction of phased SNPs is distributed over 12k phaseblocks with an average span of only 4.7kb, thus vastly increasing the phase switch error rates (Sample B, Fig 7). Based on these findings, we retained molecular phasing results only for high-coverage samples and applied statistical phasing to the rest of the samples.

4.8.2 | Comparison between statistical phasing, molecular phasing and the use of reference panels

Due to the absence of an independent haplotype reference panel for *Antirrhinum majus*, direct external validation of phase accuracy was not possible. To benchmark statistical phasing, we compared it to molecular phasing in a 1Mb region of chromosome 6 (52.5–53.5 Mb) containing 66K SNPs. Statistical phasing was performed using *SHAPEIT5* (Hofmeister et al., 2023), both with and without a reference panel.

To ensure the most accurate phase information, we constructed a reference panel from 123 samples with ≥5x coverage, each having a single phase block spanning the entire target 1Mb region (chr6:52.5–53.5Mb). Since *HAPCUT2* does not allow for missing genotypes in its reference panel, we identified 22K phased SNPs associated with the phase block and common to all 123 samples that finally served as the reference panel. Statistical phasing of the full set of 66,000 SNPs was conducted using the *phase_common_static* function in *SHAPEIT5*, with and without the *--scaffold option* (reference panel), using the recombination map from *LDhat* and an effective population size (N_e) of 406,694 (calculated as $\pi=4N_e\mu$, with $\mu = 5.7\times10^{-9}$ and π from PoolSeq data; Field et al., 2025).

Switch error rates (SER) were assessed by comparing statistical phase calls to molecular phase calls at the 22K SNPs in the reference panel, treating molecular phase as ground truth. We found a slight decrease in SER for each of the 123 samples when the reference panel was used for statistical phasing; with a reduction from an average SER of 6.62% of 5.52% (Fig 8A). SER also decreased with increasing minor allele frequency (MAF) (Fig 8B), consistent with expectations that statistical phasing is more accurate for common haplotypes. Although rare haplotypes absent from the reference panel can still lead to switch

errors, overall SER was reduced in all individuals and across most MAF bins when the reference panel was included.

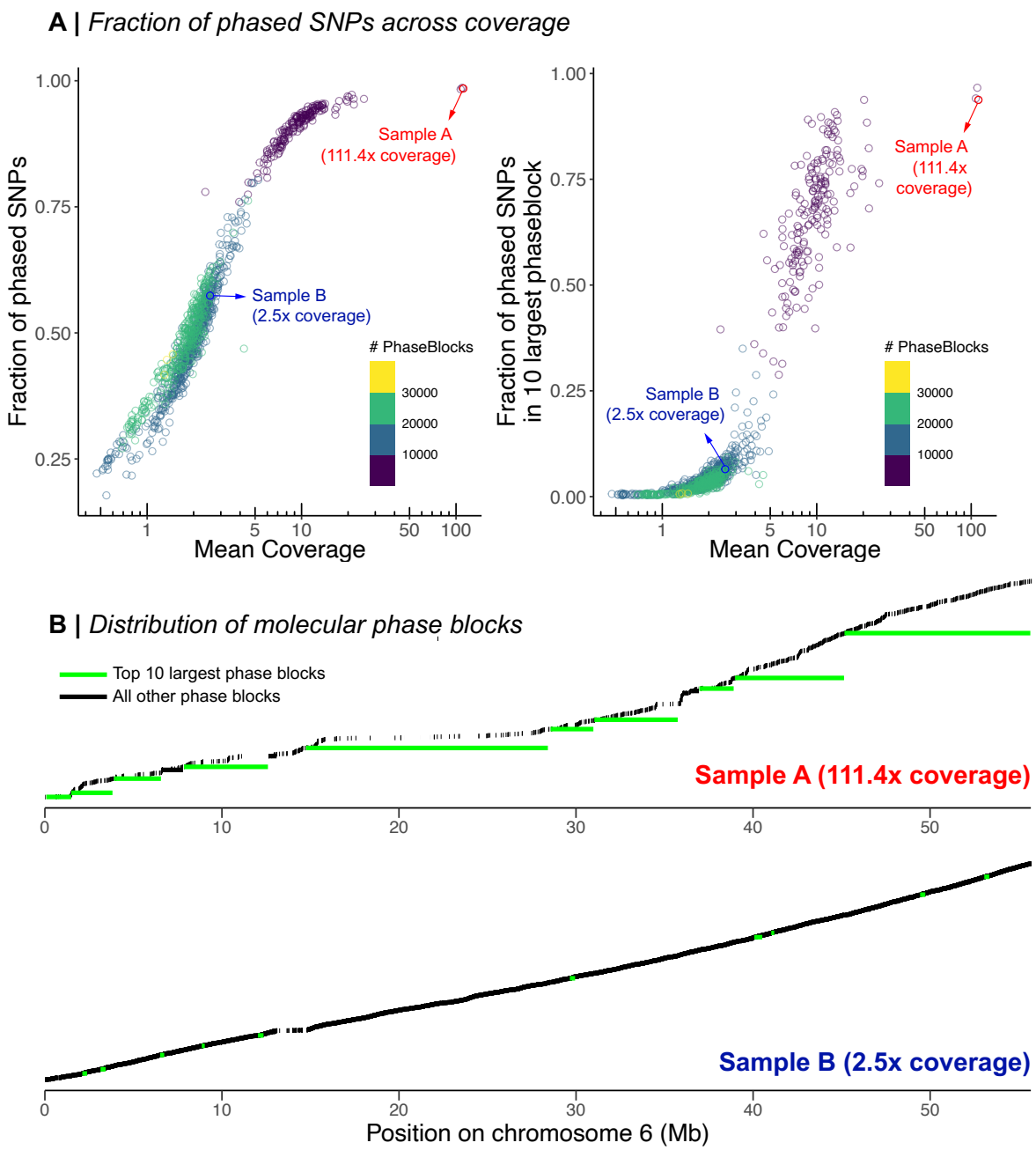

Figure 7. Performance of molecular phasing. (A) Fraction of SNPs that were phased by molecular phasing. Higher coverage contained more phase-informative SNPs associated by lesser number of phase blocks. Proportion of phased SNPs in the 10 largest phaseblocks also increased with mean sequence coverage. Two samples, A (marked in red) and B (marked in blue) are chosen to show the qualitative differences of phasing between higher and lower coverage samples. **(B)** Distribution of phaseblocks for samples A and B are shown. Each horizontal line is a phaseblock, showing their span along the chromosome 6. The 10 largest phaseblocks are marked in green. Phaseblocks of Sample A span much longer than those of Sample B, and contains 93.8% of the phased SNPs within the 10 largest phaseblocks compared to 7.8% in Sample B.

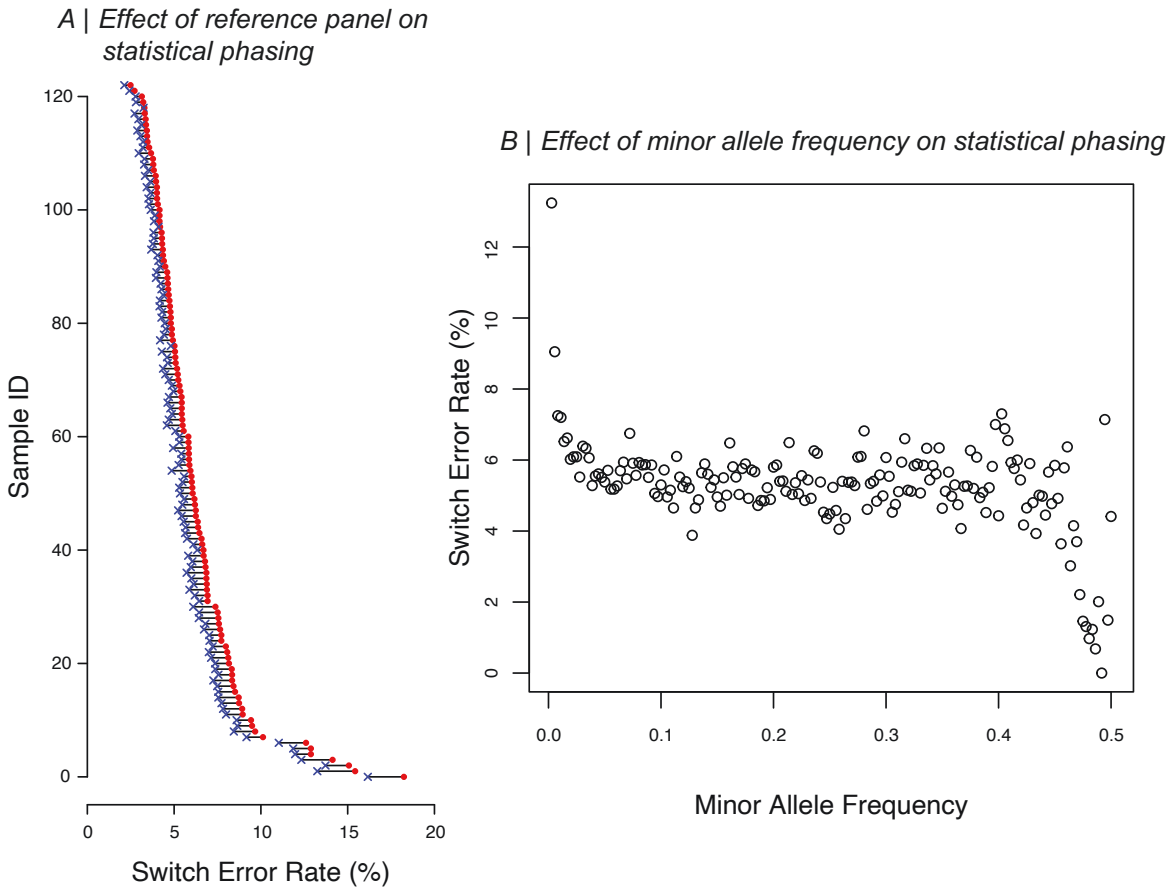

Figure 8. Performance of statistical phasing. (A) Switch Error Rate (SER) improved when using molecular phase blocks as reference panel (blue crosses) compared to without (red dots). Molecular phase was used as the ground truth for comparison to statistical phase. **(B)** SER depended on the minor allele frequency of the allele just after the switch error, and performed worse for rarer alleles.

To further assess the impact of phasing accuracy on downstream analyses, we inferred genealogies (tree sequences) using *tsinfer* (Kelleher et al., 2019) for a subset of 56 high-coverage samples with 38K phased SNPs common to all and within a single phaseblock (see Chapters 5 for details on tree inference). The intuition is that – if statistical phasing produces a large number of short phase blocks where the higher order of phase is not known, we might expect to see more marginal trees in a statistically phased region over a molecular phased region. So, we used a subset of 56 samples of the highest coverage that had an intersection of 38K SNPs all associated with 1 phase block. We compared genealogies inferred on the phased genotype called from (*i*) molecular phasing, (*ii*) statistical phasing without a reference panel, and (*iii*) statistical phasing with a reference panel. Although the number of marginal trees differed between methods, 86% of marginal tree boundaries overlapped between statistical and molecular phasing, indicating high concordance between approaches. Based on these findings, we took a hybrid approach of leveraging the advantages of molecular and statistical phasing to phase our full dataset with 20 million SNPs.

4.8.3 | Final statistical phasing with molecular phased SNPs as reference panel

Finally, all 1,074 samples were statistically phased using *SHAPEIT5* with the *phase_common_static* function, the *--scaffold* option (reference panel), the recombination map from *LDhat*, and N_e=406,694. The reference panel comprised 180 samples with ≥5x coverage and the intersection of 8,998,583 SNPs molecular-phased with *HAPCUT2*.

4.9 | Polarising alleles as ancestral or derived

We polarised alleles of the 20 million phased SNPs in *A. majus* as ancestral or derived using high-coverage PoolSeq sequence data (mean coverage = 89.97x) from a closely related outgroup species *Antirrhinum molle*. For each bi-allellic site in *A. majus*, we determined which alleles were present within *A. molle*. For 86.1% of sites, we observed 3 site-patterns that allowed us to identify which allele is derived in *A. majus* (Table 7).

Pattern A: *14,177,182 sites (64.16%) resolved*

Sites where one of the *A. majus* alleles is fixed in *A. molle*. Out of these 14.2 million sites that constituted 64.2% of all the bi-allelic *A. majus* sites, 12.9 million (58.6%) sites had the *A. majus* major allele fixed in *A. molle*, whereas the remaining 1.2 million (5.6%) sites had the *A. majus* minor allele fixed in *A. molle*. For all such sites, the fixed *A. molle* allele was assigned ancestral, while the other *A. majus* allele derived.

Pattern B: *3,921,661 sites (17.75%) resolved*

Sites where both *A. majus* alleles are present in *A. molle*, and both species share the same major allele, i.e., allele frequency ≥ 0.5. For such sites, the major allele was assigned ancestral, since the minor allele in either taxa would most likely be derived due to new mutations being relatively rare.

Pattern C: *921,886 (4.17%) resolved*

Sites where both species are polymorphic, but only share 1 allele. In this case, the shared allele is considered ancestral, while the unique allele is considered derived.

The remaining 13.9% of bi-allelic sites in *A. majus* had no logical basis for assigning ancestral or derived alleles based on the information from *A. molle* (Patterns D–F in Table 7). The most unresolvable site-pattern (Pattern D: 1,631,993 or 7.38% of all bi-alleleic sites), involved both species sharing the same 2 alleles, but the minor allele in one species was the major allele in the other. For 1,391,039 sites (6.3%**)**, there was simply no sequence information available for *A. molle* (Pattern E in Table 7). Finally for the remaining 52,338 sites (0.237%), neither of the *A. majus* alleles was present in *A. molle* (Pattern F in Table 7), suggesting both alleles being derived in an unknown order in the ancestor to *A. majus*. For all

the 3,075,714 unresolved bi-alleleic sites (13.9%), we assumed the minor allele (< 0.5 allele frequency) in *A. majus* to be the most likely derived allele. In addition, latest genealogical inference (where allele polarisation is of utmost importance) tools such as *Relate*, can alter the polarisation of alleles based on incompatibilities of mutations on genealogical tree branches.

Table 7. Overview of number of sites where *A. majus* alleles could be polarised into ancestral or derived, based on alleles present in *Antirrhinum molle*. For unresolved sites, major allele in *A.majus* was considered ancestral.

	Total sites		Site-patterns			Description of sites	Ancestral Allele
	#	%	Type	#	%		
Sites resolved	19,020,029	86.08	A	14,177,182	64.16	One of the *A. majus* alleles is FIXED in *A. molle*	Fixed allele
			B	3,921,661	17.75	Both *A. majus* alleles present in *A. molle*. Both species have the same major allele (frequency ≥ 0.5)	Major allele
			C	921,186	4.17	Both *A. majus* and *A. molle* are polymorphic, but only share 1 allele.	Shared allele
Sites NOT resolved	3,075,714	13.92	D	1,632,337	7.39	Both *A. majus* alleles present in *A. molle*. Both species do **not** have the same major allele (frequency ≥ 0.5)	–
			E	1,391,039	6.30	No genotype information for *A. molle*	–
			F	52,338	0.24	No shared allele between *A.majus* and *A.molle*	–

4.10 | Conclusions

Producing a genomic resource at scale with today's state-of-the-art methods can be somewhat inexpensive, but never error-free. Specifically, most errors arise from genotype imputation and haplotype phasing, that depends on several factors markedly different between model and non-model organisms—mainly, presence of a reference-panel or pedigree information and its size, genome complexity, diversity along the genome, and finally, sequence quality. This chapter set out to provide the first population-scale, phased whole genome dataset for the Antirrhinum majus hybrid zone and, in doing so, explore a possible

methodological pipeline for future datasets at such scale even beyond our study system. We achieved this, by cost-effectively linked-read-sequencing 1084 plants to low- and high-coverage, hoping to identify variants segregating in the population rather than an individual, that can then inform imputation and phasing of sites missing in the lower coverage samples.

The raw data yielded 429 million putative variant sites. A series of depth, quality, multiallelic, and indel-proximity filters reduced this set to 29 million high-confidence biallelic SNPs that were then used for genotype calling and imputation with *STITCH*. By benchmarking *STITCH* imputed genotype calls at 30 variant sites against "gold-standard" KASP genotypes, we identified a configuration of imputation parameters that maximises the joint objectives of completeness as well as accuracy in our dataset dominated by low coverage samples (~80%). This delivered a mean accuracy of 89 across all samples, and 94% for the samples with ≥5x coverage. While imputation accuracy of <1% is routinely reported in model organisms (humans, cattle, *Arabidopsis*) (Arouisse et al., 2020; Lee et al., 2021; Stahl et al., 2021), such accurate genotype imputation is almost unheard of in non-model systems, mainly because of the lack of external genomic resources, such as a reference panel or pedigree, and long-range information in sequencing reads. Our estimates match those of other non-model systems (~3–12%) (Money et al., 2015; Watowich et al., 2025). It is worth mentioning here that most studies do not report or estimate the individualized genotype-imputation accuracy due to the lack of any external validation. Generally, it is more common to check for correlations between expected and observed minor allele frequency bins. We could benchmark the accuracy of individualised imputation calls, owing to our concerted long-term sampling of the hybrid zone and available KASP sequences for every individual sampled. While we find very high correlation between the expected and observed MAF bins (Spearman's *rho* = ~0.99), our individualised accuracy estimates suggest mismatch between genotype calls although the over frequency is preserved. Therefore, we hope these values would provide guidelines for future studies in other systems that lack external genotype information to compare with, and inform their sampling and sequencing design.

The second accuracy bottleneck was phasing. Owing to the unique barcode information, we devised a hybrid scheme where we molecular-phased 180 samples with ≥5x coverage to be then used as a reference panel for statistically phasing all samples. Since we do not have independent sequence information for the haplotagged samples, we could not compute an absolute switch error rate. But comparing the molecular phase to statistical phase with the reference panel, we found switch errors to occur at a rate of 5.5%, which is expectedly far higher than model systems (~1%) (Hofmeister et al., 2023), but match non-model study systems (~3–15%) (Majidian and Sedlazeck, 2020).

Overall, a point of caution merits explicit acknowledgement. The deliberate inclusion of very low-coverage individuals enhances population representation but depresses imputation and phasing accuracy for samples with <5x coverage. This dictates caution in analyses that rely on long-range phase information, such as ARG inference, estimation of allele ages, etc. Therefore, analyses requiring per-site precision should either weigh genotypes by their imputation confidence (*INFO_SCORE*), restrict to the ≥5x coverage subset

or use the most accurate phase information from only molecular phaseblocks. Moreover, future method developments should also ideally incorporate locus-specific uncertainty into their models. On the other hand, analyses such as GWAS are less likely to be affected by a handful of genotype errors, given our large sample size.

Despite these limitations, the resulting catalogue of 20 million phased SNPs constitutes an unprecedented resource for the *Antirrhinum* system. *When applied to phenotypically structured systems, these workflows clarify transitions across hybrid zones.* The next chapters leverage this dataset to first study the molecular parallelism between two hybrid zones, then combines mixed-model GWAS with classic population-genetic outlier statistics to dissect the genetic architecture of flower colour and divergence, and finally, reports a preliminary analysis of genealogical history of a selected allele that controls flower colour variation. Together, this dataset enables us to study the *Antirrhinum* hybrid zone and understand how selection and gene flow shapes genomic landscapes during early stages of speciation.

4.11 | Supplementary Information

Supplementary information related to this chapter is in Appendix C.

Chapter 5
Genealogical analysis of replicate flower colour hybrid zones in Antirrhinum†

Abstract

A major goal of speciation research is identifying loci that underpin barriers to gene flow. Population genomics takes a 'bottom-up' approach, scanning the genome for molecular signatures of processes that drive or maintain divergence. However, interpreting the 'genomic landscape' of speciation is complicated, because genome scans conflate multiple processes, most of which are not informative about gene flow. However, studying replicated population contrasts, including multiple incidences of secondary contact, can strengthen inferences. In this paper, we use linked-read sequencing (haplotagging), F_{ST} scans, and genealogical methods to characterise the genomic landscape associated with replicate hybrid zone formation. We studied two flower colour varieties of the common snapdragon, *Antirrhinum majus* subspecies *majus*, that form secondary hybrid zones in multiple independent valleys in the Pyrenees. Consistent with past work, we found very low differentiation at one well-studied zone (Planoles). However, at a second zone (Avallenet), we found stronger differentiation and greater heterogeneity, which we argue is due to differences in the amount of introgression following secondary contact. Topology weighting of genealogical trees identified loci where haplotype diversity was associated with the two snapdragon varieties. Two of the strongest associations were at previously identified flower colour loci: *Flavia*, that affects yellow pigmentation, and *Rosea/Eluta*, two linked loci that affect magenta pigmentation. Preliminary analysis of coalescence times provides additional evidence for selective sweeps at these loci and barriers to gene flow. Our study highlights the impact of demographic history on the differentiation landscape, emphasizing the need to distinguish between historical divergence and recent introgression.

† This chapter is published and can be found online at: https://doi.org/10.1111/mec.70067

5.1 | Introduction

A major goal of speciation research is to identify loci underlying barriers to gene flow. Population genomic studies usually take a 'bottom-up' approach by scanning the genome for patterns of within- and between-population variation that indicate selection driving or maintaining divergence (Ravinet et al., 2017; Wolf and Ellegren, 2017). For example, during speciation with gene flow, genomic regions associated with local adaptation or genetic incompatibilities are expected to show elevated genetic differentiation (usually measured by F_{ST}), with the rest of the genome homogenised through genetic exchange (Feder et al., 2012; Wu, 2001). Indeed, numerous studies of the 'genomic landscape' have found highly heterogenous patterns of genetic differentiation and, in some cases, have shown that regions with high F_{ST} house genes underpinning adaptive traits that also act as reproductive barriers (Hooper et al., 2024; Martin et al., 2013; Poelstra et al., 2014; Todesco et al., 2020). However, we now know that interpreting the differentiation landscape is more challenging than some researchers once hoped (Ravinet et al., 2017; Wolf and Ellegren, 2017).

The main challenge is that genome scans can conflate multiple processes, some of which are not due to current heterogenous gene flow (Ravinet et al., 2017). Consider a simple model of secondary contact, where divergence builds up over a long period (which may involve intermittent isolation), and erodes following recent contact. Genome-wide divergence builds up relatively slowly, due to both drift and selection. Divergence will inevitably be heterogenous along the genome both by chance and due to intrinsic properties of the genome, such as the local density of functional elements and local recombination rate (Burri, 2017). After contact, introgression will erode divergence where the populations meet, potentially revealing the location of barrier loci (Duranton et al., 2018). While relatively fast compared to build-up of divergence, this erosion takes some time, and will be delayed if interbreeding is geographically localised (Barton and Gale, 1993). Thus, genome scans reflect both initial divergence and post-contact introgression, and these may be hard to disentangle.

Inclusion of replicate hybrid zones aids the interpretation of genome scans, allowing comparison of divergence across multiple contacts (Nadeau et al., 2014; Rancilhac et al., 2024; Vijay et al., 2016; Wilding et al., 2001). Overall divergence may reflect differences in the timing of contact or rates of gene flow. Nevertheless, parallel contacts should ultimately lead to similar differentiation landscapes if large-effect outlier loci reflect barriers to gene flow that have resisted introgression in each location. In contrast, outliers found in a single zone might reflect local demographic processes (e.g., bottlenecks), evolutionary noise, sampling effects, or population specific barriers (Westram et al., 2021). Several studies have used this logic to identify loci that underpin local adaptation and speciation. The most compelling studies combine traditional site-based genome scans with tree-based methods, which make it possible to analyse more than two populations within a single framework that acknowledges their recent shared history (Poelstra et al., 2014; Rancilhac et al., 2024).

In this paper, we study genome-wide variation associated with replicate hybrid zones in the common snapdragon, *Antirrhinum majus*, a classic model for understanding phenotypic

variation both in the laboratory and in nature (Hudson et al., 2008). We focus on two varieties of *A. majus* subspecies *majus—A.m.m* var. *pseudomajus* and *A.m.m* var. *striatum* (hereafter, var. *pseudomajus* and var. *striatum* for brevity)—that are native to France and Spain (Whibley et al., 2006). These varieties have largely non-overlapping geographic distributions, occupy similar habitats and are pollinated by the same bee species (Tavares et al., 2018). The major difference between them is their contrasting flower colour: var. *pseudomajus* has magenta flowers with a small patch of yellow pigment on the face of the flower below the bee entry point, while var. *striatum* has yellow with restricted veins of magenta colouration above the bee entry point (Fig. 1A). These differences in colour, which are thought to be alternative adaptations to attract the same bee pollinators, are caused by a small number of loci that control the production of two flavonoid pigments in floral tissue, anthocyanin (magenta) and aurone (yellow). *Rosea*, *Eluta* (Tavares et al., 2018), *and Rubia* (Field et al., 2025) affect anthocyanin production, while *Sulfurea* (Bradley et al., 2017), *Flavia* (Bradley et al., 2025), *Cremosa* and *Aurina* (Richardson et al., 2025) affect aurone production.

During the last ice age, var. *pseudomajus* and var. *striatum* are thought to have been restricted to areas of low elevation, but subsequently expanded into the Spanish Pyrenees (Vargas et al., 2004; Whibley et al., 2006). As a result, at least three separate hybrid zones have formed in separate valleys below the altitudinal limit of *A. majus* (Fig. 1A). In one such zone near the town of Planoles, a transition from yellow to magenta flowers occurs over a few kilometres (Whibley et al., 2006). Scans of genome-wide sequence variation have revealed strong allele frequency differentiation and sharp geographic clines around previously identified colour loci (Field et al., 2025; Tavares et al., 2018). In contrast, most of the surrounding genome shows low genetic differentiation, probably owing to the homogenising effects of dispersal and recombination (Ringbauer et al., 2018; Tavares et al., 2018).

Here, we expand our analysis to include individuals from the Planoles hybrid zone and a second zone near the town of Avellanet, located over 50 km to the west (Fig. 1A). By comparing their respective genomic landscapes and jointly analysing two independent hybrid zones, we hoped to disentangle ancestral divergence from the effects of recent introgression. We were especially interested in whether known flower colour loci act similarly in both localities and stand out from their genomic background. To address this, we used both traditional F_{ST} scans, and genealogical methods for studying the genome-wide distributions of tree topologies and coalescence times across the genome. As a secondary aim, we use this study as an opportunity to compare different methods for inferring genealogical trees along the genome. Several approaches are now available for inferring genealogies from phased SNP datasets (Nielsen et al., 2024). However, these methods have not yet been widely used to study adaptation and speciation. It is also unclear how they perform when applied to real datasets, and there has been limited discussion about when more sophisticated methods might be warranted over simpler ones. We hope that this study helps other researchers decide which method might be most appropriate for their data and specific goals.

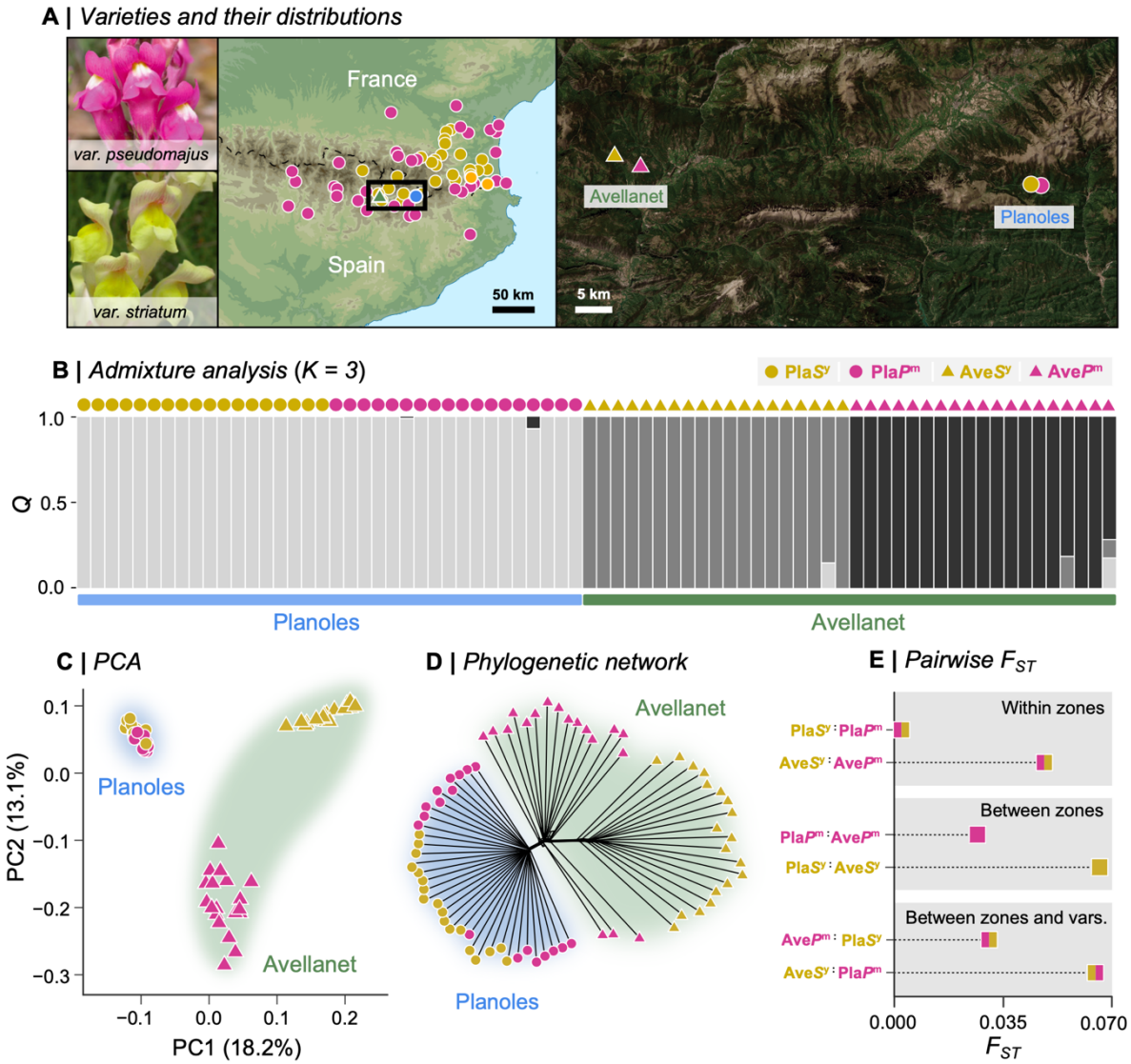

Figure 1. Evolutionary relationship between *A. majus* subspecies *majus* populations from two hybrid zones. (A) Geographic distributions of magenta flowered *A.m.m.* var. *pseudomajus* and yellow-flowered *A.m.m.* var. *striatum*. The map is based on sample locations in Whibley et al. (2006) and does not show the full distribution of either variant. Circles are coloured according to the population. Samples were collected from two hybrids zones: Avellanet (19 magenta & 19 yellow samples) and Planoles (18 magenta & 18 yellow samples). Points on the map represent the average location of each population. **(B)** Genetic structure, shown as admixture coefficient (*Q*) for *K*=3 clusters inferred by *Admixture* from 1.7 million LD-thinned SNPs. Each vertical bar is one individual. **(C)** The first two principal components of the same dataset. **(D)** Phylogenetic network (*neightbourNet*) on the same dataset. **(E)** Estimates of per-site Weir & Cockerham's F_{ST}, averaged over all 11.5 million SNPs. F_{ST} was calculated between varieties from the same hybrid zone (PlaP^m vs PlaS^y, AveP^m vs AveS^y), and between the hybrid zones (PlaP^m vs AveP^m, PlaS^y vs AveS^y, PlaS^y vs AveP^m, PlaP^m vs AveS^y). Pla: Planoles, Ave: Avellanet, P^m: magenta-coloured var. *pseudomajus, S*^{y}: yellow-coloured var. *striatum*.

5.2 | Results and Discussion

5.2.1 | Genome-wide analysis reveals different histories of post-contact gene flow across the two hybrid zones

We sampled 18 individuals of magenta-coloured var. *pseudomajus* and 18 yellow-coloured var. *striatum* from Planoles (hereafter, PlaP^m and PlaS^y), as well as 19 of each variety from Avellanet (hereafter, AveP^m and AveS^y) (Table S1). We sequenced them using haplotagging, a method of linked-read sequencing (Meier et al., 2021). An advantage of haplotagging over standard short-read sequencing is the ability to track source haplotypes by means of molecular barcoding. After mapping sequence reads to the *A. majus* reference genome v. 3.5 (Li et al., 2019), we followed the variant calling and imputation pipeline outlined in Meier et al. (2021) that leverages linked-read information to identify 11,533,030 bi-allelic SNPs (22 SNPs/kbp) across all of the samples. We then phased the SNPs using *SHAPEIT5* (Hofmeister et al., 2023) and used information from a closely related outgroup (*A. molle*; Durán-Castillo et al., 2022) to polarise variants as ancestral or derived.

Based on previous work that showed low genome-wide differentiation between the varieties in Planoles (Tavares et al., 2018), we expected to see similarly low differentiation at the previously unstudied hybrid zone at Avellanet. To test this hypothesis, we generated a LD-thinned dataset containing 1.71 million SNPs and performed *Admixture* (Fig 1B; S1), principal component (Fig. 1C) and phylogenetic (Fig. 1D, S2) analysis to characterise genetic structure. In contrast to our expectations, we found different genetic structure at each hybrid zone. Specifically, PlaP^m and PlaS^y always formed a single group, rather than clustering by flower colour. In contrast, the AveP^m and AveS^y always formed two distinct groups. This result was also supported by the average genome-wide F_{ST} estimated from all 11.5 million SNPs, which showed that genetic differentiation was much lower at Planoles (PlaP^m vs. PlaS^y: F_{ST} = 0.003) than it was at Avellanet (AveP^m vs. AveS^y: F_{ST} = 0.048) (Fig. 1E). In fact, F_{ST} between AveP^m and AveS^y was higher than between AveP^m and PlaP^m (F_{ST} = 0.027), which are separated by more than 50 km, whilst AveS^y and PlaS^y showed the highest pairwise F_{ST} (= 0.066) of all.

The above results suggest a more substantial history of hybridization and gene flow at the Planoles hybrid zone than at Avellanet. To assess this more formally, we used the program *δaδi* (Gutenkunst et al., 2010) to fit a series of demographic models to the joint site frequency spectrum separately at each hybrid zone. We first compared the fit of a model of strict isolation (SI, where two populations diverge with no gene flow) to a model of secondary contact (SC, where populations diverge in allopatry followed by gene exchange after coming back into contact). For both hybrid zones, the SC model was a far better fit to the data than the SI model (ΔAIC > 2000 for both zones), providing evidence of gene flow between the magenta and yellow populations at each zone (Fig. S3, Table S2). It also suggested a more substantial history of gene flow at Planoles characterised by a much longer period since secondary contact than at Avallenet.

Together, these results suggest strikingly different histories of gene flow at each of the hybrid zones, which is largely consistent with observations made at these hybrid zones over

more than a decade. At Planoles, plants are abundant every year, and hybrid individuals can be found over broad areas spanning more than 1 km (Whibley et al., 2006). In contrast, we do not always find a large number of plants at Avellanet (Stankowski, Barton & Field; personal observations). In some years, the plants are abundant, and in others, their distribution is patchy and hybrids are uncommon. Thus, the difference in genetic structure between the zones may reflect the demographic stability of the populations which is what ultimately provides opportunities for hybridization and subsequent gene flow across the zone.

5.2.2 | Genome scans reveal highly heterogenous differentiation landscapes with varying degrees of parallelism

Although we observed a strong difference in the magnitude of F_{ST} at each hybrid zone, it is possible that finer-scale pattern of differentiation along the genome is highly similar. Indeed, highly correlated F_{ST} landscapes have been observed in studies where multiple populations with varying levels of differentiation have been compared (Burri et al., 2015; Stankowski et al., 2019). The general explanation for observing correlated differentiation landscapes is that common evolutionary processes and intrinsic genomic properties have shaped variation across multiple incidences of divergence in isolation (Burri et al., 2015), local adaptation (Jones et al., 2012), or secondary contact (Nouhaud et al., 2022).

To test for correlated differentiation landscapes, we first calculated Hudson's F_{ST} in 10-kbp non-overlapping genomic windows for each pair of populations. This revealed highly variable patterns of differentiation among the comparisons, both in the level of F_{ST} and pattern of heterogeneity. First, comparing the genome scans between var. *pseudomajus* and var. *striatum* at each of the hybrid zones, we found little heterogeneity in the pattern of differentiation at Planoles. F_{ST} was consistently low across most of the genome (median = 0.008, sd = 0.011) (Fig. 2), with the exception of a small number of localised peaks of differentiation rising above the background. In contrast, the F_{ST} landscape was highly heterogeneous at Avellanet (median = 0.028, sd = 0.064), with far greater variability across chromosomes and many areas with pronounced differentiation (Fig. 2). The windowed F_{ST} estimates exhibited strong dissimilarity between the hybrid zones (Spearman's rho = 0.04; hereafter, ρ) (Fig. S4A, Table S6).

The remaining comparisons showed that differentiation patterns depended heavily on the populations included. Most notably, AveS^{y} showing highly parallel patterns (ρ ranging from 0.63 to 0.86) between comparisons that included it (Fig S4B, Table S6). This shows that the highly heterogeneous differentiation landscape at the Avellanet hybrid zone is driven more by the history of AveS^{y} population than by gene flow between AveP^{m} and AveS^{y}.

At both hybrid zones, elevated F_{ST} windows tended to coincide with reduced genetic diversity (π_w) in one of the two populations, and/or elevated between-population sequence divergence (d_{xy}) (Fig. S5). At Planoles, the diversity landscapes were highly similar (ρ = 0.89), with outlier regions tending to show lower π_w in PlaS^{y} (Fig S6A, Table S6). The diversity landscapes were less similar at Avellanet (ρ = 0.47), with the population that showed lower

π_w varying among the genomic regions (Table S6). For example, in the most pronounced F_{ST} island on Chr 2, AveS^y had lower π_w, whereas AveP^m had lower π_w in the outlier regions on Chr 1 (Fig. S6B, S7). At Avellanet, we found a clear negative relationship between the local recombination rate and F_{ST} (ρ = –0.30), indicating that highly differentiated regions at Avellanet tended to show lower recombination rates (Fig. S5, Table S6). Coupled with the low diversity and strongly elevated d_{xy} in highly differentiated regions (Fig. S5), this suggests that F_{ST} has been shaped primarily by widespread linked selection acting independently in the two populations. At Planoles, the relationship between recombination rate and F_{ST} is less pronounced (ρ = 0.08), with outlier regions showing variable rates of recombination and modest reductions in π_w (Fig. S5).

We next examined patterns of differentiation at known colour loci to determine whether they would be detected in outlier scans, since flower colour is expected to have evolved before the do hybrid zone formed (Tavares et al., 2018). At Planoles, two of the known colour regions were identified as F_{ST} outliers using both 95th and 99th percentile thresholds (Fig. 2, S4A). This included regions containing the *Flavia* locus (hereafter, *FLA*) on Chr 2 that controls the intensity of yellow pigmentation, and the two tightly linked loci *Rosea* and *Eluta* on Chr 6 (hereafter, *ROS/EL*) that have large effects on magenta colouration. At Avellanet, the *FLA* locus was identified at the 99th percentile threshold, while *ROS/EL* was only detected at the 95th percentile threshold (Fig. 2, S4A).

5.2.3 | Different genealogical inference methods provide vastly different numbers of trees yet similar genealogical landscapes

Given the challenges of interpreting multiple pairwise F_{ST} scans, we next shifted to genealogical tools that allowed us to jointly analyse relationships among all four populations. Among the variety of tools available, we selected and compared four that were broadly representative of the main approaches in methodological implementation: (*i*) Neighbour-joining trees in arbitrary windows (Martin and Van Belleghem, 2017), (*ii*) *tsinfer* (Kelleher et al., 2019), (*iii*) *Relate* (Speidel et al., 2019), and (*iv*) *Singer* (Deng et al., 2024). We restricted our comparison to a 2 Mbp region (45-65 Mbp in Chr 2) with 461,864 SNPs, that showed highly heterogenous patterns of F_{ST} (Fig. 2) and contains the *FLA* locus (Bradley et al., 2025).

The first method divides the genome into non-overlapping windows containing the same number of SNPs (50 SNPs in our analysis) and infers a phylogenetic tree for each region separately. We inferred neighbour-joining trees, though other methods such as maximum-likelihood have also been applied (Fontaine et al., 2015). While simple and widely used, this approach has a significant limitation. Arbitrarily defined genomic segments often span historical recombination events, where relationships between haplotypes cannot (and ideally should not) be accurately represented by a single bifurcating tree (Shipilina et al., 2023). As a result, important genealogical signals may be dampened by the clumping of unique trees into one.

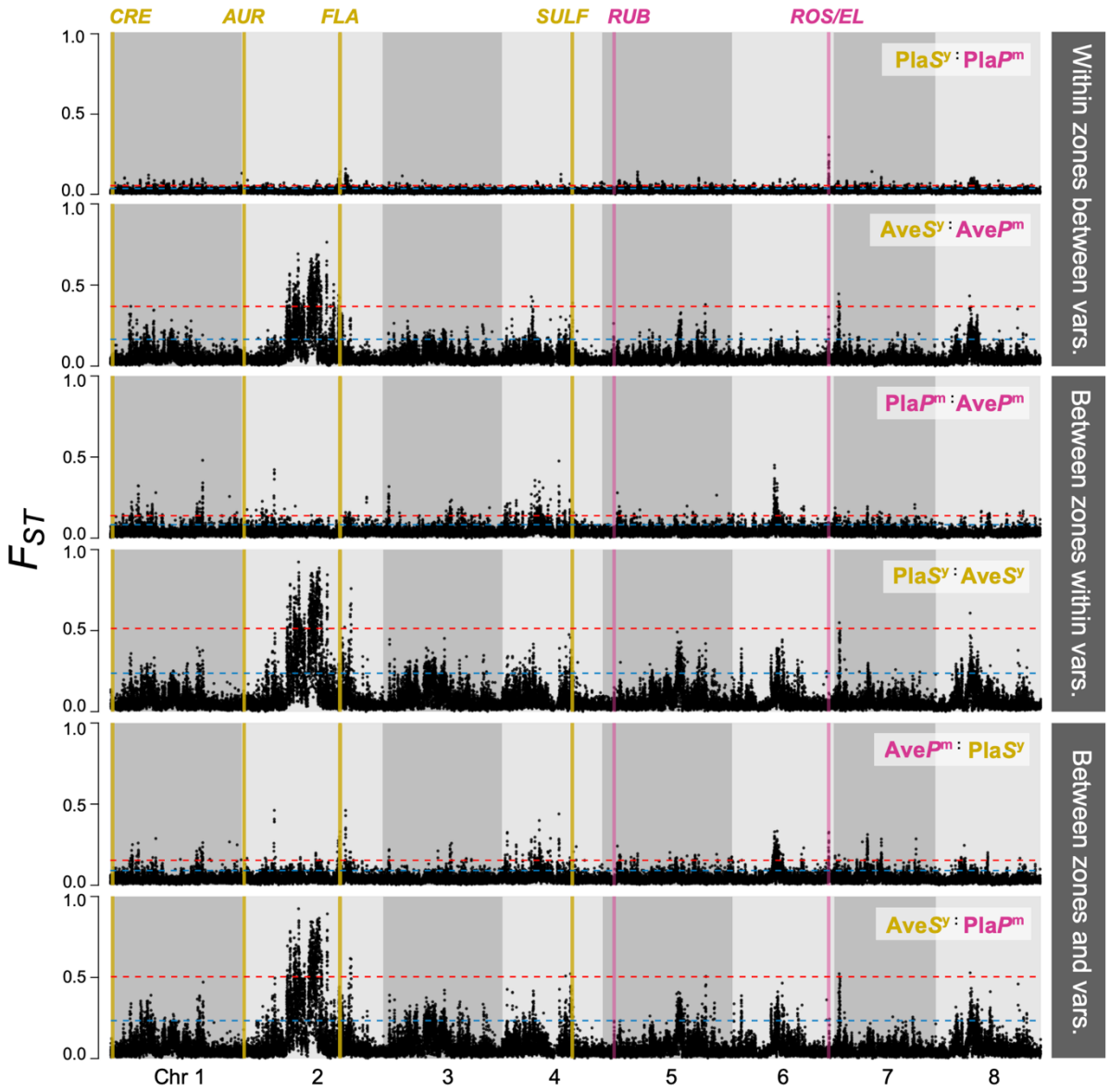

Figure 2. Genome scans show heterogenous F_{ST} landscapes with varying degrees of parallelism. F_{ST} is estimated in 10 kbp non-overlapping windows (n = 50,881) for each chromosome. Dotted blue and red lines show the 95th and 99th percentile of genome-wide F_{ST} estimates. **Top panel** (Within zones and between varieties): Comparison between varieties at each hybrid zone (PlaP^m vs PlaS^y, AveP^m vs AveS^y). **Middle panel** (Between zones and within varieties): Comparison between hybrid zones for magenta and yellow population (PlaP^m vs AveP^m , PlaS^y vs AveS^y). **Bottom panel** (Between zones and varieties): Comparison between varieties from different hybrid zones (PlaS^y vs AveP^m, PlaP^m vs AveS^y). Grey shading delimits chromosomal boundaries. Pla: Planoles, Ave: Avellanet, P^m: magenta-coloured var. *pseudomajus*, S^y: yellow-coloured var. *striatum*.

Although *tsinfer* and *Relate* accommodate the effects of past recombination, they do not explicitly model recombination. In other words, incompatible SNP patterns only imply that historical recombination occurred somewhere between the boundary of neighbouring trees. *Singer* goes further and takes a Bayesian approach, attempting to infer the full ARG by fitting a model of coalescence and recombination to the SNP data. So, unlike the deterministic topology inference of *tsinfer* and *Relate* (i.e., multiple runs will always produce the same tree topologies, though inferred branch lengths may differ between runs), each MCMC iteration

of *Singer* estimates a tree sequence that is drawn from the posterior distribution of possible trees. Thus, a region with no SNP may contain multiple inferred trees, that are not supported by any data within that region, but are a plausible outcome of the estimated model.

Examination of the resulting tree sequences shows that methods produce vastly different results (Table 1). First, we found that the number of trees varied substantially across methods. The neighbour-joining method contained the lowest number of trees at 9,237 (i.e, 1 tree for each 50 SNPs window). *Relate* and *tsinfer* inferred substantially more trees, with 198,375 and 406,135, respectively. *Singer* inferred the most trees by far, with 1,950,778. The average span of an NJ tree was 2.1 kbp (sd = 2.3 kbp), compared with 100 bp (sd = 322 bp) for *Relate*, 50 bp (sd = 222 bp) for *tsinfer*, and 9 bp (sd = 24 bp) for *Singer*. Finally, we asked how many SNPs fell within the span of each marginal tree. Marginal trees in *Relate* contained an average of 2.3 SNP (sd = 1.7) compared with 1.14 (sd = 0.47) for *tsinfer*. On average, trees for *Singer* contain less than 1 SNP (0.24, sd = 0.54), with 1950778 (80.73%) trees containing no SNP.

Table 1. Results of tree inference for four genealogical inference methods. The methods were applied to the same 2 Mbp region (45-65 Mbp in Chr 2) that contained 461,864 SNPs. The total number of trees inferred by each method, the mean span of trees in bp, and mean number of SNPs associated with each tree are provided.

Inference method	Number of trees	Mean tree span (bp)	Mean no. SNPs per tree
Neighbour-joining trees in 50 SNP windows	9,237	2126.44	50
Relate	198,375	99.81	2.33
Tsinfer	406,135	49.24	1.14
Singer	1,950,778	9.25	0.24

Overall, the characteristics of each tree sequence align with their respective methodological approaches. The number of SNPs associated with each NJ tree is defined by the user and will ultimately reflect a trade-off between information content (i.e., number of SNPs) and tree span. In an ideal world, window size would be minimised such that trees span as few recombination events as possible. However, if we assume that the transitions between trees by *tsinfer* reflect real recombination events, this would imply that the average 50 SNP window spans 43 observable recombination events. Although *Relate* and *tsinfer* define margins between trees based on incompatible SNP patterns, *Relate* produces half the number of trees. This may reflect different levels of tolerance for incompatibilities, and the number of trees may vary depending upon the parameters chosen by the user. Finally, because *Singer* models recombination explicitly, it inevitably produces far more trees than the other methods. Although many trees in the sequence are not supported by SNP data, allowing recombination to shape the sequence in the absence of polymorphism data is more consistent with reality, and may provide additional information in some inference schemes.

However, it seems reasonable to exercise caution when making detailed inferences from trees that are not supported by SNP data.

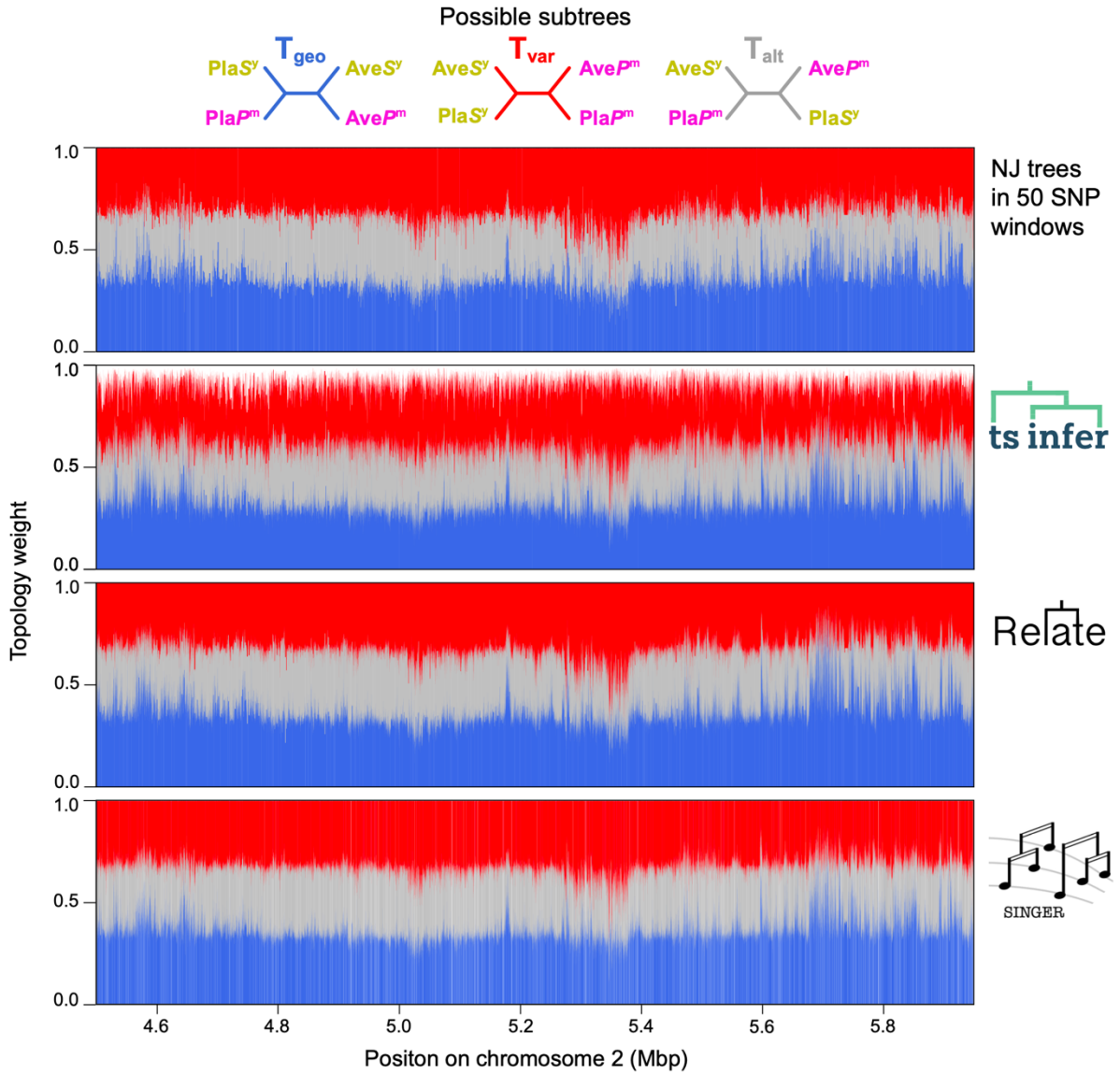

Figure 3. Topology weighting of trees sequences inferred by four different methods yield broadly similar genealogical landscapes. Topology weights of the three possible subtree topologies (T_{geo}, T_{var}, T_{alt}) are plotted for each tree in the sequence along a small section of Chr 2. Each vertical bar shows the proportions of each topology in one genealogical tree. Therefore, topology weights add up to 1 except for trees inferred by *tsinfer* since it allows polytomies.

Moving beyond the summaries of tree sequences, we next used topology weighting to compare how topologies change along the genome. Topology weighting iteratively subsamples one haplotype from each population and estimates the proportion of each subtree topology. In this dataset of four populations: AveP^m, AveS^y, PlaP^m and PlaS^y, it weighs the contribution of each of the three possible topologies (Fig. 3) – the geography tree (T_{geo}), where samples cluster by hybrid zone ((AveS^y,AveP^m)(PlaS^y,PlaP^m)), the variety topology (T_{var}), where samples cluster by the variety, ((AveS^y,PlaS^y)(AveP^m,PlaP^m)), and an alternative

topology (T_{alt}), where samples neither cluster by geography or variety ((AveP^m,PlaS^y)(AveS^y,PlaP^m)). By iteratively sampling many subtrees (in our case 10,000), we can obtain their relative frequencies (i.e., topology weights), which provide a measure of the weight (or bias) of the full tree to each group level topology (Fig. 3).

Although the characteristics of the tree sequences vary among the four methods, there is striking similarity in genomic distributions of the topology weights inferred from them. Figure 3 shows the weights for the three group-level topologies. From visual inspection alone, the topology weights are highly similar among the methods, increasing and decreasing in a coordinated way along the chromosome. Correlation analysis of the weights, performed on the topologies that coincide with SNP positions, shows the similarity is indeed quite strong among the methods (ρ value ranges for T_{geo}: 0.57–0.77, T_{var}: 0.57–0.66, T_{alt}: 0.45–0.71; Fig. S8, Table S7).

However, there are also some clear distinctions. First, the change in amplitude of the weights is not as extreme in the NJ method compared with the other methods. This is not surprising, as the 50 SNP windows span many distinct marginal trees, which we would expect to have a smoothing effect. Another major difference is that topology weights from *tsinfer* do not sum to one, implying that some subtrees cannot be classified as one of the 3 possible topologies. The reason for this is that *tsinfer* infers polytomies, while *Relate* and *Singer* force all branches to bifurcate.

In summary, the results of our comparisons show that the different methods produce vastly different tree sequences, yet largely agree on how group-level relationships change along the genome for our snapdragon dataset. It is good to know that the crudest and most sophisticated approaches give a similar picture, if only from a topological standpoint. Deciding which to use will depend on the size of the dataset and the goals of the study. However, we see little reason use the window-based approach given that more computationally efficient and precise methods are now available. *Tsinfer* and *Relate* are far better options but have different strengths. For example, *tsinfer* retains nodes and branches among trees, meaning that they can be represented as an ARG and used in analysis that leverage homology of tree features (Shipilina et al., 2023). In contrast, *Relate* has been shown to be more accurate than *tsinfer* (+*tsdate*) when it comes to estimating deeper coalescence times (Brandt et al., 2022). *Singer* is far more computationally demanding than the other methods, so it is difficult to scale to large datasets. However, for smaller datasets, and in a more defined genomic regions of interest, *Singer* allows for extremely fine-scale genealogical inference along with estimates of uncertainty.

Unlike NJ trees, *tsinfer, Relate* and *Singer* infer a sequence of trees, consistent with how historical recombination events have altered genealogical relationships across the genome. They both do this by allowing topologies to vary locally to reconcile neighbouring site patterns that cannot be represented as a single bifurcating tree, but their approach varies significantly. *Tsinfer* reconstructs plausible ancestral sequences from sampled chromosomes and then infers the relationship between those sequences, preserving the correlation between consecutive genealogical trees. Therefore, neighbouring trees inevitably share many

of the same nodes and branches. *Relate*, on the other hand, infers a completely new tree upon encountering an incompatible SNP. Therefore, consecutive trees do not share homologous nodes or branches, although this can be partly addressed by assigning the same age to nodes with identical descendant sets across adjacent trees.

5.2.4 | Topology weighting reveals regions associated with flower colour

For our purpose, *Relate* seemed to be the ideal choice to infer genome-wide genealogies, due to its scalability, efficiency, resolution of polytomies, and accuracy of inferring deeper coalescence times. Applying the algorithm to our genome-wide dataset yielded 4,975,454 trees, with an average span of 101.2bp (sd = 739.4bp, median = 47bp, max = 416.8Kbp). We again used topology weighting to quantify bias toward the three group-level relationships (T_{geo}, T_{var}, T_{alt}) for each marginal tree.

We first analysed the distribution of all topology weights in a ternary framework using the program *TwisstNtern* (Stankowski et al., 2024). The ternary plot is a natural framework for analysing the joint distribution of weights in a tree with four groups because it is possible to graphically represent each tree as a single point in an equilateral triangle based on the three weights. The three corners of the ternary plot—[1, 0, 0], [0, 1, 0], [0, 0, 1]—correspond to all of the trees where the sampled subtrees match only one of the three possible group-level subtrees. In contrast, the centre of the ternary plot—[1/3, 1/3, 1/3]—corresponds to where all three of the possible subtrees are found at equal frequency. Any other location in the ternary plot indicates an enrichment of one particular subtree topology. Previous simulations have shown that the ternary distribution of weights can be shaped by a range of factors, including population split times and effective population sizes, as well as processes that lead to haplotype sharing between non-sister groups (e.g., introgression) (Green et al., 2006; Guerrero and Hahn, 2018; Maddison, 1997; Martin and Van Belleghem, 2017; Stankowski et al., 2024).

In our analysis, we expected the ternary distribution of weights to be skewed toward the geography topology (T_{geo}, top of the triangle in Fig. 4), because this topology matches the genome-wide relationships observed between the populations (i.e., topology weighting of the genome-wide neighbour joining tree (Fig. S2) yields weights of T_{geo}=1.0, T_{var}=0.0, T_{alt}=0.0). While we did observe this skew, the bias toward T_{geo} was relatively weak (mean T_{geo}=0.367, T_{var}=0.317, T_{alt}=0.316), and similar across the 8 chromosomes. Although some of the trees showed high T_{geo} weights (max T_{geo}=0.93 with 5% of trees showing weights above 0.49), none of the 4,975,454 trees perfectly matched T_{geo}. Rather, most of the genealogies clustered near the centre of the ternary plot (i.e., most weights were near 0.33 for all three topologies), indicating that haplotype diversity is broadly shared across the four groups.

Notably, we found striking left-right asymmetry in the distribution of topology weights between the left and right halves of the ternary plot (Fig. 4). Specifically, we observed a long tail of topology weights extending toward the right-hand corner of the plot, resulting in a 1% bias in the distribution toward the variety topology (T_{var}). Such a bias is unexpected when

sharing is due to the random sorting of ancestral polymorphism, as there is an equal chance that any given tree will be biased toward either one of the discordant topologies, leading to a symmetrical distribution of weights (Stankowski et al., 2024). This asymmetry is similar to what is measured by the site-based statistic Patterson's *D* (Patterson et al., 2012).

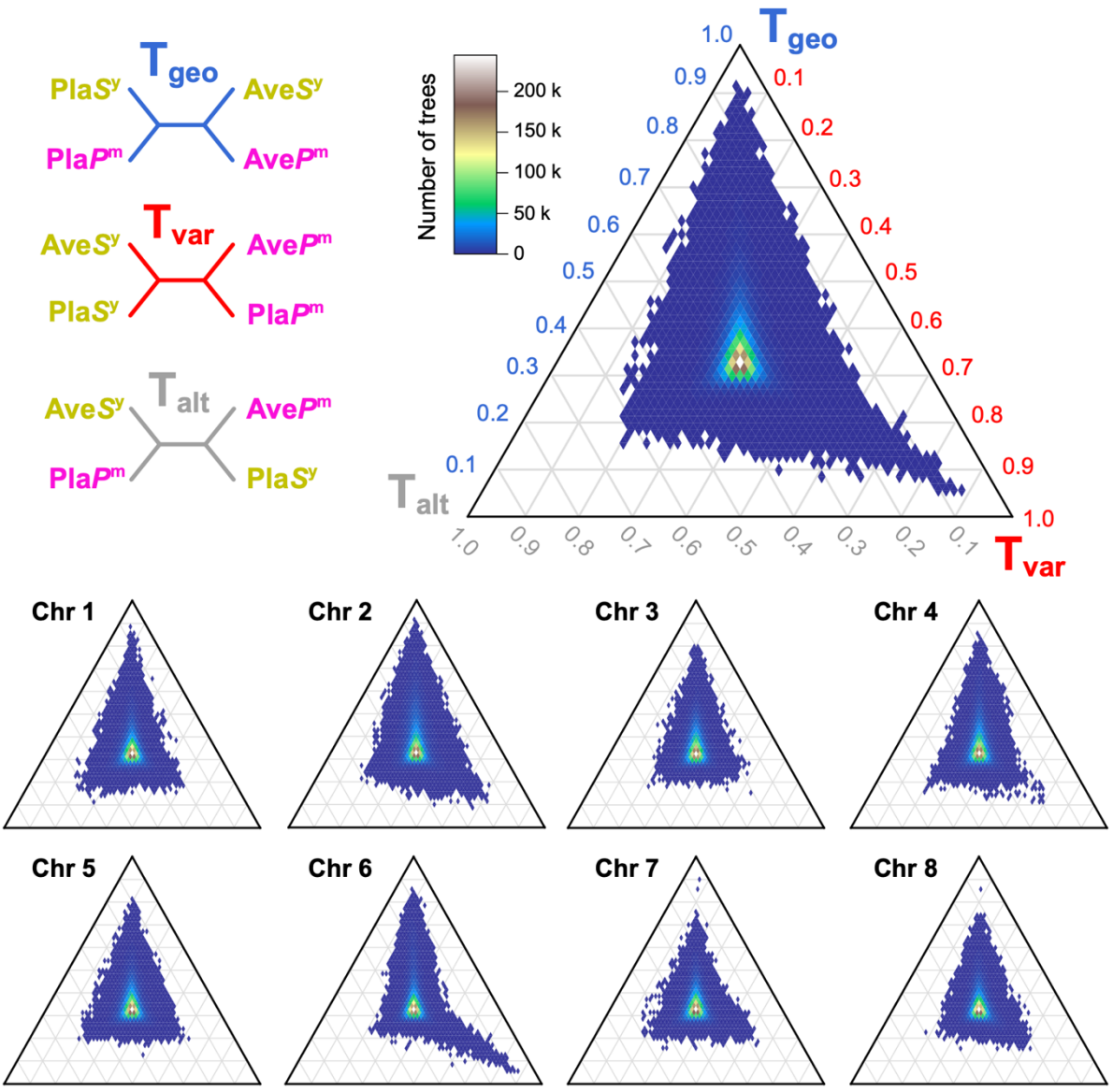

Figure 4. Ternary plots showing the joint distribution of topology weights. Empirical distributions of topology weights for each of the 4,975,454 genealogical trees inferred using *Relate*. Top right: Distribution for the whole genome; Bottom: Distribution for each chromosome. Each tile in the distribution is coloured according to the density of genealogical trees falling in that area of the distribution, as indicated in the colour scale. The three topologies associated with each axis in the ternary plot are shown in the top left of the plot. The three corners of the ternary plot—[1,0,0], [0,1,0], [0,0,1]—correspond to trees that perfectly match the three possible group-level subtrees.

Indeed, roughly symmetrical distributions were observed on several chromosomes, including Chr 1, 3 and 5 (Fig. 4). The remaining chromosomes showed significant asymmetries toward the variety topology, driven by a relatively small number of genealogies (1,490 or

0.03%) with T_{var} weights that exceeded 0.55 (Fig. S9). This indicates a bias of haplotype sharing between populations of the same variety (Fig. S9). The most striking bias was observed on Chr 6, where weights approached 0.9.

To explore regions associated with the genetic differentiation, we plotted the genomic positions of detected T_{var} outliers (defined as T_{var}>0.55, Fig. 5). T_{var} outliers were spread across multiple points along each of the chromosomes rather than clustering at a single site. Most of the known colour genes were observed near T_{var} outliers, but we also observed bias toward T_{var} in regions of the genome that have no known effect on flower colour, including regions of Chr 1, 5, 7 and 8.

5.2.5 | Coalescence times at FLA and ROS/EL differ from the surrounding background

The two genomic regions that showed the clearest association with the colour topology were on Chr 2 and Chr 6, together accounting for 66% of all T_{var} outliers using the 0.55 cutoff (or 84% using the 0.6 cutoff). The outlier region on Chr 2 includes the recently discovered *Flavia* locus, which affects the patterning of yellow colouration in the face of the flower (Bradley et al., 2025) (Fig. 6). This signal of T_{var} enrichment extends over roughly 2 Mbp of the chromosome, interrupting bias toward T_{geo} on either side of it. Within the *FLA* locus, weights for some genealogies exceed 0.7 (Fig. 6). The other region, *ROS/EL* located on Chr 6 contains the two linked colour loci, *Rosea* and *Eluta*. *Rosea* activates anthocyanin biosynthesis across the corolla, while *Eluta* modifies its distribution (Tavares et al., 2018). Within the *ROS/EL* region, T_{var} weights are strongly elevated and characterised by local peaks and troughs spanning about 1 MB. On either side of *ROS/EL*, all three topology weights hover around 0.33, indicating that haplotype variation is broadly distributed among the groups.

Given existing evidence for selection on *FLA* (Bradley et al., 2025) and *ROS/EL* (Tavares et al., 2018), we next examined the coalescence times for genealogies in and around these colour loci. Since positive selection purges haplotype diversity from the population, we may expect to find shallower coalescence times within each variety (often measured using π_w) reflecting the historical sweep of causal alleles (Hejase et al., 2020b). In addition, these loci can also generate local barrier effects in the genome, which we would expect to increase coalescence times between the varieties (often measured using d_{xy}) (Hejase et al., 2020b; Wakeley, 2009).

To test for these patterns, we first compared the median time to the most recent common ancestor (TMRCA) for genealogies inside each locus to those in the flanking regions of the loci where there was no obvious association with colour (Fig. 6). For *FLA*, the median TMRCAs for var. *pseudomajus* (i.e., PlaP^m+AveP^m) was higher inside the locus than in the flanking regions, whereas, for *ROS/EL*, there was no obvious difference (Fig. 7). For var. *striatum* (i.e., PlaS^y+AveS^y), we observed a similar result in *FLA* and *ROS/EL*, with median TMRCAs being lower in the loci than in the flanking regions. We also compared TMRCAs

between the varieties, finding higher median TMRCAs inside *FLA* and *ROS/EL* loci in comparison to the flanking regions.

We also examined the relationship between the TMRCAs and T_{var} weights inside each locus, as we expected signatures to be most pronounced for genealogies that more closely resembled the variety topology. In *FLA*, we found no relationship between the median TMRCA and T_{var} weight within var. *pseudomajus* (ρ = 0.002, Fig. 7). For *ROS/EL*, we only observed a weak negative relationship, with higher T_{var} trees showing a broad range of median TMRCAs within var. *pseudomajus* (Fig. 7). In contrast, we found a very clear negative relationship between median TMRCA and T_{var} weight within var. *striatum* for both loci, such that genealogies with high T_{var} weights tended to have shallower median TMRCAs (ρ = –0.399 for *FLA* and –0.364 for *ROS/EL*, Fig. 7). We also observed a clear positive relationship between T_{var} weights and median TMRCAs between the two varieties, such that genealogies with a high T_{var} tended to have higher TMRCA or deeper coalescence times. Similar patterns are observed when each hybrid zone was analysed separately (Fig. S10, S11).

While preliminary, these results are consistent with (i) selection having acted on haplotypes associated with colour and (ii) suggests that these loci have a local barrier effect. For *FLA*, evidence for selection mainly comes from sharp allele frequency clines at the Planoles hybrid zone (Bradley et al., 2025; Field et al., 2025). Our results provide preliminary evidence that a selective sweep has occurred on the background of var. *striatum* and the allele is present in both Avellanet and Planoles. The lack of a signature in var. *pseudomajus* is consistent with the phenotypic effect of *FLA*, as it only affects yellow colouration. The *ROS/EL* locus is more complex, with previous simulations of the build-up of F_{ST} suggesting that there have been multiple independent sweeps at the two linked loci (Tavares et al., 2018). Here, we only found evidence for selection in the yellow group, while previous work has suggested sweeps on both the yellow and magenta backgrounds. Our results do not preclude such a sweep, as footprints of selection are transient and fade with time. However, it might indicate a more recent sweep in the yellow population. Also, our analysis is coarse-grained and does not consider how fine-scale genealogical relationships change across the region. In future, we plan to use genealogical tools and a much larger dataset to dissect this region in fine detail.

5.2.6 | Conclusions and implications for genomic studies of speciation

In this paper, we studied the genomic landscape associated with replicate hybrid zone in *Antirrhinum majus*. Our study highlights many of the known challenges in interpreting genome scans in the context of adaptation and speciation, as differentiation landscapes can be shaped by a multitude of factors and processes that have nothing to do with speciation *per se* (Ravinet et al., 2017; Wolf and Ellegren, 2017). Comparing two parallel hybrid zones, we show that genome scans can be dominated by signals of historical demography, a factor less widely discussed but critical for isolating speciation related patterns. At Planoles, var. *pseudomajus* and var. *striatum* show very little differentiation, as we would expect between taxa that were described as varieties based only a difference in flower colour. At a second

previously unstudied hybrid zone at Avallenet, differentiation was far more striking and characteristic of more divergent taxa (Bolnick et al., 2023; Stankowski and Ravinet, 2021). This shows that differentiation landscapes can be extremely variable within species, highlighting the dangers of generalising about broader processes from a single pair of samples.

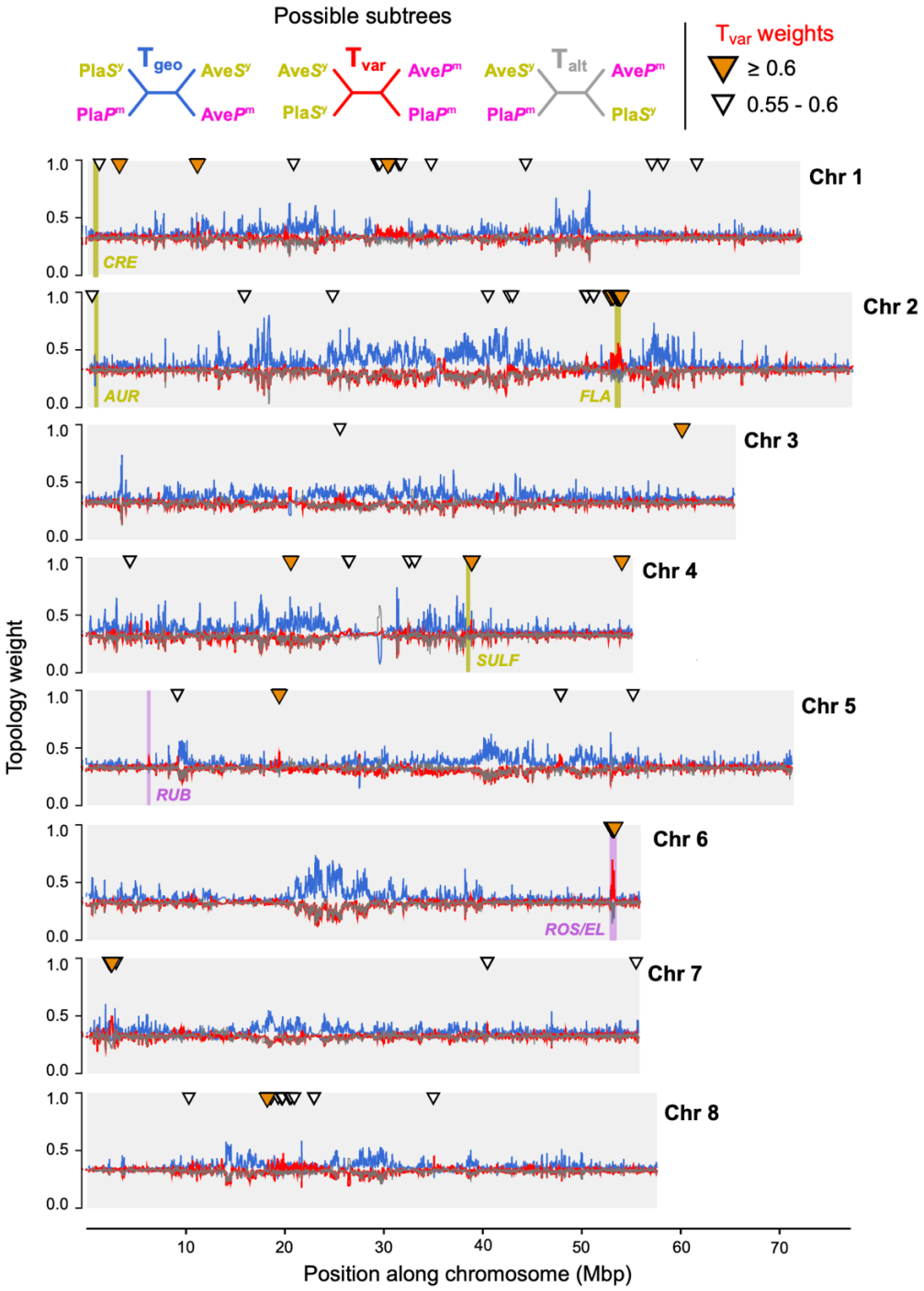

Figure 5. Genealogical landscape of parallel hybrid zone formation revealed by topology weighting. Topology weights (loess smoothed, span = 50 Kbp) for the 4,975,454 trees inferred by *Relate* plotted along each chromosome. 7 loci controlling flower colour are highlighted in yellow or magenta. White triangles indicate trees with raw T_{var} weight between 0.55 and 0.6. Orange triangles indicate trees with raw T_{var} weights ≥ 0.60. Smoothed tracks of T_{geo}, T_{var} and T_{alt} weights are drawn in blue, red and grey respectively.

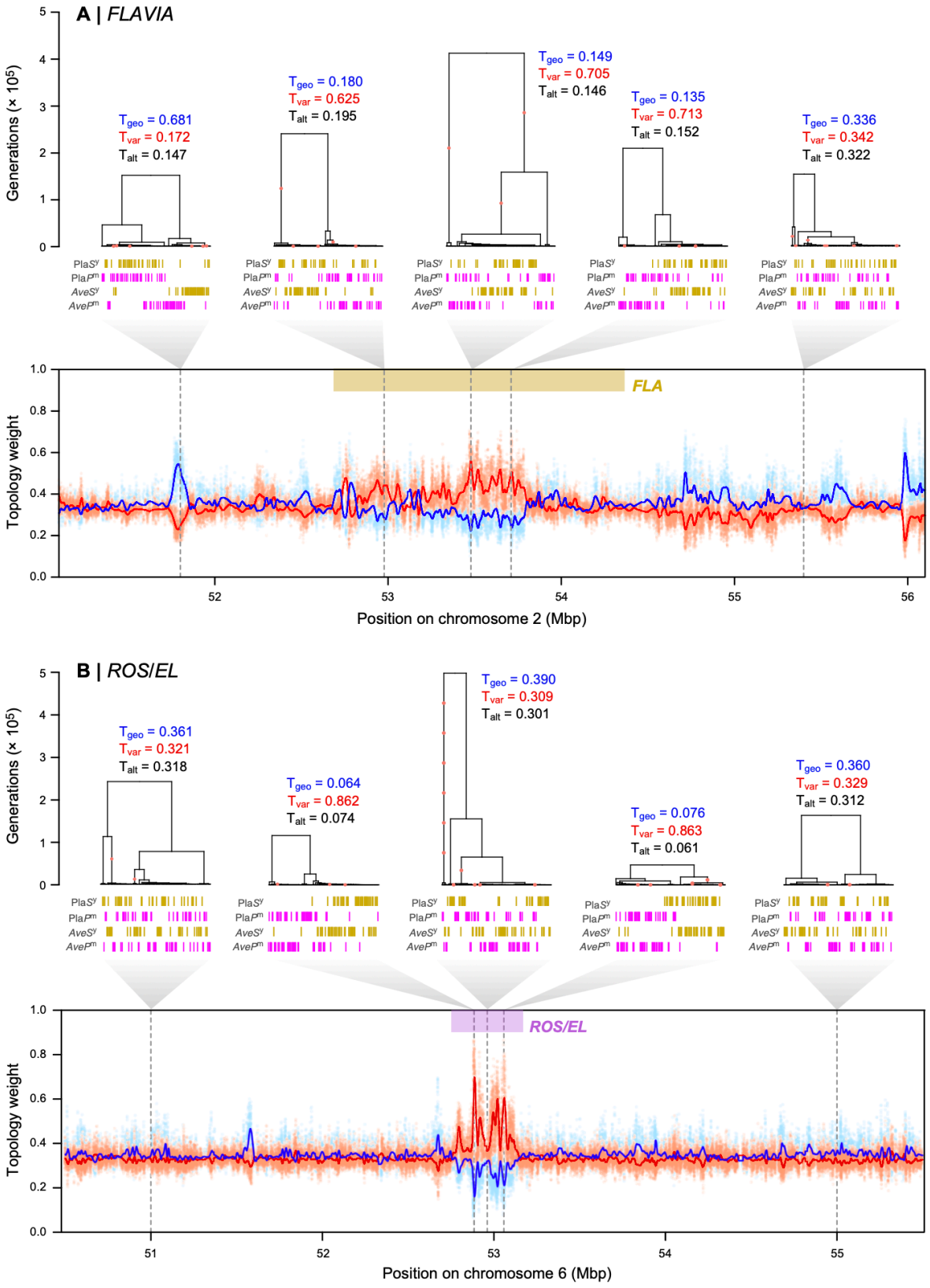

Figure 6. Fine-scale genealogical landscape at the *FLAVIA* and *ROS/EL* loci. (A) 5 Mbp genomic region centred around *Flavia*. Trees show relationships at various points along the sequence, with red circles indicating mutations associated with each tree. Vertical bars represent haplotypes, coloured according to the populations (PlaS^y, PlaP^m, AveS^y, AveP^m). From left to right (1-5), trees 1 and 5 are chosen arbitrarily, but equally distant from the locus. Trees 3 and 4 are trees have the highest smoothed and raw T_{var} weights,

respectively. Bottom panel shows topology weights (T_{geo}, blue; T_{var}, red; T_{alt}, black) through the region. Solid lines are loess smoothed weights (span = 50 kbp), while dots are raw weights. *FLA* locus (Chr2:52560000-54050000) is marked in a yellow bar, while the rest is considered as flank in TMRCA calculations. **(B)** Same as (A), but for *ROS/EL* locus. From left to right, trees 1 and 5 are equally distant from the colour locus, while trees 2-4 are within in. Tree 2 and 4 have the highest raw T_{var} weights at the *ROS1* and *EL* loci. Tree 3 shows a low T_{var} likely due to recombination between the 2 linked loci. *ROS/EL* locus (Chr6:52775000-53150000) is marked in a magenta bar, while the rest is considered as flank in TMRCA calculations.

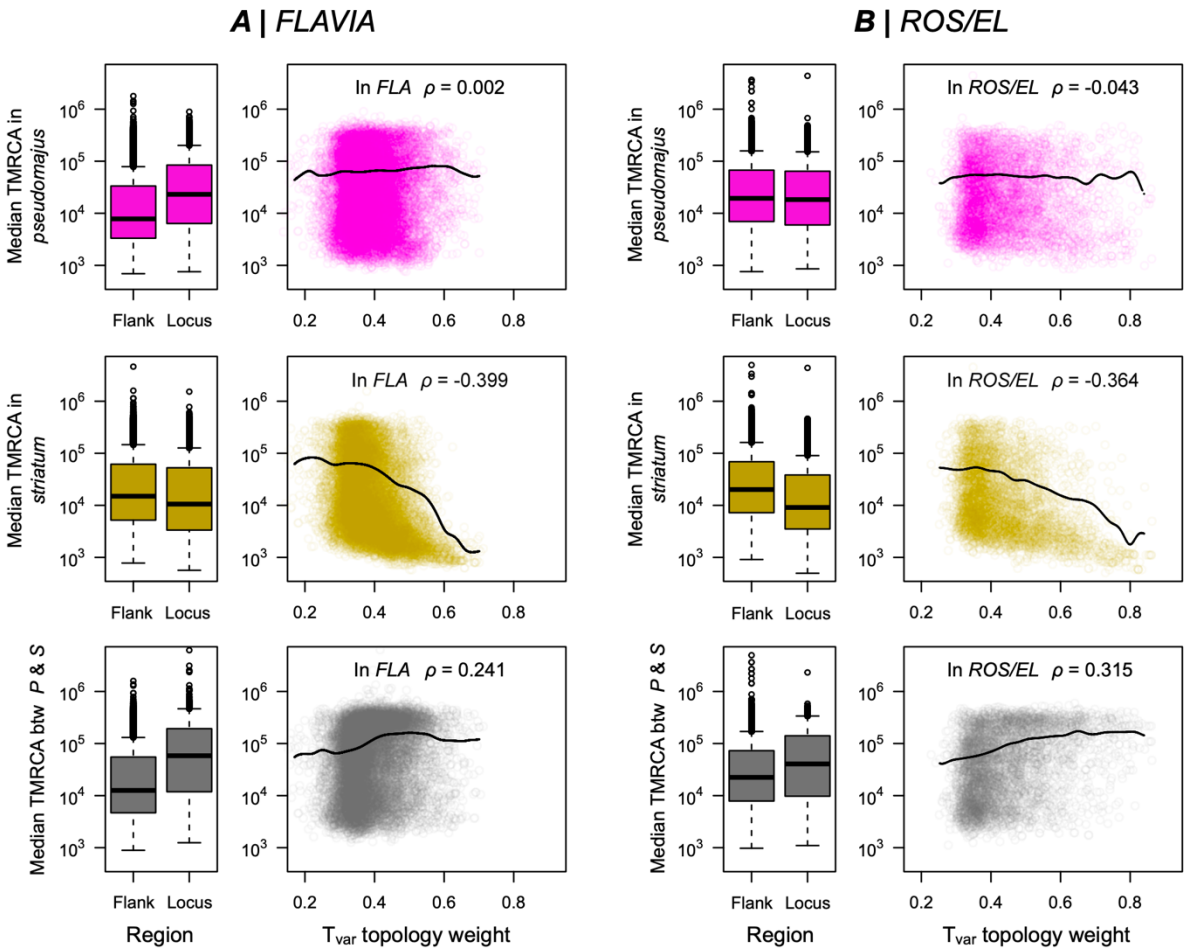

Figure 7. Coalescence times at *FLAVIA* and *ROS/EL*. Boxplots for (A) *FLAVIA* and (B) *ROS/EL* show the median time to most recent common ancestor (TMRCA) within var. *pseudomajus* (top row), within var. *striatum* (middle row), and between var. *pseudomajus* and var. *striatum* (bottom row), all on a log scale. The left boxes ('Flank') show the TMRCAs in the flanking regions around the locus, while the right boxes ('Locus') show the values inside the locus. The scatterplots and dashed black lines show the full distribution and the overall smoothed trend between the T_{var} weight and median TMRCA within the locus, (rows are as indicated for the boxplots). ρ is the correlation coefficient from a Spearman's rank correlation.

Our analyses suggest that different levels of divergence at the two hybrid zones are primarily due to variation in the timing and/or rate of gene flow following secondary contact. This raises important questions about what has caused this difference, and, more broadly, the long-term dynamics of gene flow between *Antirrhinum* varieties in the Pyrenees. The

strongest patterns of differentiation were always observed in comparisons that included var. *striatum* from Avellanet. This result can be explained in several ways. For example, varieties at Avellanet may exhibit historical divergence that was typical of allopatric var. *striatum* and var. *pseudomajus.* This divergence may have been maintained at Avellanet, either by a lower rate of gene flow than at Planoles, or by secondary contact being much more recent than at Planoles. The current spatial distribution of *A. majus ssp. pseudomajus* hints at the first possibility, as the continuous populations around Planoles contrast with the patchier distribution at Avellanet. Moreover, a slightly distant population of *var. striatum,* isolated from Planoles by a mountain pass, shows a similar pattern of differentiation when compared to populations at Planoles (see Figure 5 in Field et al., 2025). Another possibility is that the var. *striatum* population from Avellanet was historically isolated, causing its demographic history to be distinct from other populations. These questions are beyond the scope of our current dataset, and more detailed work is needed to understand the biogeographic and evolutionary history of *A. majus*.

Topology weighting of marginal trees inferred from genealogies allowed us to identify loci associated with the two snapdragon varieties. Flower colour is the only trait that consistently differs between the varieties, suggesting that the T_{var} outliers may underpin this variation. Two of the loci known to cause differences in pigmentation, *Flavia* (Bradley et al., 2025) and *Rosea/Eluta* (Tavares et al., 2018), were identified as T_{var} outliers; whilst others, *Cremosa* (Richardson et al., 2025), *Rubia* (Field et al., 2025), *Sufurea* (Bradley et al., 2017) and *Aurina* (Richardson et al., 2025) did not show clear associations. This could be due to several reasons, including proximity of samples to the core of the hybrid zones, effect size of the loci and sequencing coverage. Most of the T_{var} outliers have not been previously associated with colour. It is possible that these outliers underpin some other trait (floral or non-floral) that differs between the varieties, or they may simply be spurious associations reflecting the highly stochastic nature of the coalescent process. This highlights a more general limitation of *all* genome scans: they detect regions of elevated differentiation between populations, and more detailed mapping studies and functional work are needed to demonstrate causality. *Interpreting these regions also requires understanding the genetic architecture underlying the phenotypes.*

Finally, we encourage others to explore and critically evaluate the utility of genealogical methods in their research. Several recent studies, mostly focusing on human populations, suggest that genealogical tools can lead to more accurate inferences about past evolutionary processes (Fan et al., 2023, 2022; Speidel et al., 2019; Stern et al., 2019; Wohns et al., 2022). However, relatively few studies have used genealogical methods to study adaptation and speciation (Campagna et al., 2017; Hejase et al., 2022, 2020b; Hooper et al., 2024; Meyer et al., 2024; Rueda-M et al., 2024; Stankowski et al., 2024; Wang and Coop, 2022). Our preliminary genealogical comparison of known adaptive loci with the surrounding genomic background further highlights the potential of these tools for studying the interplay between selective sweeps and barriers to gene flow. ARGs and tree sequences are very rich structures that are complex and challenging to interpret (Shipilina et al., 2023). However,

paired with new linked-read sequencing methods, we think there is tremendous scope for creativity around how we can best visualise local genealogical relationships, account for uncertainty, and identify signatures that are associated with the speciation process.

5.3 | Materials and Methods

5.3.1 | Sample collection and DNA extraction

Leaf material was collected from individuals of *A.majus* at two hybrid zones near the towns of Planoles (42.3162ºN, 2.1039ºE) and Avellanet (42.3503ºN, 1.3288ºE) (Table S1). Several leaves were collected from each individual and refrigerated at 4ºC before further processing. DNA was preserved by placing leaf tissue in a paper envelope, and envelopes into an air-tight plastic bag with silica gel. DNA was extracted using a custom protocol optimised for isolating high molecular weight DNA (Supplementary Methods).

5.3.2 | Library preparation and sequencing

Sequencing libraries were constructed by mixing genomic DNA with a pool of haplotagging beads with a different set of A and C barcode oligos (see Supplementary Table 1 for oligonucleotide sequences). This modification shifts the barcode position for the A/C segment from the original i7 index position into Read 2, followed by a mutated Tn5-17A/18G-MEmut sequence – ACTTGTGTATAAGAGACAG (Steiniger-White et al., 2002). The mutated Tn5-MEmut sequence allows tagmentation but does not otherwise interfere with Illumina sequencing. An additional standard 8-bp i7 Illumina index barcode was added during the final PCR amplification to introduce a fifth barcode segment to allow multiplexing of more than 384 samples. Amplified libraries were cleaned up and size-selected using Ampure magnetic beads (Beckman Coulter), Qubit quantified, and adjusted with 10 mM Tris, pH 8, 0.1 mM EDTA to 2.5 nM concentration for sequencing. Libraries were sequenced aiming for 2x coverage with Illumina paired-end sequencing (2x 150 bp) across a lane of Novaseq 6000 S4 by Azenta Life Sciences (Leipzig Germany). The sequences were then demultiplexed by recognising and trimming away the Tn5-MEmut sequence from R2 and the remaining B/D and A/C along with the Plate barcodes. The remaining sequences were processed as previously described in Meier et al., 2021

5.3.3 | Processing of raw reads and read mapping

Raw reads were mapped to the *A. majus* reference genome v3.5 (Li et al., 2019) using *EMA v0.7.0* (Shajii et al., 2018b), a BX-tag-aware modification of *BWA* (Li, 2013). First, haplotag barcodes with BX tags were converted to 16-basepair barcodes using 16BaseBCGen (https://tinyurl.com/SamHaplotag). Reads with correct BX-tags (98.14%) were then mapped with *EMA*, which favours alignments where reads with the same barcode group together. Reads with faulty BX-tags (1.86%) were mapped to the genome using *BWA v0.7.17*. The

resulting BAM files were combined and checked for quality using the *multi-bamqc* command in *qualimap v2.2.1* (Okonechnikov et al., 2016) (Table S1). PCR and optical duplicates were marked and removed using the *markdup* tool in *sambamba* (Tarasov et al., 2015).

5.3.4 | Variant discovery, imputation, phasing and allele polarisation

We used the *mpileup* and *call* commands in *bcftools v1.18* (Danecek et al., 2021) to identify candidate sites that were then used for final genotype inference and imputation by *STITCH v1.6.10* (Davies et al., 2016). Variant calling was performed with the *bcftools* multiallelic calling program using the flags *-m* and *--annotate AD,ADF,ADR,DP,QS,SP*. The resulting VCF was filtered to remove low-quality and potentially erroneous variant sites (Table S2). We first removed all INDELs (*bcftools view -V indels*), all SNPs within 5 basepairs of INDELs (*bcftools filter –SnpGap 5*), all monomorphic REF or ALT sites (*bcftools view -m2 -e "AC==AN || AC==0*), and all sites with more than 2 alleles (*bcftools view -M2*). Next, we removed all sites with >2.5 times the mean coverage across all samples (130x), sites with a genotype quality score <20 and a mapping quality score <30 (*bcftools filter -e "INFO/DP>130 | QUAL<20 | MQ<30*). Finally, bi-alleleic sites with >0.8 of missing genotypes were removed (*bcftools view -e "F_MISSING>0.80"*), producing a set 11,574,426 candidate sites (Table S2).

We applied *STITCH* to impute variants for the 11 million sites described above. *STITCH* models each chromosome as a mosaic of *K* founding haplotypes using both the underlying sequence reads and the linked-read information encoded in the BX-tag. Unlike traditional callers, *STITCH* imputes genotypes in the presence of missing data based on haplotype information from all sequenced individuals. Following guidelines and informed by pilot *STITCH* runs, we used the following parameters: *--K=75, --nGen=100, --niter=40, --expRate=0.5, --downsampleToCov 10 --use_bx_tag TRUE*. To optimise computational resources and runtime, we performed *STITCH* with the above parameters on 1 Mbp regions with an overlap of 100kb overhang allowing them to be combined afterwards. Out of the 11 million sites, 41,396 (0.4%) sites were deemed invariant by *STITCH* (*i.e.,* the *bcftools* and *STITCH* calls disagreed) and were removed leaving a final set of 11,533,030 SNPs of which 93.9% had an INFO score ≥ 0.8, computed by *STITCH* as a proxy for imputation confidence (Table S3). Moreover, the observed and imputed allele frequency were highly correlated (R^2 = 0.87).

Finally, we used the *phase_common_static* from *SHAPEIT5 v 5.1.1* (Hofmeister et al., 2023) to statistically phase genotypes without a reference panel. We polarised alleles in *A. majus* as ancestral or derived using high-coverage PoolSeq sequence data (mean coverage = 89.97x) from multiple populations of the closely related outgroup species *A. molle* (Durán-Castillo et al., 2022). Detailed information on the logic used can be found in the supplementary methods.

5.3.5 | Genome-wide evolutionary relationships and demographic inference

We used three methods to infer genome-wide evolutionary relationships among the sequenced samples. First, we estimated principal components of the genotype matrix. Prior

to analysis, we pruned the dataset to reduce linkage disequilibrium (LD) between neighbouring SNPs (r^2 threshold of 0.1, window size = 50 SNPs, step size = 10 SNPs). This was done using *Plink v2.0 (Chang et al., 2015)* using the command *--indep-pairwise 50 10 0.1*, yielding 1,710,010 SNPs.

We used the model-based clustering program *Admixture v1.3* (Alexander et al., 2009), to assess the genetic structure. *Plink v2.0* was first used to produce BED files from the original VCF file. We ran *Admixture* on the LD-pruned dataset using the unsupervised model for all values of *K* ranging from 2 to 6.

We also inferred a phylogenetic network using the R package *phangorn v2.12* (Schliep, 2011). The LD-pruned dataset was converted to PHYLIP format using the script *vcf2phylip*. We then calculated a distance matrix from all aligned SNPs using the *dist.ml* function with *model = "JC69"*. The phylogenetic network was then inferred using the *neighborNet* function and drawn with *Splitstree v4.19.1* (Huson and Bryant, 2006).

We calculated per-site F_{ST} between each pair of populations on the full SNP dataset, using the approach described by Weir and Cockerham (1984), implemented in *vcftools v0.1.16* (Danecek et al., 2011) using the *--weir-fst-pop* flag. Site-based estimates were averaged to obtain a genome-wide estimate.

We estimated gene flow between the two varieties independently in each locality using the diffusion approximation, implemented in the software program *δaδi* (Gutenkunst et al., 2010). Since secondary contact is considered the most likely explanation for the current distribution of the two varieties (Tavares et al., 2018), we focused on comparing secondary contact models (SC) with strict isolation models (SI). All models included variation in the ancestral population's effective size prior to population split, following Momigliano et al. (2021). In their basic form, both SI and SC models represent a population split into two populations with specific effective population sizes (N1 and N2) that diverge for a period without gene flow (Ts). In the SC model, these populations then begin exchanging migrants during a secondary contact phase (Tsc), with potentially asymmetric migration (M1 and M2). We expanded these models to account for recent population growth (p1 and p2) and/or Hill-Robertson interference by fitting a genome fraction (P) where the effective population is only a fraction (hrf) of what is found in the rest of the genome. In total, we tested 8 distinct models, including 4 modifications of the SI and SC models: (1) standard model (2) model with population growth in the daughter populations (3) a standard model with Hill-Robertson interference and (4) a combined model that included both population growth and Hill-Robertson interference. Each model was fitted 30 times to the data to ensure convergence, and model comparison was performed using the Akaike Information Criterion (AIC). The importance of gene flow in each locality was then compared by calculating the ratio between Tsc/Ts.

5.3.6 | Genome-wide differentiation, diversity and recombination rate

We calculated Hudson's F_{ST} in 10 kb windows for each pair of populations using the script *popgenWindows.py*[1]. Genetic diversity was measured for each site using the *-site-pi* function in *vcftools*.

We used *LDhat v2.2* (Auton and McVean, 2007) to calculate the population-scaled recombination rate (ρ) between each SNP, separately for each population. We first used the *lkgen* function in *LDhat* to generate a log-likelihood lookup table for the number of haplotypes in each population, with $\vartheta = 0.009$ (calculated from average genome-wide π in 10kb windows from previously published study (Tavares et al., 2018). We then used the *interval* function with the parameters: *-its 10000000 -samp 5000 -bpen 5*, to estimate variable recombination rates. Finally, we summarised results from the MCMC iterations to estimate mean ρ between each SNP using the *stat* function with the parameters: *--burnin 1000*. *LDhat* was performed on windows of 2000 variants with an overlap of 100 variants at each end and combined afterwards.

5.3.7 | Genealogical inference

We used four methods to infer trees from our data. First, we inferred neighbour-joining trees for 50 SNP non-overlapping windows using the script *phyml_sliding_windows py*[1] with *–minPerInd = 15.*

The second method used was *tsinfer* (Kelleher et al., 2019). We used a custom script to convert phased, polarised SNPs into the *tskit.samples* format, which was then used to infer tree topologies with the *tsinfer.infer* function in *tsinfer v0.3.2* library[2], followed by the *TreeSequence.simplify* function in *tskit v0.5.8* library[2] to remove unary nodes.

Third, we inferred a tree sequence using *Relate v1.1.8* (Speidel et al., 2019). We assumed $\mu = 5.7 \times 10^{-9}$/bp/generation and uniform recombination rate of 1cM. We initially ran Relate separately on each chromosome setting the haploid N_e to 813388, as derived earlier from $\pi = 4N_e\mu$ where $\pi = 0.009$. We then used the *EstimatePopulationSize.sh* script to jointly infer a time-varying population size history and branch lengths under that history. For this step, we used a *--threshold 0* to ensure that no trees were excluded in the joint-fitting and *--num_iter = 10*. We also included each population in the argument. Finally, we converted the genealogical trees stored in *.anc* and *.mut* format to *.newick* format with the *RelateExtract –mode AncToNewick* function. We focused our analysis on a 5 Mbp region around two flower colour loci: *FLAVIA* (locus – Chr2:52650000-54050000; region including locus and the flanking sequence on either side – Chr2:51100000-56100000) and *ROS/EL* (locus – Chr6:52775000-53150000; region including locus and the flanking sequence on either side – Chr2:50500000-55500000). Specific genealogical trees were plotted using a custom script modified from *Treeview.sh* in *Relate* library. Time to the most recent ancestor (TMRCA) are computed using

[1] https://github.com/simonhmartin/genomics_general
[2] https://github.com/tskit-dev/tskit

a custom modified script from *tskit* library for 3 subsampled groups: within all var. *pseudomajus* individuals, within all var. *striatum* individuals and between var. *pseduomajus* and var. *striatum* individuals. For each case, TMRCA is first computed for all pairwise combinations of individuals, followed by calculating the median.

Finally, we ran *Singer v0.1.7* (Deng et al., 2024) on 500 kbp genomic windows. For each window, we calculated average π with *VCFtools*, which was then used to calculate N_e from $\pi = 4N_e\mu$. We ran *singer_master* with the parameters: *-m = 5.7e-9, -ratio = 1, -mcmc_iter = 100, -thin = 20, -polar = 0.9.* We then used the function *convert_to_tskit* to convert the last MCMC iteration to *tskit* format and to extract trees in *newick* format.

5.3.8 | Topology weighting and ternary analysis

Topology weighting was performed on sequences of trees derived from the various genealogical inference methods using *Twisst* (Martin and Van Belleghem, 2017). Due to the large number of trees and haplotypes, we followed standard *Twisst* guidelines and limited the topology sampling to 10,000 subtrees using the flag *--method fixed*. Genome-wide topology weights were plotted with loess smoothing (span = 50 kbp). We used the *TwisstNTern* framework (Lifchitz et al., 2025) to visualise and calculate asymmetry in the distribution of topology weights for the whole genome, and for each chromosome separately, using the *--superfine* granularity.

5.4 | Supplementary Information

Supplementary information related to this chapter is detailed in Appendix D.

Chapter 6
Dissecting the genetic basis of flower colour in a hybrid zone: Integrating top-down and bottom-up approaches

Abstract

This chapter integrates top-down genome-wide association study (GWAS) with bottom-up genome scans of F_{ST}, D_{XY}, π_w and ARG-based topology weighting to infer the genetic architecture of flower colour across an *Antirrhinum* hybrid zone. After correcting for genetic relatedness, correlations between colour traits and long-range LD with causal loci, our series of GWA analysis found that most of the variation in flower colour can be explained by a handful of loci. For magenta colouration, we confirmed *Rosea*/*Eluta* to have the largest contribution, followed by epistatic effects of *Rubia* on the magenta background. For yellow colouration, we confirmed *Sulfurea*, *Flavia*, and *Cremosa* to contribute the most. Conditional GWAS within the island of divergence that contains *Rosea/Eluta* identified a new potential candidate: a bHLH transcription factor that is associated with magenta colouration. Concordant peaks in differentiation and topological clustering by colour validated barriers loci amid gene flow and showed evidence of selection.

6.1 | Introduction

Studying the genetic basis of traits under selection is a key objective in evolutionary biology, that underpins our understanding of how adaptation, divergence and ultimately, speciation occur (Nosil, 2012; Orr, 2001). Traits that contribute to reproductive isolation, barriers to gene flow or local adaptation—such as beak morphology in Darwin's finches (Podos and Schroeder, 2024), wing patterns in Heliconius butterflies (Jiggins et al., 2001), or armour plates in sticklebacks (Colosimo et al., 2005)—are often direct targets of strong selection. Identifying the genetic variants and architectures underlying such traits provide crucial insight into the mechanisms of how adaptive variation appears and is maintained in natural populations and involved in species divergence. This understanding remains central in linking genetics with evolutionary theory (Barrett and Hoekstra, 2011; Bomblies and Peichel, 2022; Savolainen et al., 2013).

Approaches to dissect the genetic basis of adaptive traits broadly fall into two complementary paradigms referred to as "top-down" (phenotype-driven) and "bottom-up" (genotype-driven) approaches (Barrett and Hoekstra, 2011; Stankowski et al., 2023; Stinchcombe and Hoekstra, 2008). "Top-down" methods, such as Quantitative Trait Locus (QTL) mapping and Genome-Wide Association Studies (GWAS), start with phenotypic variation and scan the genome for markers that show statistical associations with trait values. QTL mapping, typically conducted in controlled crosses, excels at detecting loci with large effects but suffers from low resolution and is limited to segregating variation present in the mapping populations (Lander and Botstein, 1989; Mackay et al., 2009). GWAS, on the other hand, leverages historical recombination in natural or diverse populations, offering potentially higher resolution and ability to detect multiple alleles in fine scale, including those with smaller effects (Atwell et al., 2010; Burton et al., 2007; Visscher et al., 2017). But, while powerful, GWAS results are sensitive to population structure, genetic relatedness, and environmental correlations, which can generate spurious associations, if not rigorously controlled (Berg et al., 2019; Clauw et al., 2024; Price et al., 2006; Uffelmann et al., 2021). Consequently, GWAS associations may require robust validation, typically through molecular genetic work, for traits or genetic regions previously under-studied. The increasing prevalence of mixed-model GWAS (Zhou and Stephens, 2012), Bayesian approaches (e.g., BSLMM) (Zhou et al., 2013), and relatedness matrices (Kang et al., 2010) has mitigated some of these limitations, but the challenge of accurately distinguishing causal from correlated variants persists.

"Bottom-up" approaches, conversely, start with a genome scan without prior phenotypic information, to identify regions exhibiting population-level patterns like elevated genetic differentiation (e.g, F_{ST}), reduced diversity (π_w), elevated divergence (D_{XY}), or specific haplotype patterns (e.g., extended haplotype homozygosity, EHH) (Ravinet et al., 2017; Tang et al., 2007; Wolf and Ellegren, 2017). These methods are particularly adept at highlighting genomic regions involved in local adaptation or reproductive isolation across environmental gradients or hybrid zones (Hooper et al., 2024; Martin et al., 2013; Poelstra et al., 2014;

Todesco et al., 2020). The logic hinges on the expectation that loci under divergent selection will exhibit patterns distinct from the neutral background, such as, F_{ST} outliers or steeper and sharper clines in allele frequency (Barton and Hewitt, 1985; Feder et al., 2012; Wu, 2001). More recently, methods leveraging ancestral recombination graphs (ARGs) complement the above site-based summaries with deeper ancestry reconstruction (Nielsen et al., 2024), by either identifying regions where genealogical histories cluster by phenotype or geography (Martin and Van Belleghem, 2017; Pal et al., 2025; Stankowski et al., 2024), or offering a powerful window into the history of selection, gene flow and introgression (Hejase et al., 2020a; Hooper et al., 2024; Hubisz et al., 2020; Stern et al., 2019, p. 202; Wang and Coop, 2022). While these bottom-up approaches can efficiently scan the genome for regions of elevated differentiation, they may struggle to resolve between demography and selection, or between causal and linked loci—particularly in systems where population structure, population-specific demographic events or hybridisation is prevalent (Ravinet et al., 2017; Wolf and Ellegren, 2017). More crucially, they do not directly link these regions to specific phenotypes, creating a crucial gap. Therefore, to fully study the genetic basis of a trait under selection while also understanding the evolutionary history of the underlying alleles, it is ideal to somewhat integrate the "top-down" and "bottom-up" approaches.

Hybrid zones, where genetically distinct populations meet and interbreed, are a unique setting which allows such integration of top-down and bottom-up approaches to be applied and compared. Gene flow across the zone exposes barrier loci which can be identified by bottom-up approaches (e.g., F_{ST} outliers and cline analysis). Hybridisation breaks up trait associations making it possible to identify genotype phenotype associations with methods like GWAS provided that genome-wide LD is not too strong (Brelsford et al., 2017; Buerkle and Lexer, 2008). The mosaic of genotypes and phenotypes generated by secondary contact and restricted but non-zero gene flow allows direct observation of the effects of selection, recombination, and introgression on trait divergence (Barton, 1979; Barton and Gale, 1993; Gompert et al., 2017; Stankowski et al., 2015; Surendranadh et al., 2025; Westram et al., 2018).

In this chapter, I used an integrative genomic approach to study the genetic basis of a trait that differentiates two varieties of the common snapdragon in a hybrid zone in the Spanish Pyrenees, despite widespread gene flow between them. This hybrid zone between two closely related varieties of the *Antirrhinum majus* subspecies *majus*—var. *pseudomajus* and var. *striatum*—has been studied for 2 decades (Bradley et al., 2017; Field et al., 2025; Ringbauer et al., 2018; Surendranadh et al., 2022, 2022; Tavares et al., 2018; Whibley et al., 2006). They differ in their flower colour patterns that signpost the bee entry point. Flowers of the *striatum* variety exhibit a largely yellow aurone background, with restricted veins of magenta anthocyanin on the upper petals, while var. *pseudomajus* exhibits a complementary pattern with magenta anthocyanin contrasted with a restricted patch of yellow at the bee entry point on lower petals. The contrasting flower colour patterns serve as effective pollinator guides and are maintained by strong selection (Surendranadh et al., 2025; Tavares et al., 2018; Whibley et al., 2006). While flower colour shows a steep clinal pattern along the

hybrid zone, the genome background is largely homogenized by gene flow, except a few discrete “islands” of elevated divergence coinciding with loci controlling floral pigmentation, highlighting their role as barrier loci (Ringbauer et al., 2018; Surendranadh et al., 2025).

Molecular genetic studies through hybrid crosses and functional analyses between the varieties have established that flower colour in *Antirrhinum* is regulated predominantly by 7 loci acting on the mostly independent branches of the broader flavonol biosynthesis pathway—anthocyanin (magenta) (Field et al., 2025; Schwinn et al., 2006; Tavares et al., 2018) and aurone (yellow) (Bradley et al., 2025, 2017; Richardson et al., 2025). The magenta anthocyanin pathway is controlled by the transcription factors *Rosea* (*ROS*) that promotes pigment biosynthesis predominantly in the flower's petals, and *Eluta* (*EL*), which restricts pigment expression spatially (Ono et al., 2006; Schwinn et al., 2006; Tavares et al., 2018). A recently identified gene, *Rubia* (*RUB*), modifies magenta intensity in interaction with *ROS* (Field et al., 2025). The yellow aurone pigmentation pathway is regulated by 4 distinct loci including *Sulfurea* (*SULF*), which represses the biosynthesis enzyme 4’-O-glucosyltransferase (Am4’CGT) via small RNAs, resulting in variation of yellow pigment distribution (Bradley et al., 2017). Additional loci—*Flavia* (*FLA*), *Cremosa* (*CRE*), and *Aurina* (*AUR*) further modify yellow pigmentation gradients and intensity through complex cis-regulatory and trans-acting mechanisms (Bradley et al., 2025; Richardson et al., 2025). A salient feature of these 7 loci is their genomic clustering into distinct divergence islands with variable size and recombination rates. The interplay between these loci shapes spatial pigment gradients along the flower and help maintain divergence between the varieties.

Despite prior knowledge from molecular genetic studies using controlled greenhouse crosses, which have primarily identified major-effect loci through targeted analyses, questions remain about the genetic architecture of flower colour variation in *Antirrhinum*. Lab-based approaches, while powerful for detecting large-effect variants in simplified crossing schemes, may overlook minor loci, polygenic background effects, or context-dependent interactions that manifest in natural populations under gene flow and selection. Specifically, what is the full genetic architecture of flower colour, including the number, genomic distribution, and effect sizes of causal variants, as well as any polygenic components contributing to quantitative variation? Are causal variants limited to the known major loci, or do additional minor effect loci and polygenic effects also influence flower colour patterns? Furthermore, the correspondence between signals from “top-down” association methods and “bottom-up” population genomic scans is unclear, as is the extent to which genealogical topology analysis corroborates barriers identified by genetic divergence. Lastly, the degree to which these loci function as reproductive barriers under contemporary natural selection and gene flow has yet to be fully elucidated.

In this chapter, I used a combination of top-down and bottom-up approaches to characterize the genetic basis and evolutionary history of flower colour variation across the *Antirrhinum majus* hybrid zone. Specifically, I conducted the first GWAS in *A. majus* with the aims of confirming known causal loci and identifying novel loci. Then, working from the bottom up, I used population genomic scans and ARG-based topology weighting to determine

if genome-wide associations act as barriers to gene flow in the hybrid zone. These analyses solidify existing knowledge about the genetic basis of flower colour in *Antirrhinum*, while highlighting the potential pitfalls of using GWAS in natural populations.

6.2 | Study system and flower colour variation along a hybrid zone transect

We sampled 1,084 individuals of *Antirrhinum majus* subspecies *majus*, of which 1003 were located in the Planoles hybrid zone. 744 of these plants were sampled along a ~1Km transect through the core of the hybrid zone, where there are higher fractions of plants of hybrid origin, and the rest (~20-25) from 3 distinct demes on either flank (magenta flank: MF1/2/3 and yellow flank: YF1/2/3) (see Fig 1 of Chapter 4 on sampling location). Colloquially, we will hereafter refer to the transect of 1003 plants as the Lower Road. These samples were sequenced using *haplotagging* (Meier et al., 2021) and processed following the pipeline described in Chapter 4 and in Pal et. al., 2025, which yielded 20,043,334 SNPs across the 8 *A. majus* chromosomes and are used in the subsequent analyses.

6.2.1 | Quantifying flower colour

For each sampled plant, we collected and photographed a fully open flower in order to quantify its colour once in the field and later from the photographs in three different ways. As a first method to quantify colour, we scored each flower in the field with a magenta and yellow score (hereafter, 'field' colour scores), based on the position and intensity of each pigment. This was done visually by two samplers according to a colour scoring system established over the ~15 years of fieldwork. Each flower is provided one magenta score (0.5–5) and one yellow score (0.5–3) (Fig S1). The field scores are very coarse and subjective, and do not inform much about pigment variation across the flower, which is crucial for both varieties to signpost entry points to their pollinators. However, it has been the most efficient scoring system for 55,290 flowers collected across 17 years of sampling (beyond the scope and dataset of this specific project).

The second method of colour quantification is also manual, but from photographs and scored by two fixed scorers (Arka Pal and Sean Stankowski), therefore reducing the amount of human noise. This method involved dividing the top-view photograph of each flower into 7 boxes of interest (Box A-B: Upper dorsal petal lobes, C-E: Ventral petal lobe, F-G: Lower lateral lobes), and scoring each box for magenta and yellow colour on a fixed scale of 0-4 (Fig S2). A total possible score of 28 across all boxes was then normalised between 0 and 1, based on the total number of boxes scored. While the box colour scores were also subjective, they describe variation across the face of the flower unlike field colour scores.

For the third method, we extracted quantitative measures of colour directly from images of the flowers using the program snapPallete[3]. SnapPallete allows the user to obtain measures of hue, brightness, and saturation from defined areas of the flower. For each image, we outlined the same 7 areas as were used above for the colour box scores. We then aggregate the data for all 7 circles and use Principal Components Analysis to reduce the dimensionality of these data. Prior to analysis, we explored PCs 1 to 10, but here focus on PC1 and PC2 which show the strongest associations with the red and yellow scores.

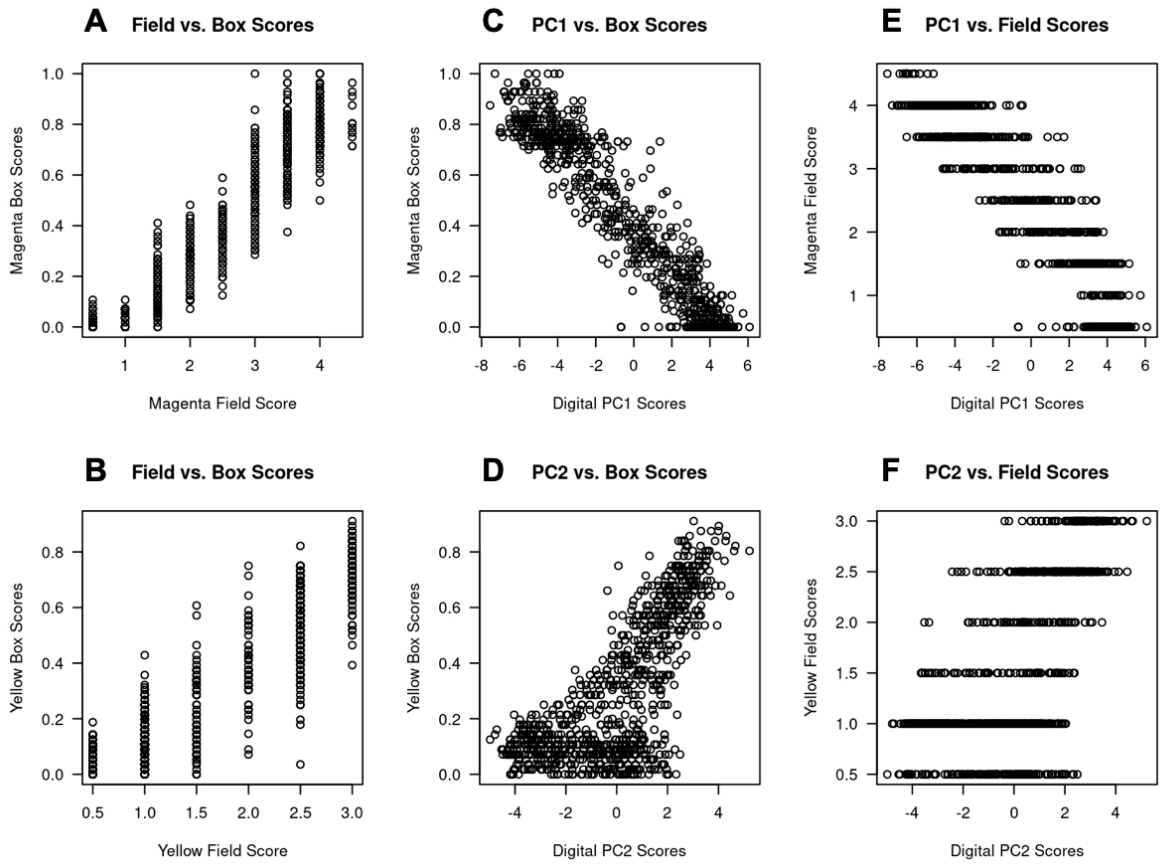

Figure 1. Relationship between colour scores obtained from different scoring methods. Both magenta **(A)** and yellow colour scores **(B)** are highly correlated between the field and box scoring methods. PC1 from the digital scoring method is highly correlated with the magenta scores **(C, E)**, while PC2 is correlated with the yellow scores **(D, F)**.

6.2.2 | Choosing a colour quantification method

Although the three approaches for quantifying colour are different, we found that their colour scores were highly correlated. First, we observed a strong positive correlation between the field scores and colour box scores for both magenta (Spearman's *rho* = 0.92) and yellow (Spearman's *rho* = 0.86) (Fig 1A). A scatterplot of these scores highlights that the colour box scores are more continuously distributed than the much coarser field scores. Since the

[3] https://github.com/seanstankowski/SnapPallete

digital scoring system does not have a specific magenta or yellow score, we performed a principal component (PC) analysis on the mean hue, saturation and value (HSV) from all the circles. The first two PCs explained 45% of the total variation in flower colour (Fig S4C). We found that the PC1 score was highly correlated with the magenta field (Spearman's *rho* = 0.92) and colour box scores (Spearman's *rho* = 0.93) (Fig 1B). However, PC2 was relatively correlated with the yellow scores (Spearman's *rho* for field scores = 0.63, box scores = 0.68) (Fig 1C). Visual inspection of the PC space shows that the digital scores fail to disentangle the presence of high levels of yellow and magenta pigmentation observed in some hybrid phenotypes, with these instead only having intermediate scores on PC2 (Fig S4C).

Based on the above results, we chose to proceed with the box colour scores for our downstream analysis. Although the digital scores provide a more objective and continuous description of the flower colour space, it is difficult to obtain quantitative measures that relate to the underlying colour pigments. Moreover, the individual elements of the digital score can have unintended outcomes, such as averaging hue over all circles in a flower with high expression of both magenta and yellow pigments, results in a hue value that is "blue" due to the continuous nature of HSV colour scale. The field scores are intuitive like the box colour scores, but they are extremely coarse-grained and do not capture spatial variation in colouration across the flower. The colour box scores provide a nice compromise between resolution and interpretability.

6.2.3 | Flower colour varies across the hybrid zone transect

To look at the variation of colour across the hybrid zone, we divided the transect into 51 demes of 30m radius using a hierarchical clustering algorithm to check how magenta and yellow colour varied across the transect. Both magenta and yellow colour showed clinal patterns along the transect, in opposing directions. Mean magenta colour increased from 0.11 to 0.72 from west to east of the transect, while mean yellow colour decreased from 0.54 to 0.09 (Fig S5). The reason for colour scores not reaching 0 or 1 on the opposing ends of the transect is two-fold. First, *striatum* flowers often exhibit the highest aurone intensity in the central part of the petals (see colour scoring in Boxes in Fig S2), while the top and bottom petals lack similar intensity. Thus, when normalised across the whole flower, yellow colour scores never reach a value of 1 for most of the "pure" *striatum* flowers. Second, introgression of colour alleles across the zone leads to variation in scores within each deme. Appreciable rates of long-range pollen and seed dispersal lead to plants with high magenta score to be in the yellow flank or vice-versa (Surendranadh et al., 2025). However, the variance in colour scores in the flanks is much lower (s.d. for both colours ranging 0.1–0.17) compared to the hybrid core (s.d. ranging 0.32–0.54 in the core), where there is an excess of plants of hybrid origin (Fig S5).

To quantitatively describe the colour variation across the hybrid transect, we also fitted a sigmoid cline to the mean colour scores at each deme and found that the narrow clines in magenta and yellow colouration are roughly coincident (cline center for magenta =

2.3Km, yellow = 2.0Km) and concordant (width for magenta = 1.1Km, yellow = 1.3Km) (Fig 2). While the cline shape qualitatively matches previous results, the estimates somewhat differ (width for magenta = 0.77Km, yellow = 0.18Km, in Surendranadh et al., 2025). These differences stem from the large gap in sample size (~1K in this study vs. ~26K in Surendranadh et al, 2025), lack of data in the cline tails (3 demes in the flanks till ~1.5 Km away from the core vs. 999 demes till ~15Km away) and also the cline model fitted (sigmoid in this study vs. stepped in Surendranadh et al., 2025).

Yet, the transition of flower colour scores along the transect confirms the expected clinal pattern across this hybrid zone providing a clear phenotypic foundation for further investigation. Having established this spatial variation in flower colour, we next sought to investigate the genetic architecture underlying these phenotypic differences in *A. m. majus* through genome-wide association analysis.

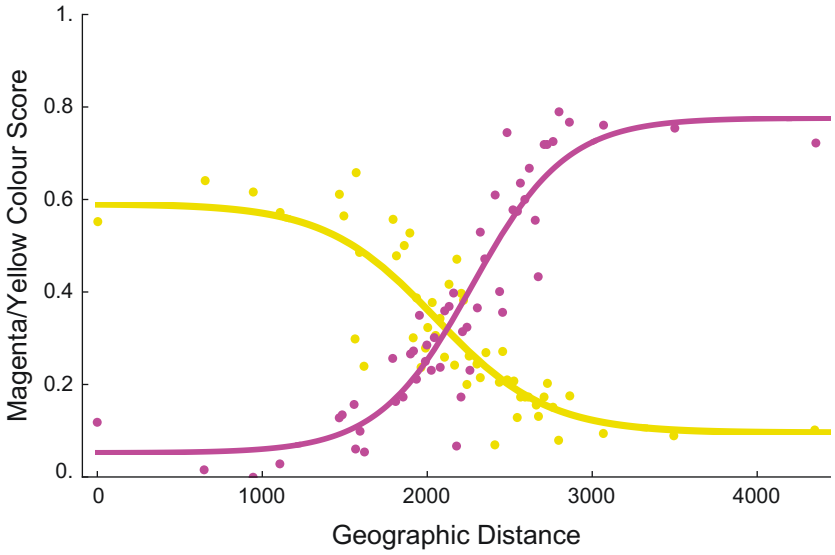

Figure 2. Sigmoid cline fit for the mean normalised magenta (in magenta) and yellow (in yellow) box scores across the 51 demes in the hybrid zones. Each coloured circle denotes the mean corresponding colour score in each deme, and the coloured curves represent the best fit cline from Metropolis-Hastings algorithm. Magenta scores increase and yellow scores decrease from west to east of the transect in a clinal pattern.

6.3 | Disentangling the genetic architecture of flower colour variation

We next investigated the genetic architecture of the flower colour variation. We only retained 915 samples from the full dataset, since they had the "box" colour score information. 744 of these samples came from the hybrid core ('Core' samples) and the remaining 171 from the previously described 3 outer demes on each side of the core ('Flank' samples). Of the 20 million SNPs we initially identified, 15.8 million SNPs segregated in this set of samples. I next performed genome-wide association mapping GEMMA (Zhou and Stephens, 2012) to identify

the loci involved in flower colour variation and better understand the genetic architecture of the colour traits. I formulated a step-wise strategy of conducting a series of frequentist and Bayesian models that can account for possible confounders.

6.3.1 | Linear models (LM) of genome-wide association highlight the need to account for possible confounders

First in our stepwise strategy, we performed a linear model (LM) separately on the magenta and yellow box scores. For both colours, we found significant associations above the Bonferroni corrected significance of likelihood ratio test (LRT) *P*-value threshold ($-\log_{10}P$ = 8.5) at all the loci previously shown to control flower colour variation (Fig 3A, 4A). In the LM for magenta colour, we found significant associations at the three previously identified loci that underpin anthocyanin production. The strongest association was a ~300Kb region around *ROS/EL* on chr 6 (*P*-value at the most associated SNP = 9.6×10^{-238}), followed by a ~100Kb region at the recently discovered *RUB* locus on chr 5 (*P*-value = 1.9×10^{-46}) (Fig 3A). For yellow colour, we also found significant associations at all the previously known yellow-controlling loci, the strongest being a large ~600Kb region around *SULF* locus in chr 4 (*P*-value = 5.5×10^{-55}), followed by a larger ~700Kb region at the *FLA* locus in chr 2 (*P*-value = 1.1×10^{-54}). There were smaller peaks of significant association at the *CRE* locus in chr 1 (~33Kb; *P*-value = 5.4×10^{-27}) and *AUR* locus in chr 2 (~23Kb; *P*-value = 1.5×10^{-14}) (Fig 4A). It was surprising to find all the magenta and yellow-controlling loci significantly associated in both LMs for magenta and yellow, although anthocyanin and aurone biosynthesis pathways are independent (Whibley et al., 2006). Even more surprisingly, in the LM for yellow colour, *ROS/EL* had a higher association signal than *CRE* and *AUR* (Note that association signal refers to higher $-\log_{10}P$, and not to be confused with higher effect size or regression coefficient) (Fig 4A). This overlap of significantly associated regions hinted at possible confounding due to correlation between the colour scores, most evident in the flanks. For example, individuals in the magenta side also tend to have low yellow scores and vice versa (Fig S3), which could create spurious statistical associations. Moreover, the LD at the hybrid core is also quite weak (r ~ 0.03) (Surendranadh et al. 2025).

In addition to associations at the known colour loci, both the LMs for magenta and yellow colour exhibited a rather noisy significance landscape, with 0.17% (~27.7K) and 0.2% (~32.2K) of all SNPs above the $-\log_{10}P$ significance threshold, and 26 and 33 of these significant association peaks separated by at least 1Mb and within the top 0.05% of all associations that visibility rose above the bulk of the genome-wide background. These associated regions have not been previously linked to flower colour variation. While most of these regions were just above the significance threshold, there were 2 regions in chr 3 with stronger significance than *RUB* in the LM for magenta (Fig 3A), and similarly, 15 regions with stronger significance than *AUR* in the LM for yellow (Fig 4A).

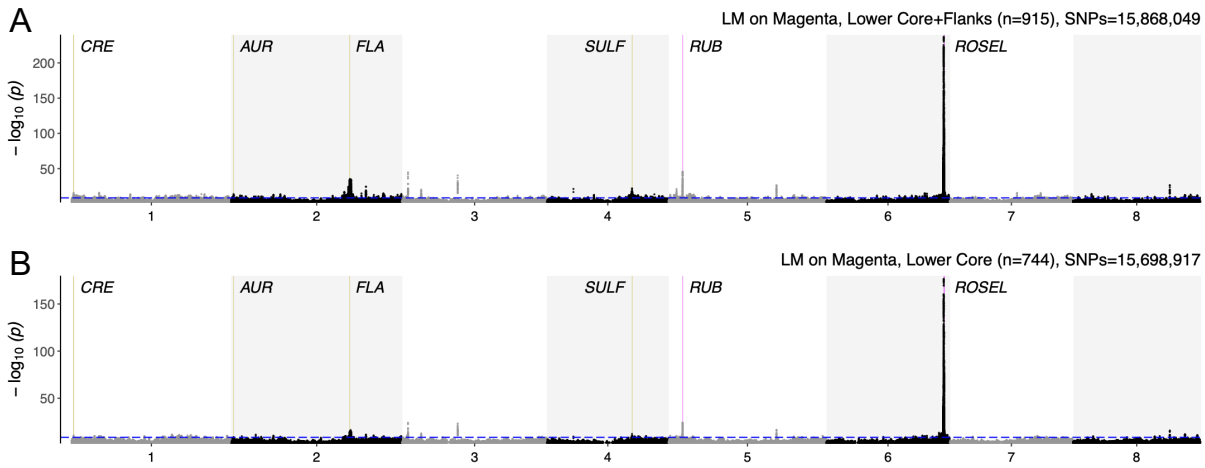

Figure 3. Genome-wide linear model (LM) association with normalised magenta box colour, for samples from the **(A)** whole Lower Road (samples from the core and flanks; *n* = 915) and **(B)** only the Lower Core (*n* = 744). For both LMMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow on based their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

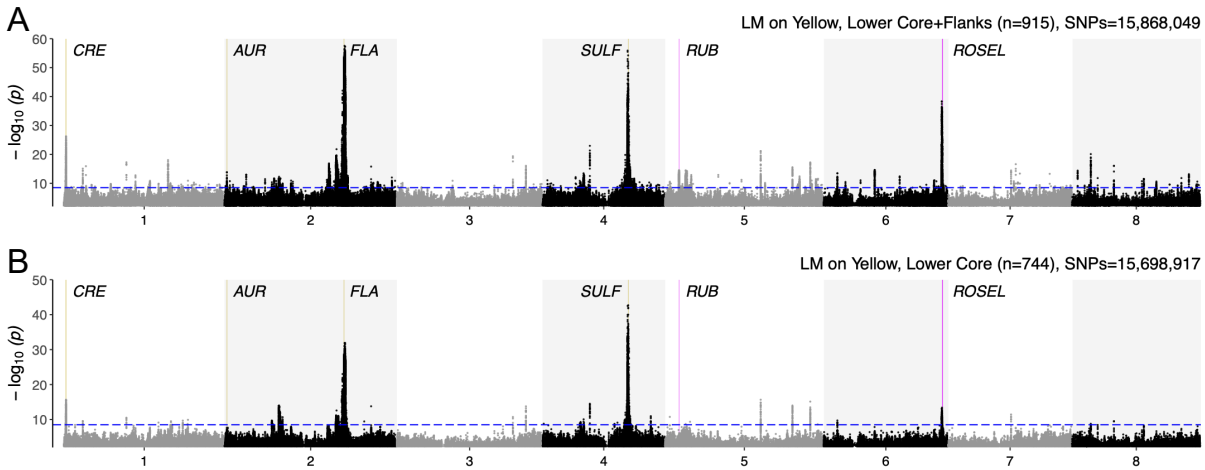

Figure 4. Genome-wide linear model (LM) association with normalised yellow box colour, for samples from the **(A)** whole Lower Road (samples from the core and flanks; *n* = 915) and **(B)** only the Lower Core (*n* = 744). For both LMMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

While all the 915 samples from both core and flanks provided great statistical power to detect SNP associations, samples in the core are crucial because the mixing of "pure" alleles from the *pseudomajus* and *striatum* offers more power to detect independent effects. Therefore, we ran a second LM regression with only the core samples (*n*=744) in order to check if the SNP associations would persist (Fig 3B, 4B). This reduced the overall statistical significance of the associations (i.e., now 0.08% and 0.09% above $-\log_{10}P$ significance threshold in the LM for magenta and yellow colour respectively) that reflects the loss of statistical power due to lower sample size. But all significant associations were congruent

between the two LMs suggesting that these associations did not arise due to over-representation of samples of parental *pseudomajus* or *striatum* genomic background in the flank (Fig 3, 4).

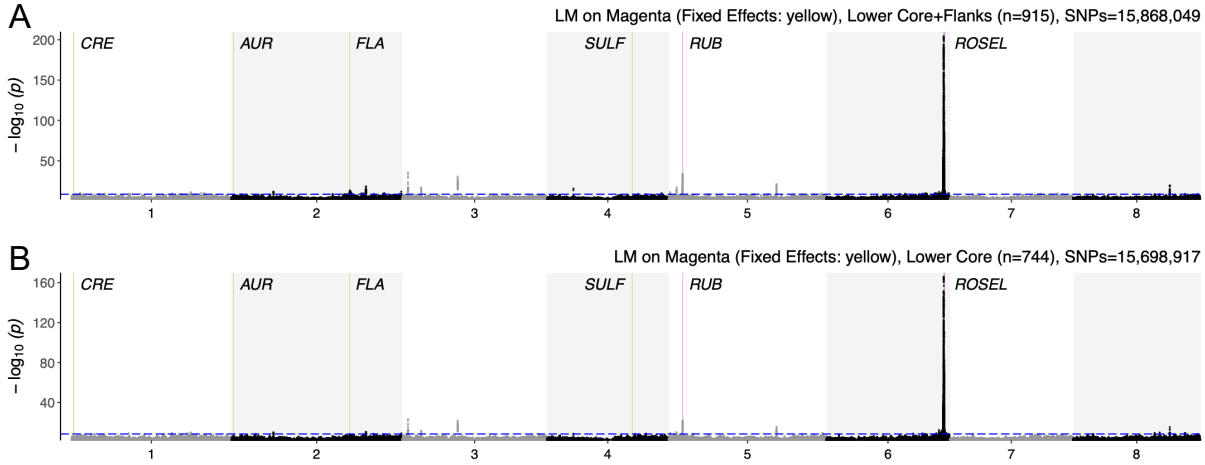

Figure 5. Genome-wide linear model (LM) association with normalised magenta box colour with yellow box colour as fixed-effect covariate, for samples from the **(A)** whole Lower Road (samples from the core and flanks; *n* = 915) and **(B)** only the Lower Core (*n* = 744). For both LMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

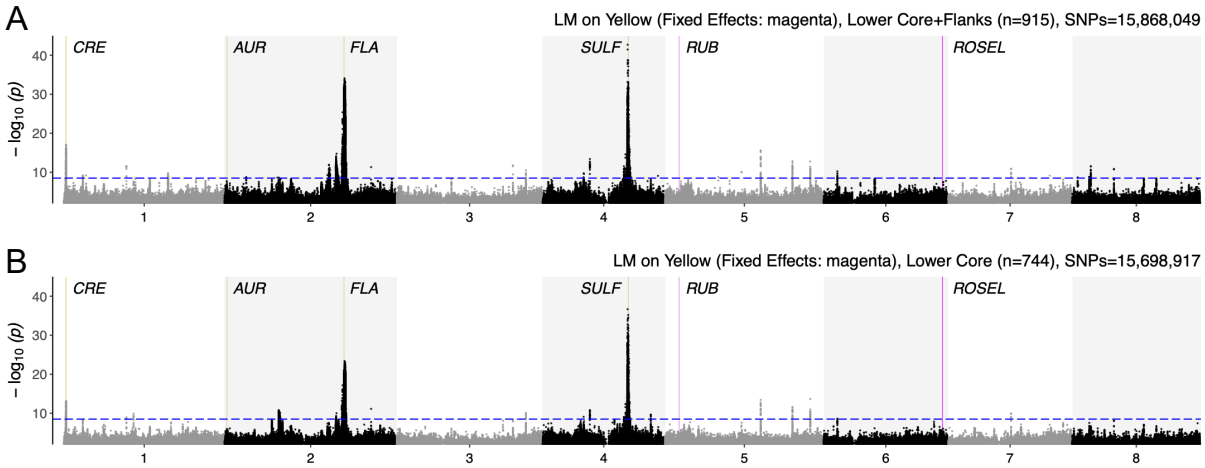

Figure 6. Genome-wide linear model (LM) association with normalised yellow box colour with magenta box colour as fixed-effect covariate, for samples from the **(A)** whole Lower Road (samples from the core and flanks; *n* = 915) and **(B)** only the Lower Core (*n* = 744). For both LMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

To first check if correlation of flower colour scores is driving the unexpected associations at all colour loci in both magenta and yellow LMs, we added the alternate colour

scores as fixed covariate in two new LMs, i.e., we performed a LM for magenta colour with yellow score as a fixed effect; and vice-versa (Fig 5, 6). Meeting our expectation, we found the associations at loci controlling the alternate flower colour either disappeared or were drastically reduced in their significance, i.e., the major association with magenta colour was at *ROS/EL* and *RUB*; while with yellow colour was at *CRE*, *FLA* and *SULF*. However, we surprisingly lost the significant association at *AUR* for yellow colour (most associated SNP *P*-value = 1.8×10^{-8}). Regions previously not linked to flower colour variation still remained significantly associated in this new set of LMs with the alternate colours as fixed covariates (Fig 5, 6).

To check whether these previously uncharacterised regions could be novel candidates for flower colour or just spurious associations stemming from unaccounted confounders, we inspected the quantile-quantile (QQ) plot for *P*-values, and found that observed *P*-values were highly inflated in comparison to the theoretical expectations across almost all quantiles and chromosomes (Fig S6, S7). This highly suggested potential confounding, driven by population structure, genetic relatedness and/or other unaccounted covariates, rather than novel candidate or polygenicity of the flower colour trait.

6.3.2 | Controlling for relatedness and effects of known colour loci

Although the new associations that we found could be causally associated with colour, it is also probable that they are spurious associations due to unaccounted confounding factors. We, therefore, performed linear mixed models (LMMs) to account for genetic relatedness that could arise from several factors such a population structure, assortative mating, etc, that prior work in this hybrid zone has shown evidence for. We computed a genetic kinship matrix from a subset of 0.62 million SNPs which were chosen based on their minor allele frequency, LD and proximity to the known colour loci. This kinship matrix was included as a random factor to perform a linear mixed model (LMM), again separately for magenta and yellow colour.

The overall statistical significance for all SNPs significantly decreased (i.e., lower -$\log_{10}P$ values), with only 0.04% (6,283 SNPs) and 0.04% (6,542 SNPs) above the significance threshold for magenta and yellow colour respectively (Fig 7A, 8A). In the magenta LMM, we again found significant associations at all three magenta loci – *ROS/EL* in chr 6 (*P*-value = 1.1×10^{-157}) and *RUB* in chr 5 (*P*-value = 3.2×10^{-23}) (Fig 7A); while in the yellow LMM, there were significant associations at *CRE* on chr 1 (*P*-value = 2.7×10^{-11}), *FLA* on chr 2 (*P*-value = 2.6×10^{-18}) and *SULF* on chr 4 (*P*-value = 2.5×10^{-35}) (Fig 8A). *AUR* was again not recovered as a significant association (*P*-value = 5.8×10^{-6}). Adding the relatedness matrix removed 21 and 29 associated regions in magenta and yellow respectively, while retaining the highest of those – 5 in magenta on chr 3, 5, 8 and 4 in yellow on chr 5 and 7. The QQ plots showed a better fit between the observed and expected *P*-value quantiles (Fig S8, S9). Unlike the results of LMs, most chromosomes had their observed *P*-values distributed uniformly in the lower quantiles (bottom left of QQ plots in Fig S8, S9) as expected, and a tail of observed *P*-values inflated at

the highest quantiles as would be the case in chromosomes with loci causally linked to a trait (top right corner of QQ-plots in Fig S8, S9).

After correcting for genetic relatedness and correlation between the colour scores, I tested if the newly identified associations are causal or spurious due to long range LD with significant SNPs at previously known colour-associated loci (see associations on chromosomes 3, 5, 8 in LMM for magenta in Fig 7A and chromosomes 5, 7 in LMM for yellow in Fig 8A). To test this, we performed a series of LMMs, where we incrementally included the SNP most associated with the trait as a fixed effect covariate (Fig 7B-E, 8B-G). The rationale is that if a significant association at a previously unidentified locus is causal, it would persist in LMMs that include other known colour loci as fixed effect covariates. Alternatively, any spurious associations that arise due to long range LD with SNPs at known colour loci would disappear if the said colour locus is included as a covariate in the model.

Overall, serially accounting for all the SNPs associated with flower colour indicated that all significant associations at the previously unidentified regions were spurious. Walking through the serial steps for the magenta LMM, we first included the two highest SNPs at the *ROS/EL* region (highest SNP for *ROS* locus and highest SNP for *EL* locus) (Fig 7B). All previously unidentified regions were no longer significantly associated, while known colour-controlling locus *RUB* still was. An association with a slightly reduced statistical significance still existed in the *ROS/EL* region. In the next two LMMs separately, we added the highest SNP at *RUB* and the highest SNP at the remaining association within *ROS/EL*, and found that each of those LMMs retained the association other than the one added as fixed covariate (Fig 7C-D). As a sanity check, a final LMM with all the 4 SNPs above included as fixed covariate left us with no significant associations suggesting that the major causal associations to magenta colour are all at the *ROS/EL* and *RUB* loci (Fig 7E).

Similarly walking through the serial steps for the yellow LMM, we first added the highest SNP in the *SULF* locus as a fixed covariate, and found that associations at the previously known colour-controlling loci *FLA* and *CRE* were retained while the rest were no longer significant. A weaker yet significant association still remained at the *SULF* locus (Fig 8B). Next, in 3 separate LMMs, we added the highest SNP at *CRE*, *FLA* or the remaining association at *SULF*, thus finding that all associations except the SNPs included as covariates were retained (Fig 8C-E). A final LMM that included all the above four SNPs at *CRE*, *FLA* and *SULF* finally revealed no significant associations, thus reiterating that *CRE*, *FLA* and *SULF* are regions that cause yellow colour variation (Fig 8F). We never recovered *AUR* in any of the LMMs (Fig 8).

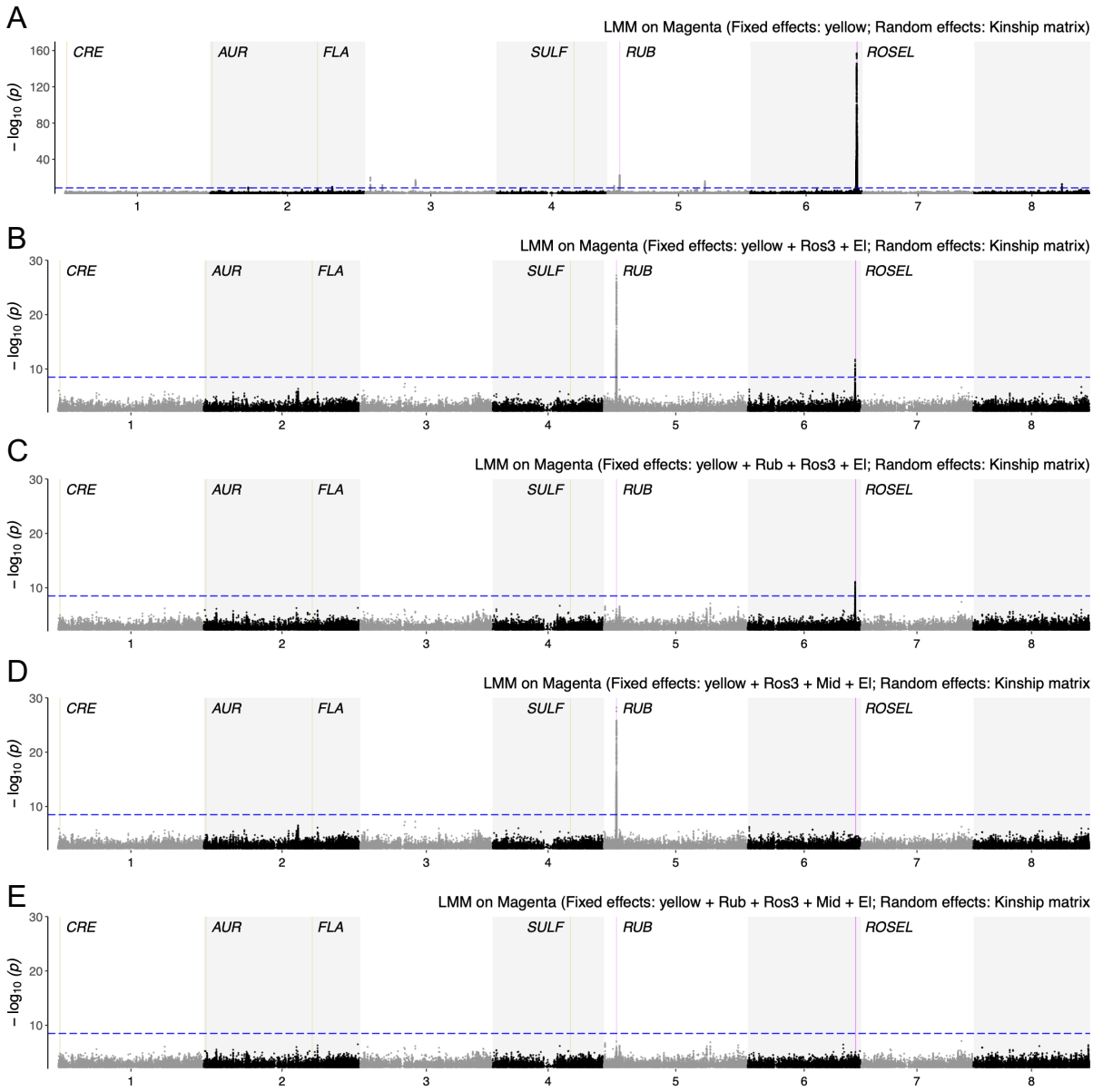

Figure 7. Stepwise genome-wide linear mixed models (LMM) with normalized magenta box scores for samples from the Lower Road (*n* = 915). **(A)** Yellow box colour as fixed-effect covariate and kinship matrix as a random covariate. **(B)** Highest SNP at *ROS3* and *EL* added as fixed-effect covariate to the model in A. **(C)** Highest SNP at *RUB* added as fixed-effect covariate to the model in B. **(D)** Highest SNP at *MID* added as fixed-effect covariate to the model in B. **(E)** Highest SNPs at RUB, ROS3, MID and EL as well as yellow colour added as fixed-effect covariates; kinship matrix added as random effect. For all LMMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr 2, *SULF* in chr 4.

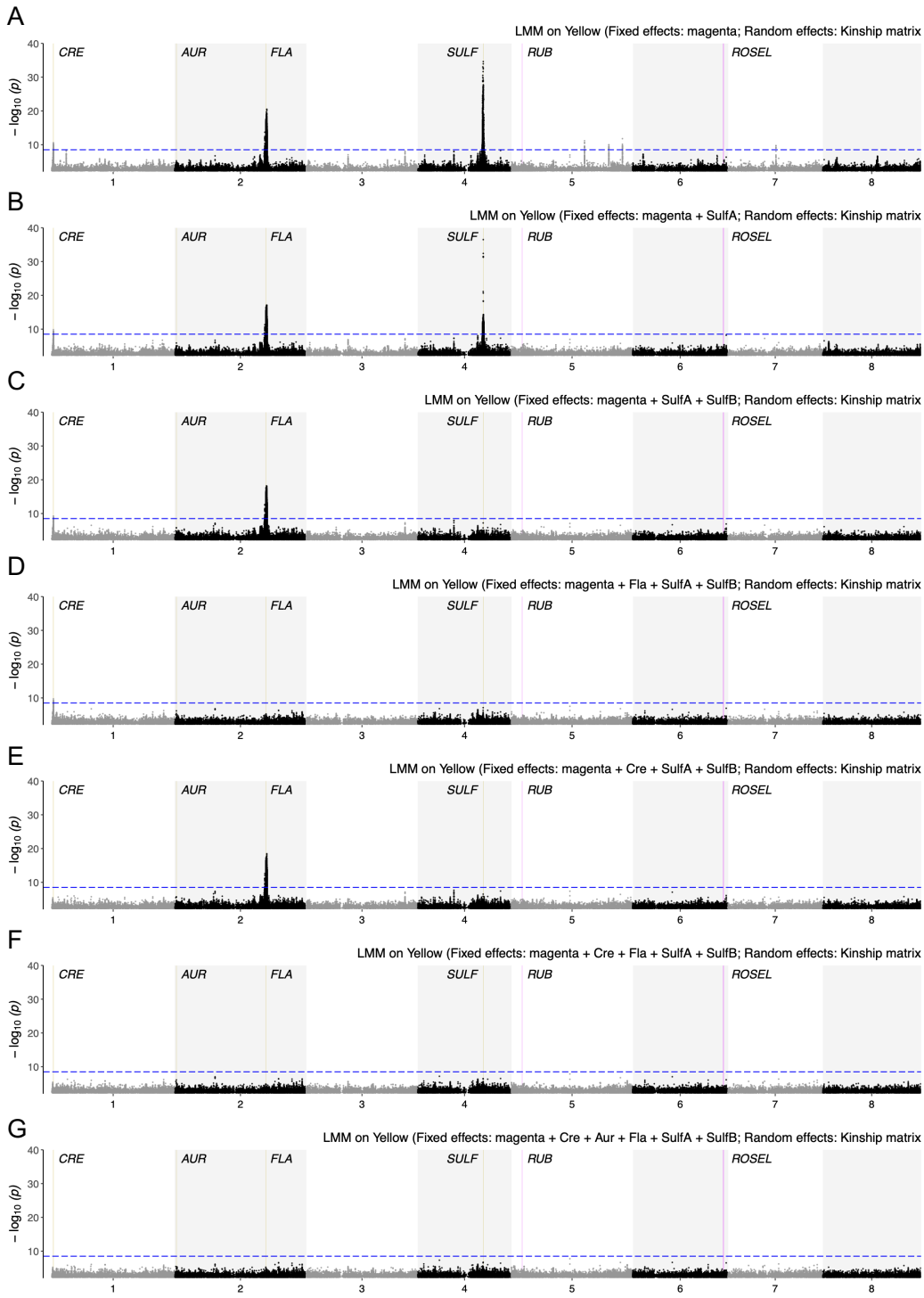

Figure 8. Stepwise genome-wide linear mixed models (LMM) with normalized yellow box scores for samples from the Lower Road (*n* = 915). **(A)** Magenta box colour as fixed-effect covariate and kinship matrix as a random covariate. **(B)** Highest SNP at *SULF* added as fixed-effect covariate to the model in A. **(C)** Highest SNP at the remaining peak at *SULF* added as fixed-effect covariate to the model in B. **(D)** Highest SNP at *FLA* added as fixed-effect covariate to the model in C. **(E)** Highest SNP at *CRE* added as fixed-effect covariate to the model in C. **(F)** Highest SNPs at both *FLA* and *CRE* added as fixed-effect covariate to the model in C. **(G)** Highest SNPs at *AUR* added as fixed-effect covariate to the model in F. For all LMMs, likelihood ratio test (LRT) *P* values with Bonferroni correction are used for detecting significant association ($-\log_{10}P$ threshold = 8.5). All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr 2, *SULF* in chr 4.

6.3.3 | Bayesian analysis sheds light on the genetic architecture while confirms the same loci to have a large-effect on flower colour

While LMMs are useful to uncover associations between a trait and SNP genotypes, they can neither distinguish between polygenic and oligogenic architectures, nor quantify the relative contributions of small versus large effect variants. LMMs assume that all variants contribute a small additive effect on the trait, which is normally distributed with the same variance. The genetic background is modelled as a random effect while each SNP genotype is tested as a fixed effect. On the other hand, models such as Bayesian Sparse Linear Mixed Model (BSLMM) uses Bayesian inference to model jointly the near-infinitesimal effects of genome-wide small polygenic effects (like a LMM) and the additional sparse effects of some loci on the trait in consideration (Bayesian variable selection regression (BVSR)). Therefore, on one hand, BSLMM sheds light into the genetic architecture of the traits, BSLMM provides quantitative estimates for the total phenotypic variance explained (PVE) by all SNP genotypes and the proportion of PVE explained by sparse/large genetic effects (PGE) and the number of SNPs inferred to have sparse-effect in each MCMC iteration. Therefore, LMM and BSLMM can together reveal both the architecture and specific loci involved in a trait, providing a cross validation for SNPs identified by their LMM *P*-values to the posterior probability inclusion (PIP) from BSLMM, and a sanity check to compare the effect sizes of BSLMM and LMM.

Using BSLMM we found that 89% of the total variance in magenta colour was explained by SNP genotypes, of which 88% was explained by loci with effects larger than the joint contributions of the near-infinitesimal genome-wide effects (Fig 9A, Table S1). 17 SNPs genome-wide had a Bayesian posterior inclusion probability (PIP) ≥ 0.05, of which 14 were concentrated in the *RUB* locus in chr 5 and *ROS/EL* in chr 6 (Fig 9B). Similarly for yellow colour, we found 86% of the phenotypic variance being explained by SNP genotypes, of which 74% was attributed to sparse effect SNPs (Fig 10A, Table S1). 27 SNPs had a PIP score ≥ 0.05, of which 22 were concentrated around *CRE*, *FLA* and *SULF* on chr 1, 2, and 4 (Fig 10B).

6.3.4 | Bayesian and frequentist GWAS models confirm the effects of all previously described flower colour loci except one

The results from LMM and BSLMM broadly agree with one another. Both analyses recovered all the previously-identified magenta-controlling loci while 3 out of 4 previously described yellow-controlling loci (Fig 7A, 8A, 9, 10). For magenta, the most associated region was the ~300Kb *ROS/EL* locus on chr 6 (Fig 7A, 9), which upon closer inspection, revealed 4 independent association peaks separated by sharp drops (Fig 11A, S12). The 2 leftmost association peaks within 60Kb of each other contained 3 previously characterised MYB-like transcription factors, *ROS1*, *2*, and *3* which act as a master promoter for genes in the anthocyanin biosynthesis pathway and controls the intensity and overall presence of magenta pigments across all petal lobes (Schwinn et al., 2006; Whibley et al., 2006). *ROS1*, *2*, and *3* share ~90% sequence identity indicating they arose from relatively recent gene duplication, of which *ROS1* has been previously described to be functionally most important. (Schwinn et

al., 2006). But the SNP from the left association peak most associated in the LMM (chr6: 52917323; $P = 1.9\times10^{-46}$) and another one with the highest PIP score in the BSLMM (chr6: 52916405; PIP = 0.71) were both closest (2.2–3.1Kb upstream) to *ROS3* (chr6: 52919567–52922055) (Fig 11A, S12). While the significance of association differs minimally between the 3 homologs (Fig 11A, S12), the higher -$\log_{10}P$ near *ROS3* could just be a statistical effect due to LD with *ROS1* and *2* (84% of samples are either 000, 111, or 222 for the ancestral/derived alleles at *ROS1*/*2*/*3*), rather than evidence for a stronger causal effect.

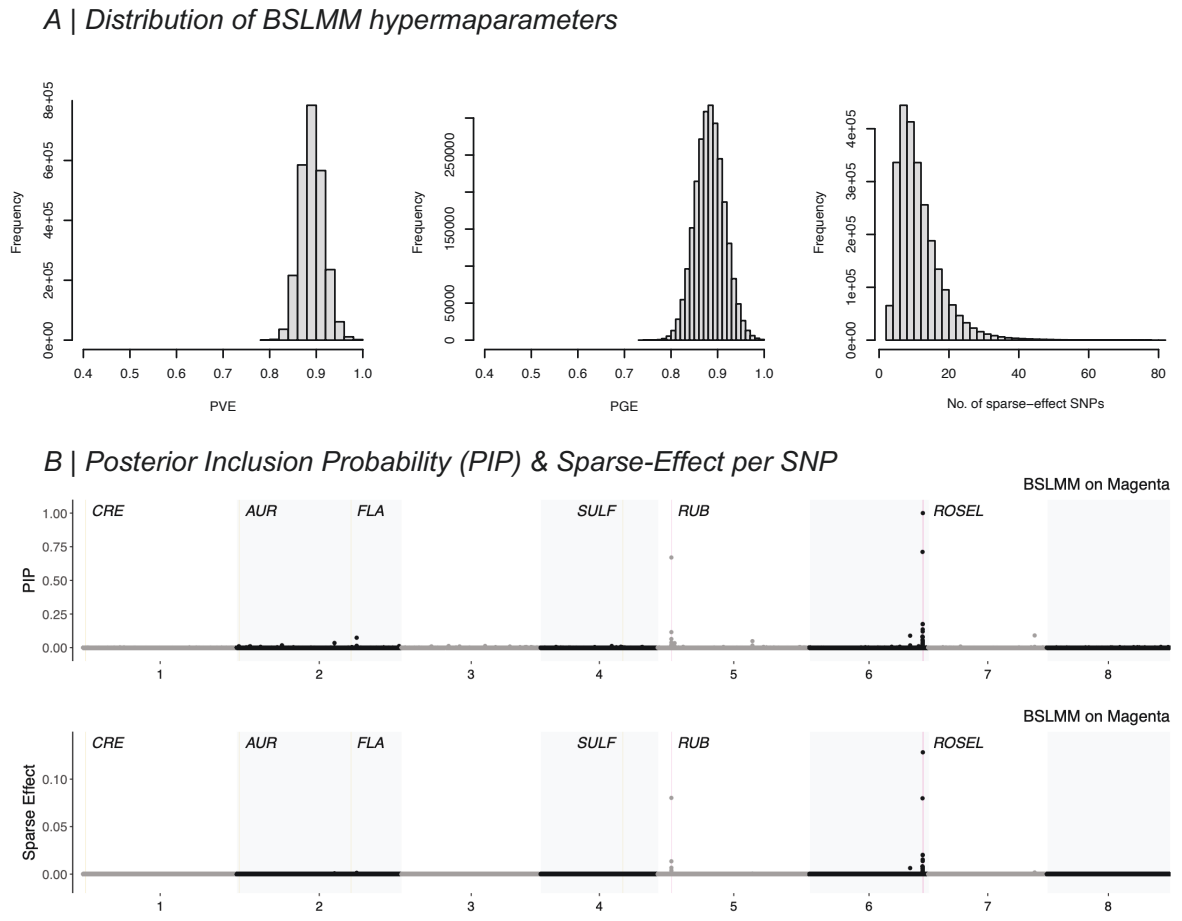

Figure 9. Results of Bayesian Sparse Linear Mixed Model (BSLMM) with normalized magenta box scores for samples from the Lower Road (*n* = 915). (A) Distribution of Bayesian posteriori hyperparameters estimates—proportion of phenotypic variance explained by all SNPs (PVE), proportion of variance explained SNPs with sparse or measurably large effects (PGE) and no. of sparse-effect SNPs included in each iteration of the model. (B) Genome-wide distribution of posterior inclusion probability (PIP) and sparse effect sizes of each SNP. All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: RUB in chr 5, ROS/EL in chr 6; and yellow-controlling loci: CRE in chr 1, AUR in chr 2, FLA in chr 2, SULF in chr 4.

In the most associated SNP from the rightmost peak of the *ROS/EL* region, the most associated SNP in LMM (chr6: 53064556; *P* = 1.1e-145) was located ~2Kb upstream of another previously characterised MYB-like transcription factor, *EL* (chr6: 53060761–53062505). *EL* that acts as a repressor to control dorsoventral spatial patterning of anthocyanin conferred

by *ROS* to lie over the bee entry point and act as a floral guide (Tavares et al., 2018). The same SNP was also the only SNP genome-wide with a perfect PIP score according to the magenta BSLMM (PIP = 1.0), i.e., it was included to have a sparse effect in 100% of all MCMC iterations (Fig 11A, S12).

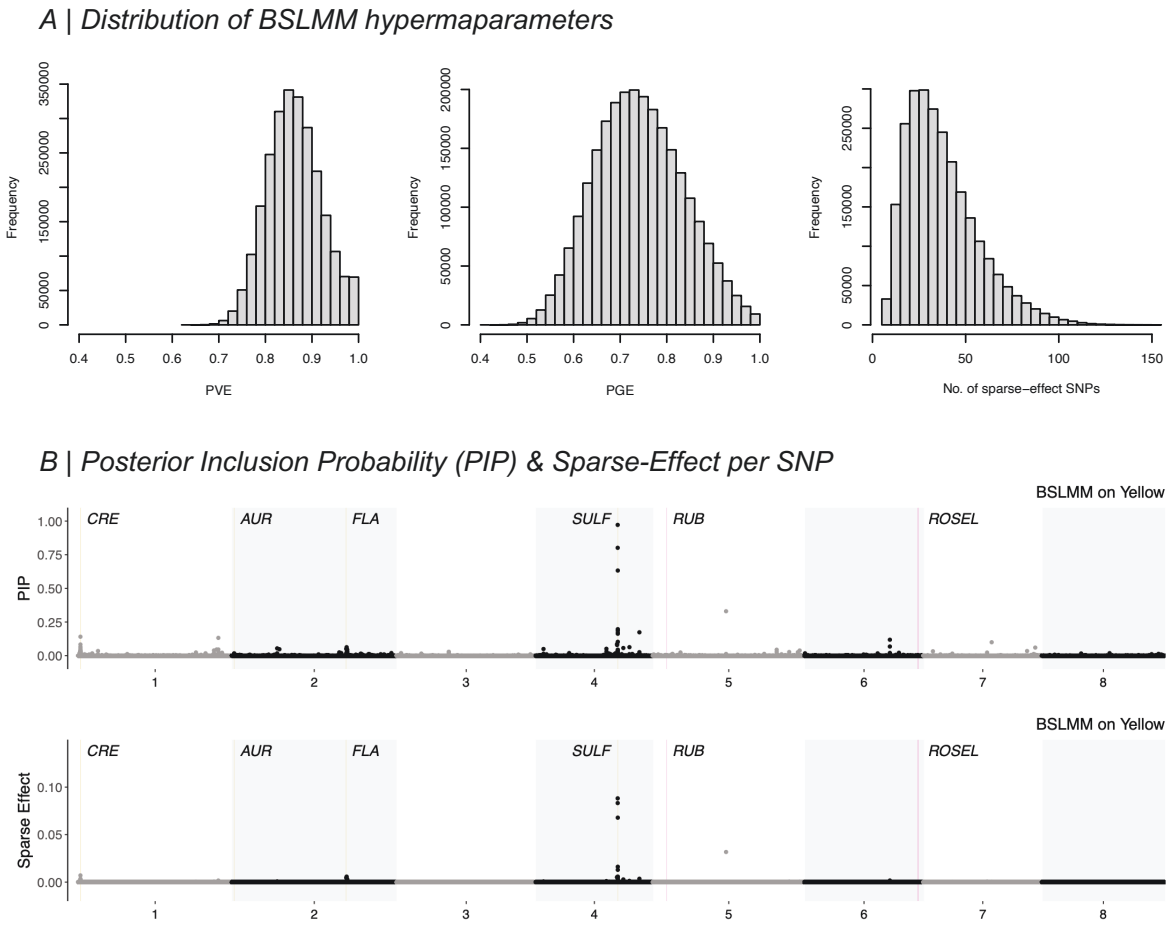

Figure 10. Results of Bayesian Sparse Linear Mixed Model (BSLMM) with normalized yellow box scores for samples from the Lower Road (*n* = 915). **(A)** Distribution of Bayesian posteriori hyperparameters estimates—proportion of phenotypic variance explained by all SNPs (PVE), proportion of variance explained SNPs with sparse or measurably large effects (PGE) and no. of sparse-effect SNPs included in each iteration of the model. **(B)** Genome-wide distribution of posterior inclusion probability (PIP) and sparse effect sizes of each SNP. All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

The third association peak between *ROS* and *EL* had the highest signal of association across the whole genome according to LMM (chr6: 52986462, $P = 1.1\times10^{-157}$) (Fig 11A, S12). 8 of the 12 high-PIP (≥ 0.05) SNPs (PIP = 0.5–0.18) in the *ROS/EL* locus (3 in *ROS* and 1 in *EL*) were also between *ROS* and *EL* (Fig 11A, S12). I further checked with conditional LMMs if this association could be due to a potential candidate locus for magenta colour in Section xx. However, irrespective of being causal or spurious, this explains why we found a reduced but significant association at the *ROS/EL* region, after including the highest associated SNPs at *ROS3* and *EL* as fixed covariates into our LMM (compare *ROS/EL* region in Fig 7A with 7B).

The second region associated with magenta colour was in chr 5, where the SNP that was the most significant (chr5: 6270535, $P = 3.2\times10^{-23}$) was within the coding region of *RUB* (chr5: 6269966–6272151) (Fig 11B, S13). *RUB* encodes flavonol synthase (a key enzyme in the anthocyanin pathway) and modifies the magenta intensity on high-magenta backgrounds (Field et al., 2025). This peak also contained 3 high-PIP SNPs with the highest SNP (chr5: 6269484, PIP = 0.67) 482bp upstream of the gene (Fig 11B, S13). All four loci above were also supported by BSLMM to have the highest sparse-effects on the magenta colour. The top four sparse-effect SNPs were respectively in the vicinity of *EL* (sparse effect size (β) = 0.13), *RUB* (β = 0.08), upstream of *ROS3* (β = 0.079), and the previously uncharacterised locus (hereafter colloquially, *MID*) between *ROS* and *EL* (β = 0.02) (Fig 9B).

The phenotypic effect of the 3 previously known colour loci—*ROS*, *EL* and *RUB*—is clearly visible in the change in magenta scores against their genotypes (Fig S10A). *ROS* and *EL* are known to work together as a complementary regulatory system—while *ROS* determines the capacity for magenta production by turning on the anthocyanin biosynthesis machinery, *EL* determines where magenta appears by spatially restricting the pigmentation regulated by *ROS* (Schwinn et al., 2006; Tavares et al., 2018; Whibley et al., 2006). Together, this results in high anthocyanin production in samples homozygous for the *pseudomajus* alleles at ($ROS^{ps}el^{ps}/ROS^{ps}el^{ps}$; mean ± s.e.magenta score = 0.72 ± 0.01) while little to none in those homozygous for the *striatum* alleles ($ros^{st}EL^{st}/ros^{st}EL^{st}$; mean ± s.e.magenta score = 0.01 ± 0.004) (compare top-left and bottom-right plots in Fig S10A). Heterozygous individuals ($ROS^{ps}el^{ps}/ros^{st}EL^{st}$) showed somewhat intermediate levels of anthocyanin (0.24 ± 0.01), aligning with the semi-dominant nature of *ROS* (center plot in Fig S10A). *RUB*, on the other hand, had minimal effect on the ros^{st}/ros^{st} background (mean magenta scores ranged 0–0.03), while noticeably boosting anthocyanin intensity on the ROS^{ps}/ROS^{ps} background (mean magenta score for rub^{st}/rub^{st} = 0.30 ± 0.07; RUB^{ps}/rub^{st} = 0.55 ± 0.03; RUB^{ps}/RUB^{ps} = 0.72 ± 0.01) (compare left and right column in Fig S10A). This confirms the interaction between *RUB* and *ROS* (Field et al., 2025). *B*eing a transcription factor, *ROS* directly regulates downstream anthocyanin biosynthesis genes, while *RUB* appears to modify the output of the anthocyanin pathway that *ROS* regulates.

The strongest association in the yellow LMM was the ~500Kb region around the *SULF* locus on chr 4. The *SULF* locus is an inverted duplication, homologous to *A. majus* chalcone 4'-O-glucosyltransferase (Am4'CGT), with both left (chr4: 38306079–38306459) and right arm (chr4: 38337252–38337713) of the duplication carrying deletions compared to original 4'CGT (Bradley et al., 2017). The most associated SNP (chr4: 38396398, $P = 2.4\times10^{-35}$) was 58Kb upstream of the right arm of the duplication (Fig 12A, S14). 9 out of the 14 SNPs with PIP ≥ 0.05 in yellow BSLMM were concentrated 11–63Kb upstream of the right arm of the *SULF* locus, with the one with the highest PIP (0.97) closest to the locus (~11Kb) (Fig 12A, S14). Since the *SULF* locus contains a deletion in the *striatum* background and the *Antirrhinum* reference genome is a cultivar originating from the *pseudomajus* background (Li et al., 2019), this region lacks good read coverage. In addition, we lose SNPs in the *SULF* locus (note the noticeable gaps in 38–38.5Mb in Fig 12A, S14) due to filtering based on quality of variant calls.

While we did find SNPs in the *SULF* locus most associated with yellow, it is possible that these are just linked to some other causal SNP. This could explain why we found a significant association at *SULF* at a previously non-significant SNP, after we accounted for the highest associated *SULF* SNP in our step-wise conditional LMMs (compare *SULF* locus in Fig 8A with 8B).

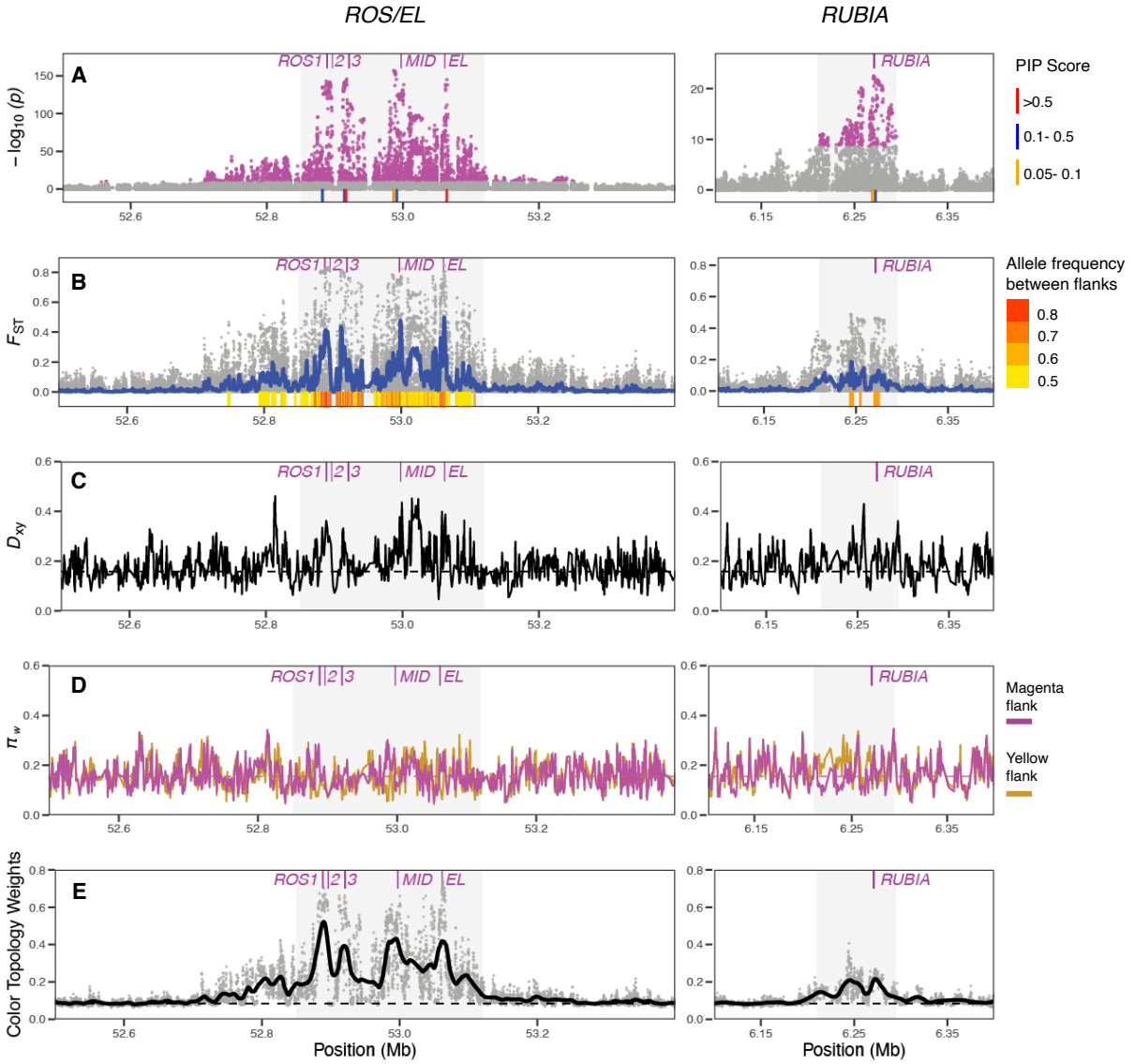

Figure 11. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on magenta colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in *ROS/EL* and *RUBIA* locus.

The second most associated region for yellow was a ~650Kb region at the *FLA* locus in chr 2 (Fig 12B, S15). While the most associated SNP in LMM (chr2: 53986221, $P = 3.2\times10^{-21}$) was ~500Kb upstream of the coding sequence of *FLA* (chr2: 53468062–52469435), this associated region was an extended plateau (~650Kb) rather than a narrow peak (Fig 12B, S15). The *FLA* locus is in a recombination desert possibly owing to its location in the pericentromeric region, and also exhibits a recombinant haplotype prevalent at high frequencies in the center

of the hybrid zone (Bradley et al., 2025). This explains why we find such an extended association instead of a narrow peak. 2 SNPs with PIP ≥ 0.05 were 247Kb (PIP = 0.06) and 66Kb (PIP = 0.05) upstream to the *FLA* gene (Fig 12B, S15). *FLA* and *SULF* act multiplicatively to shape the yellow gradient in the *A. majus* flowers (Bradley et al., 2025, 2017). *SULF* acts in trans to encode small RNAs (sRNAs), which in turn targets and represses *FLA* that encodes chalcone 4'–O–glucosytransferase, a key enzyme in the aurone biosynthesis pathway, This is well supported in our GWAS by being the top 2 loci associated with yellow colour, including 16 out of 20 SNPs with the highest sparse-effects (β = 0.008–0.88) according to BSLMM concentrated between these two loci (Fig 10B).

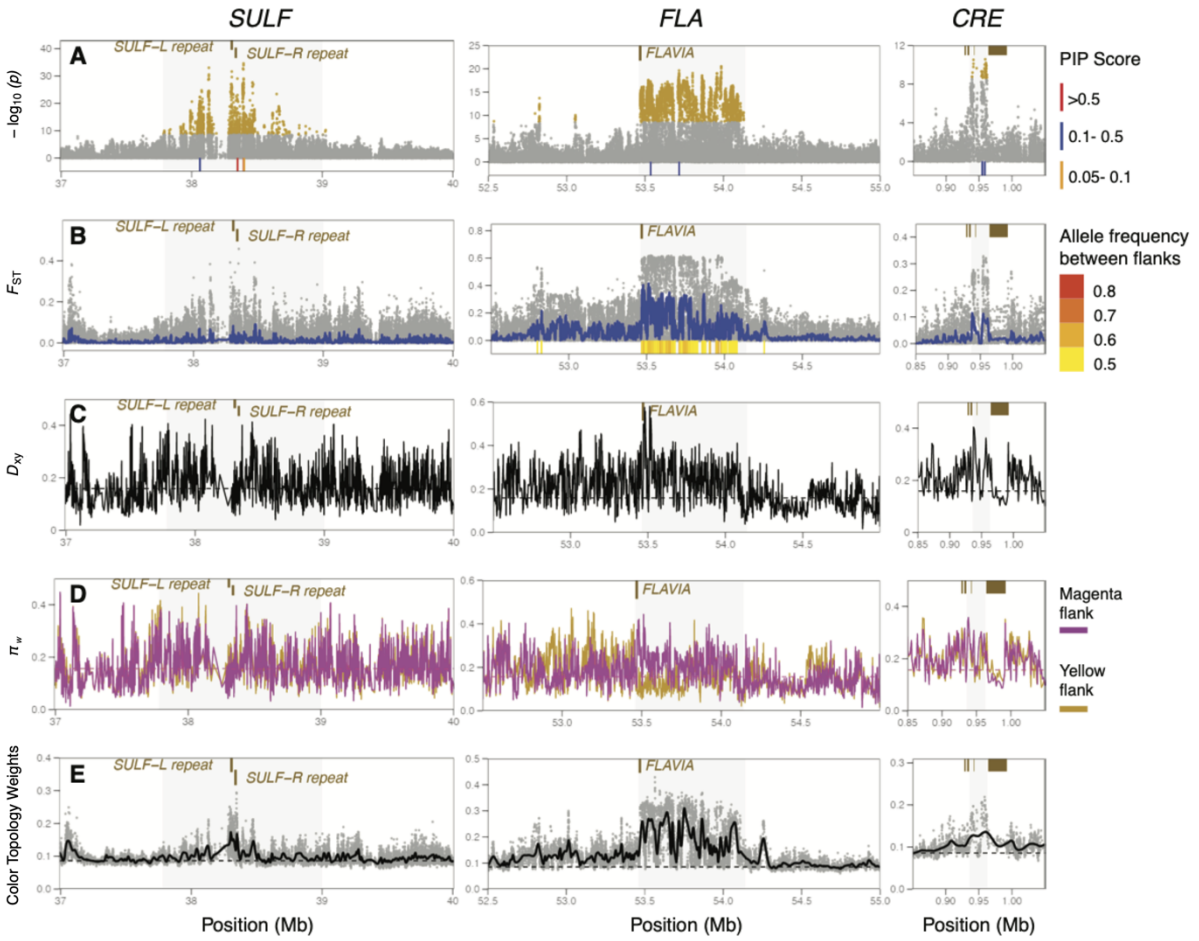

Figure 12. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (T_c) weights in *SULF*, *FLA* and *CRE* locus.

The third most associated region for yellow was a short ~30Kb region in chr 1 around *CRE* locus (Fig 12C, S16), that contains several methyltransferase genes, which have been previously implicated in controlling flower colour by altering upstream precursors in the flavonoid pathways (Bradley et al., 2025; Du et al., 2015; Okitsu et al., 2018; Richardson et al., 2025). Only 53 SNPs were significant according to the LMM (highest associated SNP: chr1:

959480, P = 2.7×10^{-11}), with 4 of those SNPs with PIP ≥ 0.05 (chr1: 954722: PIP = 0.14, β = 0.01) (Fig 12B, S16). Overall, both LMM and BSLMM supported 3 out of the 4 previously described yellow loci to be significantly associated with yellow colour (Fig 8A,10B).

The fourth previously described yellow colour locus, *AUR*, also a target of *FLA*, encodes Aureusidin synthase (AS1) that catalyses the step immediately before 4'CGT in the aurone synthesis pathway (Bradley et al. 2025). *AUR* was not found to be significant in our GWAS by either LMM or BSLMM (Fig 8, 10B). *AUR* interacts multiplicatively with *SULF*, *FLA* and *CRE* to steepen the gradient of aurone pigmentation in the frontal petal lobes along a dorsoventral axis (Bradley et al. 2025). Since the colour score used in the LMMs is an average over the whole flower, it is highly possible that our phenotypic scores do not provide enough resolution to detect the minor effect conferred by *AUR* in controlling the gradient of yellow pigment. Future work should consider regressing colour scores from each part of the flower separately.

The phenotypic effects of 3 out of the 4 yellow-controlling loci can be seen in the mean yellow scores against the genotypes at *CRE*, *FLA* and *SULF* (Fig S11). Individuals homozygous for the *pseudomajus* alleles ($SULF^{ps}fla^{ps}/SULF^{ps}fla^{ps}$) represses aurone production (mean ± s.e. yellow score = 0.08 ± 0.01) compared to those homozygous for the *striatum* alleles ($sulf^{st}FLA^{st}/sulf^{st}FLA^{st}$) (0.62 ± 0.02) (*compare top left and bottom right plot* in Fig S11). Similarly, CRE^{ps}, further reduces aurone production on all backgrounds, but visible most notably when at least minimal amounts of aurone was produced either on the homozygous $sulf^{st}$ background (mean yellow score for CRE^{ps}/CRE^{ps} = 0.33 ± 0.04; CRE^{ps}/cre^{st} = 0.43 ± 0.02; cre^{st}/cre^{st} = 0.59 ± 0.02) or heterozygous $SULF^{ps}/sulf^{st}$ background (mean yellow score for CRE^{ps}/CRE^{ps} = 0.18 ± 0.02; CRE^{ps}/cre^{st} = 0.23 ± 0.01; cre^{st}/cre^{st} = 0.34 ± 0.02) (*see middle and right column* in Fig S11). The yellow-controlling loci interact multiplicatively to inhibit aurone production, i.e., each gene contributes a proportional effect to yellow pigment reduction to precisely control the yellow gradient along the dorsoventral axis in the petals (Bradley et al. 2025). The dominant $SULF^{ps}$ is the master inhibitor that targets *FLA* in trans, followed by FLA^{st} that positively regulates the production of 4'CGT in cis, and finally CRE^{ps} confers further inhibition and is a target of *FLA*. This is also confirmed by our GWAS for yellow colour, where we find that the statistical significance of the association follows the same order—*SULF*, *FLA* and *CRE* (Fig 8A). Moreover, in the raw colour scores, we found that individuals heterozygous for *SULF* on a homozygous FLA^{st} background ($FLA^{st}SULF^{ps}/FLA^{st}sulf^{st}$) showed a stronger reduction in aurone pigments (i.e., lower yellow score, mean ± s.e. score = 0.37 ± 0.02) than those heterozygous for *FLA* on a homozygote $sulf^{st}$ background ($fla^{ps}sulf^{st}/FLA^{st}sulf^{st}$) (mean ± s.e. yellow score = 0.47 ± 0.02) (Fig S11), suggesting a stronger effect of *SULF* on colour. Finally, the interaction between the 3 loci to reduce the overall aurone production can be seen somewhat in gradual decrease in mean yellow scores on samples with genotypes that differ gradually by one *pseudomajus* allele, starting from the *striatum* background. For example, samples that are homozygous for the *striatum* in all 3 loci ($cre^{st}FLA^{st}sulf^{st}/cre^{st}FLA^{st}sulf^{st}$) had the highest yellow scores (0.69 ± 0.01) (Fig S11). Adding *pseudomajus* alleles to *CRE*, *FLA* and SULF in order gradually decreases the yellow colour score ($cre^{st}FLA^{st}sulf^{st}/cre^{st}FLA^{st}sulf^{st}$ = 0.38

± 0.07; *crestFLAstsulfst/crestFLAstsulfst* = 0.13 ± 0.03; *crestFLAstsulfst/crestFLAstsulfst* = 0.08 ± 0.02) (Fig S11). Although this gradual decrease in order does not quantitatively prove the proposed multiplicative interaction between the loci (Bradley et al. 2025), it at least further confirms the probable order in which these loci are important in shaping the yellow colour distribution. It should be noted that we are using the mean yellow colour score from the entire flower instead of a gradient in colour intensity used in Bradley et al (2025). Future work could look at the yellow scores at each part of the flower separately, to disentangle the nature of interactions more closely.

Overall, the first GWAS to understand the genetic architecture of flower colour variation that is a target of strong selection in *A. majus* recovered all (except *AUR*) the previously described major colour controlling loci, but surprisingly returned no novel minor modifier loci (e.g., like evidence of minor modifier alleles for adaptive traits like wing pattern in *Heliconius* (Papa et al. 2013)). However, the significant association in the region in between two of the most important loci—*ROS* and *EL*—prompted further dissection to test if it could potentially be a causal association to magenta colouration.

6.3.5 | Dissection of the ROS/EL locus hints at a new candidate locus controlling magenta flower colour

Large-effect loci often generate the sharpest signals of restricted gene flow, motivating a closer look at their genomic context. To further test if the significant association in the region in between *ROS* and *EL* could potentially be candidate associated with magenta colour, we performed another LMM in GEMMA where we regressed the magenta box colour scores with 18.9K SNP genotypes within ±5Mb of the *ROS/EL* locus while including the same genetic relatedness matrix computed above as random effect, and the corresponding yellow box colour scores as a fixed covariate (Fig 13). In agreement with the above LMMs (Fig 7A, 11A), we recovered the same associations at *ROS1*, *2*, *3*, *EL* and in between *ROS* and *EL* (hereafter, referred to as *MID* for being almost in the middle between *ROS* and *EL*) (Fig 13A). We then performed five LMMs where, in each, we added the highest associated SNP at the five loci while still accounting for the genetic relatedness and yellow scores as covariates. In all of these five LMMs, the association at the SNP included as fixed covariate expectedly disappeared, while the rest remained significant, but with a weaker signal of association (i.e., the strength of evidence for an association as seen by $-\log_{10}P$ value) (Fig 13B-G).

In the LMMs with SNPs from *ROS1*, *2*, or *3* included as covariates, we found that both *MID* and *EL* remained significant (with $-\log_{10}P$ values decreasing from ~150 to ~40, 4-fold decrease in signal for association), while the SNPs at *ROS* not included as covariate had their signal for association further reduced from ~150 to ~20 (i.e., 8-fold decrease in $-\log_{10}P$ values) (Fig 13B-D). Adding the highest associated SNP at *MID* as a fixed covariate also reduced their signal at both *ROS* and *EL* peaks by ~7 folds (Fig 13E). Finally, adding the highest associated SNP at *EL* as covariate in the LMM still reduced the signal at *ROS* and *MID* peak by a much smaller ~4 fold (Fig 13F). While SNPs at all associated loci except the ones included as fixed

covariates always stayed significant, their signals for association always got reduced by varying degrees (Fig 13B-F).

This reduction in statistical significance values for association could be for several reasons. First, this could reflect the overall high levels of correlation between the SNPs in the *ROS/MID/EL* region, i.e, they are in higher LD with each other compared to the genomic background. This region is only ~0.5cM wide, and 74.2% of samples have 000/111/222 genotypes for *ROS3*/*MID*/*EL*, where 0, 1, 2 refers to the ancestral homozygote, heterozygote and derived homozygote. This is expected given that 2 out of the 3 genes (*ROS* and *EL*) in this region are the known to be the main magenta-controlling genes, and hence are selected to be in the same background the confer high levels of anthocyanin pigments in the *pseudomajus* variety. Second, it is possible that the SNPs from each locus chosen to be included as covariates are not causal SNPs, although they show the highest statistical association to the magenta colour. Therefore, these SNPs and their corresponding associations are retained with lower significance values (i.e., lower $-\log_{10}P$ value). The third likely possibility is that each of these associated loci individually affect magenta colour by interacting with one another. This possibility seems likely since the associations never disappear even though they are reduced in their signals for association. This reduction reflects that the SNPs at *ROS*, *MID* and *EL* are in high LD with one another owing to their physical proximity, but the significance of the association hints at their individual effects on the magenta colouration (Note, in contrast, how peaks at chromosomes 3, 5, 8 completely disappeared when SNPs at ROS, MID, EL, and RUB were included as covariates in the LMM since SNPs at those *other* associations were only correlated to the *known* associations, and did not affect flower colour, Fig 7B-E). A final LMM which includes all the five SNPs as fixed covariates removes all significant associations, suggesting that the magenta colour variation can fully be explained by the above SNPs (Fig 13G).

Taken altogether, first, this series of LMMs verify the individual effects of *ROS* and *EL*. Second and more importantly, it hints at a third locus, *MID* in between ROS and EL, that could either act independently or in conjunction with *ROS/EL* or both to control magenta colour variation. The highest associated SNP in the *MID* peak is located 9.9Kb upstream of a gene encoding a beta-helix-loop-helix domain (bHLH) transcription factor (chr6: 52996409–52998275). While that has not been directly characterised previously, there is substantial evidence for bHLH transcription factors play crucial roles in regulating flower colour and enhancing anthocyanin biosynthesis in snapdragons as well as in other plant species (Naing et al. 2017; Sharma et al. 2020; Albert et al. 2021; Song et al. 2021). bHLH works together as part of the larger MYB-bHLH-WD40 (MBW) transcriptional complex that targets and activates the transcription of structural genes in the anthocyanin pathway (Xu et al. 2015). This regulatory mechanism is highly conserved across plant species and is essential for anthocyanin biosynthesis (Kodama et al. 2018). Therefore it explains the association peak in between the two well characterised MYB-transcription factors—*ROS* and *EL*. However, this novel candidate will require functional validation through targeted molecular genetic techniques in order to establish causality (Flint and Mackay 2009; Mackay et al. 2009).

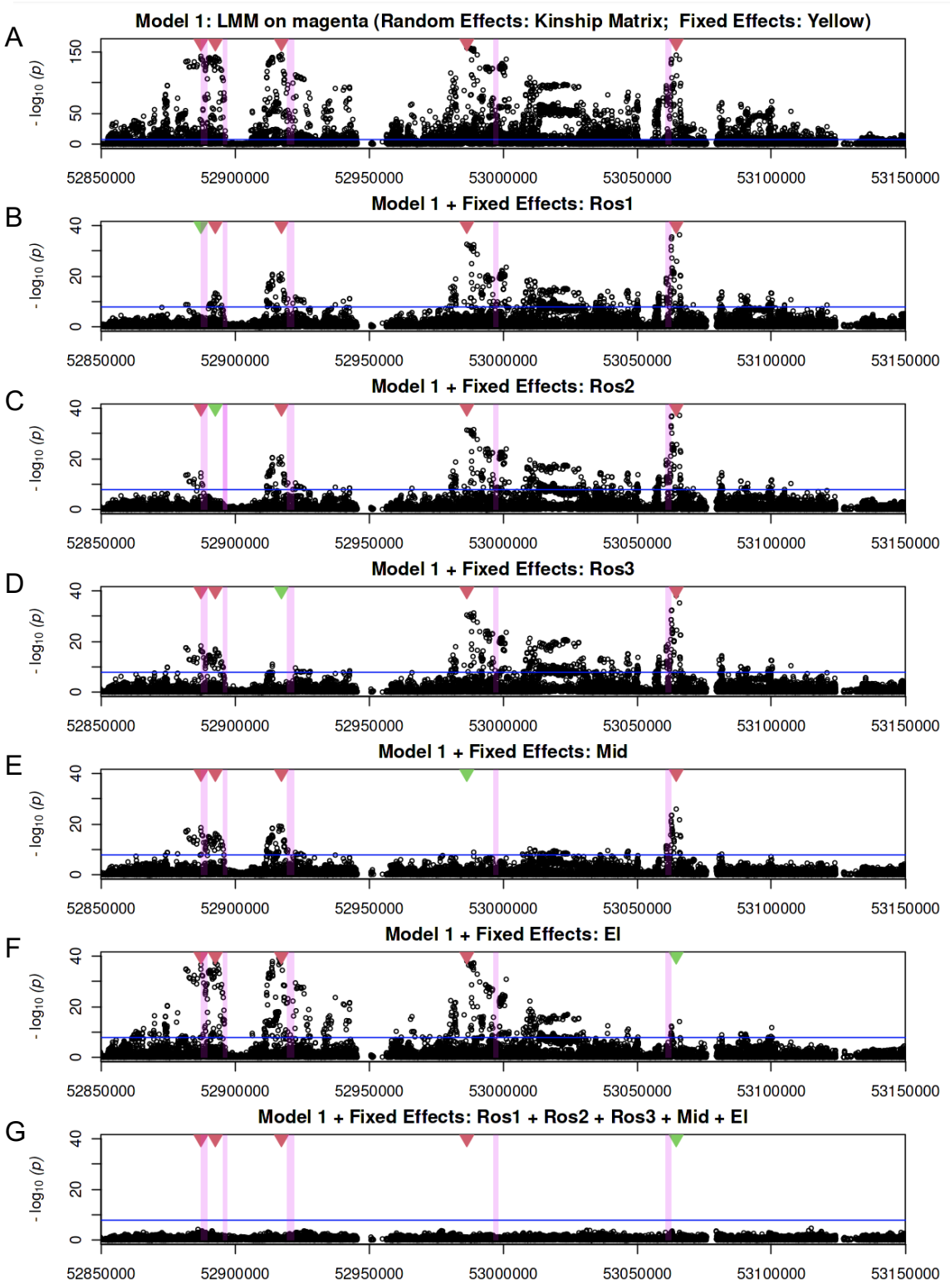

Figure 13. Conditional linear mixed model (LMM) association of *ROS/EL* region with magenta colour for samples from the Lower Road (*n* = 915). **(A)** LMM with kinship matrix as random effect and yellow colour score as fixed effect covariate. Inverted triangles denote SNPs with the highest signal (lowest *P*-value) at the five visible peaks of associations at *ROS1*, *2*, *3*, *MID* and *EL*. Each of these SNPs are individually added as fixed effect covariate in the LMM in **B** to **F** (indicated in green inverted triangle). All 5 SNPs are included as fixed effect covariates in **G**. Previously described loci and potential candidate are highlighted in pink. Left to right: *ROS1*, *ROS2*, *ROS3*, bHLH transcription factor, *EL*.

The phenotypic effect of *MID* on the genotypic backgrounds of *ROS*/*EL* is not as clearly visible as it was for *RUB/ROS/EL* (Fig S10). This is mostly because there are not enough samples with all possible genotype combinations owing to *ROS*, *MID*, *EL* being packed in a 0.5cM region. However, summing over all *ROS*/*EL* backgrounds, magenta colour scores increase with the inclusion of each MID^{ps} allele, hinting at a semidominant nature of *MID* (mean ± s.e. magenta scores for MID^{ps}/MID^{ps} = 0.02 ± 0.01; MID^{ps}/MID^{ps} = 0.24 ± 0.01; MID^{ps}/MID^{ps} = 0.69 ± 0.01) (Fig S10B). However, the more proper way to study the phenotypic effects would be introgress the MID allele into all ROS/EL backgrounds with crosses, which is beyond the scope of this thesis.

6.4 | Bottom-up approaches identify all of the colour associated loci

Having dissected the genetic architecture of flower colour, a key question emerges—do results from top-down GWAS align with signals detected by bottom-up genome-wide scans? This integration is essential for distinguishing genuine targets of selection from statistical artifacts and for studying the evolutionary processes that maintain colour polymorphisms across populations. To this end, I employed two complementary bottom-up approaches: (*i*) site/window-based summary statistics (F_{ST}, π_w, and D_{XY}) that quantify allele frequency differences and diversity patterns; and (*ii*) tree-based topological weighting that quantify clustering patterns between individuals.

For the summary statistic approach, I estimated Weir and Cockerham's F_{ST} for each SNP between individuals from the magenta and yellow flanks, and further in 10 Kb non-overlapping windows (*n*=50,881 windows) to explore the genomic landscape of differentiation between these groups. For the topology-weighted approach, I reconstructed an ancestral recombination graph (ARG) from high-coverage samples (≥5x) from the opposite flanks using *Relate* (Speidel et al. 2019). Then, for each of the 3.2 million genealogical trees along the genome, I performed topology weighting by iterative subsampling of haplotypes from each population (Martin and Van Belleghem 2017). Treating each deme from both flanks as distinct populations, I identified 9 out of 105 possible unrooted topologies in which all samples cluster strictly within their respective flank—these are reported as the "colour topology" (T_c) weight. The combined application of these two bottom-up approaches offers a robust perspective on the evolutionary mechanisms maintaining colour polymorphisms and further validates the functional relevance of GWAS-identified loci.

Across the genome, I observed minimal genetic differentiation between flanks (F_{ST} = 0.001 ± 0.02, median ± sd) and low levels of within-flank clustering (T_c weight = 0.085 ± 0.014), consistent with widespread gene flow and haplotype sharing (Fig 14) (Ringbauer et al. 2018; Tavares et al. 2018; Field et al. 2025). However, all five loci associated with flower colour exceeded this genomic background, showing elevated median F_{ST} and T_c weights (Fig. 14). Examining the genomic landscape, 109 out of 442 F_{ST} outlier windows that exceeded the 99th

percentile threshold (≥0.037) and 2,758 out of 3,212 genealogical trees with the highest 0.1% T_c weights were within ±10 kb of the colour loci (Fig. S17, S18). Both these signals coalesced into four prominent F_{ST} peaks (≥5 consecutive outlier windows spanning ≥50Kb, Fig. S17) and four high-T_c clusters (≥50 trees with outlier T_c weights spanning ≥50Kb, Fig. S18), three of which overlapped major colour associated loci: *ROS/EL*, *FLA*, and *RUB*. Notably, the largest F_{ST} and T_c signals coincided with *ROS/EL* (maximum F_{ST} = 0.85, T_c weight = 0.75) and *FLA* (maximum F_{ST} = 0.62, T_c weight = 0.43), consistent with their known major roles in magenta and yellow pigmentation, respectively. Moreover, owing to the location of *FLA* in the pericentromeric region of low recombination and *ROS/EL* containing 2 major and 1 potential candidate loci for magenta colouration, signals of F_{ST} and T_c weights show prominent peaks extended over ~650Kb around both loci. The remaining two colour associated loci, *SULF* and *CRE* exhibited weaker signatures: *SULF* fell among F_{ST} and T_c outliers but lacked a distinct peak, while *CRE* was highlighted only by F_{ST} (Fig S17, S18). These patterns mirror previous findings of minimal genome-wide differentiation coupled with pronounced barriers to gene flow at select loci in the Planoles hybrid zone, underscoring that colour-associated regions are the primary exception to otherwise high levels of gene flow between flanks (Ringbauer et al. 2018; Tavares et al. 2018; Field et al. 2025). Therefore, between the two bottom-up approaches, all 5 colour associated loci were successfully identified, validating the evolutionary as well as functional significance of these loci.

The natural follow-up question is—would we have invariably detected the spurious GWAS associations (Fig 7A, 8A) in the absence of prior knowledge on the colour associated loci. Additionally, are there other regions that we would have misidentified as regions associated with colour? To determine this, I assessed positional overlap between outlier F_{ST} windows, T_c clusters, and SNPs with significant GWA association in the same 10 Kb windows (Fig 15). Outlier GWAS windows that intersected with either F_{ST} (n = 14) or T_c (n = 11) weight outliers or both (n = 101) were all associated with flower colour. Strikingly, all 19 10Kb windows containing the spuriously associated (non-colour-associated) SNPs did not overlap with either F_{ST} or T_c outliers, demonstrating that bottom-up scans would not validate these false positives (Fig 15). However, some discordance emerged between the two bottom-up scans—out of the 34 T_c weights outliers and 266 F_{ST} outlier, 28 coincided at a prominent peak ~3Mb downstream of the *RUB* locus that was not associated with flower colour (Fig 15, S17-S18). The remaining unique outlier windows (6 for T_c weights and 238 for F_{ST}) were scattered across the genome without forming prominent peaks (Fig 15, S17-S18). This discordance in outliers highlights important differences between approaches. GWAS, on one hand, is sensitive to long range LD between SNPs that does not necessarily reflect either selection or demographic effects. On the other hand, F_{ST} scans conflate multiple evolutionary processes—including demography, gene flow, and selection—producing a noisier genomic landscape, whereas topology weighting identifies regions where individuals cluster by shared ancestry, suggesting the reuse of haplotypes that are selected.

In summary, strong convergence between bottom-up and top-down approaches validates the major colour loci as genuine evolutionary targets, while the absence of spurious

GWAS associations among bottom-up outliers provides crucial negative validation. The successful recovery of all colour loci through multiple independent approaches, coupled with effective filtering of false positives, demonstrates the power of integrative genomic analyses for distinguishing genuine selection targets from statistical artifacts.

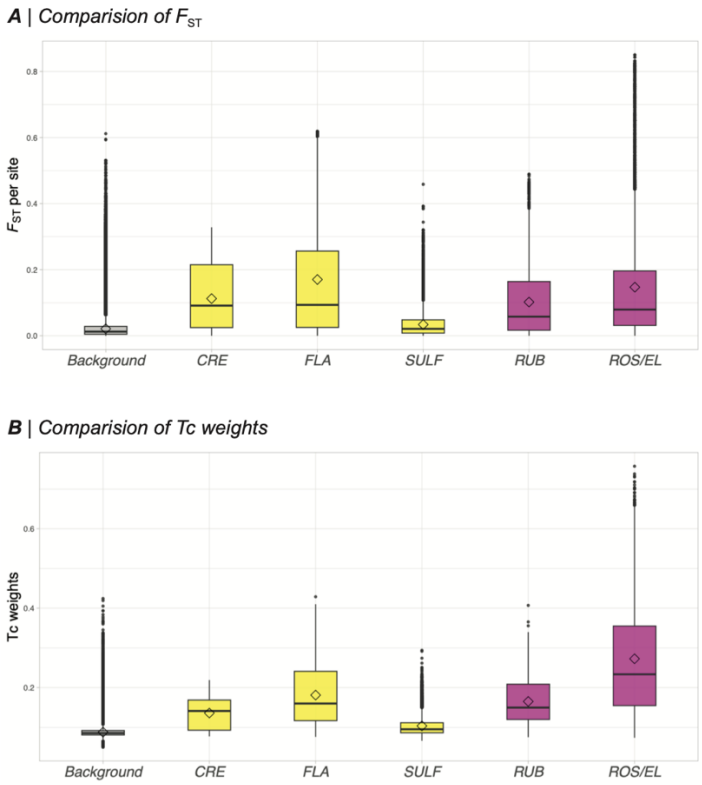

Figure 14. Comparison of bottom-up signals between colour associated loci and genomic background. Boxplots showing range of **(A)** Weir and Cockerham's F_{ST} for each SNP between samples of magenta and yellow flank and **(B)** colour topology (T_c) weight for each genealogical tree at the colour associated loci (*CRE*, *FLA*, *SULF*, *RUB* and *ROS/EL*) and background (all SNPs that are not associated to colour). Boxplots are coloured according to which colour pigment the locus is associated with; magenta for anthocyanin and yellow for aurone.

6.5 | Evidence of selection at the colour associated loci

To investigate the evolutionary mechanisms underlying colour differentiation, I further examined patterns of divergence (D_{XY}), diversity (π_w) and allele frequency differentiation between the two flanks at each colour associated locus. First, the *ROS/EL* region harboured three independent peaks of differentiation (*ROS*, *MID*, *EL*), all with striking allele frequency differentiation (≥0.8) and aligned precisely with the GWAS association peaks (Fig 11, S12). Notably, the substantial peak at *MID* further strengthens its candidacy for magenta pigmentation. While D_{XY} and π_w were highly correlated genome-wide (Spearman's *rho* = 0.99, Fig S19A), windows with elevated F_{ST} (> 0.6) within *ROS/EL* showed reduced diversity primarily in the yellow flank (correspondingly, *striatum* background) rather than increased divergence (Fig S19C), suggesting recent selective sweeps specifically on the *striatum* background. This

contrasts with previous hypotheses of sweeps on both *pseudomajus* and *striatum* backgrounds (Fig 11D, S12D) (Tavares et al. 2018), although it does not preclude an older or a soft sweep on *pseudomajus*. The asymmetric diversity reduction may reflect more recent selection on the *striatum* background, as sweep signatures erode over time through recombination and gene flow.

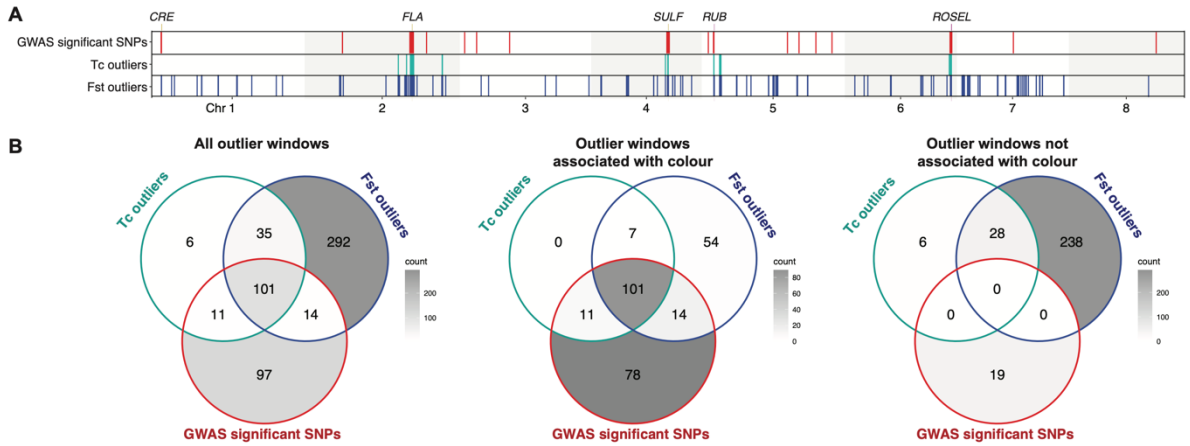

Figure 15. Comparison of signals from GWAS, F_{ST} and topology weighting. All SNPs with significant association (i.e., outliers) in GWAS and all trees with the top 0.1% of colour topology (T_c) weights are converted to the same 10 Kb non-overlapping windows on which F_{ST} is calculated. **(A)** Distribution of outlier windows along the genome. **(B)** Venn diagrams showing the overlap of outlier windows among the three analyses for all outlier windows, windows within ±1Mb to colour associated regions, and windows in regions not associated with colour.

At *SULF*—despite being the top genome-wide association for yellow—neither F_{ST} nor T_c signals were pronounced, likely due to a structural deletion in *striatum* that limits read mapping and SNP discovery (Bradley et al. 2017). Previous studies have recovered elevated F_{ST} at *SULF* only when comparing *striatum* and *pseudomajus* varieties at greater geographic distance (Field et al. 2025). Although clustering or differentiation signals were weak, the highest-F_{ST} SNP and highest-T_c tree aligned precisely with the duplicated repeat hinting at localised differentiation (Fig 12, S14).

FLA displayed an ~650 Kb region of elevated differentiation, increased divergence and a reduced diversity within the yellow flank (Fig 12, S15, S19E). The reduced π_w suggests selective sweeps on the *striatum* background, consistent with *FLA*'s role in shaping yellow gradients in the flower (Fig S19E). The increased divergence and extended differentiation likely reflects its location in a recombination desert as well as the maintenance of a selected recombinant haplotype (Fig S19E) (Bradley et al. 2025).

Lastly, *RUB* showed moderate F_{ST} elevation and allele frequency differences, with reduced diversity in *pseudomajus*, consistent with epistatic interactions with *ROS* (Fig 11, S13, S19B) (Field et al. 2025), while *CRE* displayed minimal differentiation, aligning with its modest phenotypic effect and primarily *striatum*-specific expression (Fig 12, S16, S19D) (Bradley et al. 2025).

Together, these locus-specific patterns reveal how localized selective sweeps on different genetic backgrounds, structural variation, and epistatic interactions interact with genome-wide gene flow to maintain adaptive flower colour polymorphisms in *A. majus*. The differences in diversity patterns suggest complex histories of recurrent selection events, with more recent sweeps on *striatum* at *ROS/EL* and *FLA*, and selection on *pseudomajus* at *RUB*. This comprehensive view demonstrates how integrative genomic approaches can uncover the evolutionary forces shaping adaptive variation in the face of ongoing gene flow.

6.6 | Conclusions

This chapter integrated top-down and bottom-up genomic approaches to dissect the genetic basis of flower colour variation in a hybrid zone between two varieties of *Antirrhinum majus*, confirming the role of *ROS/EL* (Schwinn et al. 2006; Whibley et al. 2006) and *RUB* (Field et al. 2025) for magenta colouration, and *SULF* (Bradley et al. 2017), *FLA* (Bradley et al. 2025) and *CRE* (Richardson et al. 2025) for yellow pigmentation, while additionally identifying a potential candidate bHLH transcription factor influencing magenta pigmentation. These findings, drawn from GWAS on 1,003 plants and population scans of differentiation and ancestry, emphasized that colour is predominantly controlled by a few large-effect loci that maintain divergence despite gene flow across most of the genome. The absence of association at *AUR*—implicated in spatial patterning of yellow pigment—likely reflects its small effect size, epistatic interactions, or limitations in our whole-flower colour quantification, where spatial averaging may have masked the resolution to detect localized gradients in pigmentation. Overall, our findings not only clarify the genetic basis of flower colour in the snapdragon system, but also offer broader lessons for evolutionary genomics, particularly regarding the strengths, limitations, and complementarity of different methodological paradigms.

Beyond dissecting the genetic basis of flower colour in snapdragons, our rigorous GWAS highlights both the power and pitfalls of association studies in natural populations — uncontrolled population structure, relatedness and correlated trait or genotypes can generate spurious peaks that disappear under appropriate statistical control (Price et al. 2006; Visscher et al. 2017). Conditional GWAS and fine-mapping proved essential for distinguishing primary causal loci from linked signals—a standard practice mirrored in previous work across systems, such as, major-effect pigmentation genes such as *Agouti* and *MC1R* in vertebrates (Linnen et al. 2009, 2013; Mallarino et al. 2016), crop adaptation genes in *Arabidopsis* (Atwell et al. 2010) and barley (Bustos-Korts et al. 2019), as well as autoimmunity-related genes in humans such as *HLA* (Raychaudhuri et al. 2012). Serial conditional analyses are particularly important when multiple loci control related traits, as stepwise approaches help isolate true causal variants from background linkage (Yang et al. 2012).

The convergence of GWAS signals with bottom-up F_{ST} outliers and ARG-based topology weighting provided orthogonal validation of the regions associated with flower colour. This integrative approach corroborated the barrier loci, but also illuminated the evolutionary dynamics by examining divergence and diversity, such as possible selective

sweeps on the *striatum* background at *ROS/EL* and *FLA*, and on the *pseudomajus* background at *RUB* warranting further investigation. Such integration overcomes the limitations of either method alone—GWAS's sensitivity to confounding and genomic scans' inability to distinguish demography from selection (Lotterhos and Whitlock 2015)—and has yielded similar success in other studies, from conifer hybrid zones where adaptive introgression drives divergence along environmental gradients (Yeaman et al. 2016) to monkeyflower speciation, where combined approaches reveal the interplay of selection and gene flow (Stankowski et al. 2023). Additionally, the lack of uniquely spurious outliers in the topology weighting highlights the value of adopting genealogical approaches in studying genomic regions under selection. This has already been demonstrated in Chapter 5 by recovering the parallel genetic basis of flower colour in replicate hybrid zones of *A. majus*, despite patterns of differentiation largely reflecting demographic effects (Pal et al. 2025).

However, in the absence of prior candidate genes, initial bottom-up genomic scans can prioritize genomic regions for targeted association testing. This strategy has been effective in studies of mammalian coat colour (Linnen et al. 2009, 2013; Mallarino et al. 2016), bird plumage (Bruders et al. 2020) and insect pigmentation (Bastide et al. 2013) among others. This contrasts with molecular genetics in controlled crosses, which excel at resolving major-effect variants but may miss context-dependent effects in hybrid zones (Manceau et al. 2010); integrating both, as in our study, provides a robust framework for linking genotypes to phenotypes under real-world selection (Linnen et al. 2013; Stankowski et al. 2023).

Overall, by revealing how large-effect barrier loci persist in the background of genome-wide gene flow, our findings provide a general framework for understanding how genetic architectures can maintain local adaptation and drive the early stages of speciation in wild systems. Looking forward, the incorporation of ARG-based approaches in speciation genomics open new avenues for dissecting the evolutionary history and ages of these barrier loci and their role in maintaining species divergence. Our findings help illustrate the power of hybrid zones as natural laboratories for studying adaptation and divergence, while emphasizing the critical importance of methodological rigor in interpreting genomic associations in these complex evolutionary contexts.

6.7 | Methods Summary

6.7.1 | Quantification of flower colour

Quantification of flower colour was done in three ways. First, for the field colour scores, each flower was visually scored in the field for magenta and yellow colouration based on the intensity and spread of the pigments (Fig S1). The magenta scores ranged from 0.5 to 5, with a score of 0.5 showing no presence of magenta pigment to intense magenta pigmentation throughout the corolla. Similarly, each flower was scored for yellow colouration, ranging from 0.5 (no yellow or yellow colour restricted to the bee entry point) to 3 (full yellow).

The two other colour scoring methods were performed on photographs taken in the field following the protocol described in Matejovičová (2022). For the box colour scoring, the front face of each flower photograph was divided into 7 boxes to cover most of the variation seen in pigmentation across the petals (Fig S2). Each box was then scored for magenta and yellow intensity on a scale of 0 to 4 (Fig S2). Finally, the total score for each flower was normalised to have a value between 0 and 1. For digital colour scores, we used *SnapPallete* (https://github.com/seanstankowski/SnapPallete) to extract colour from the images. We used manual mode (-m) in *SnapPallete* which allows the user to manually draw any number circles of varying sizes by mouse dragging. We placed the circles in the areas shown in Fig S3.

6.7.2 | Genome-wide association mapping of flower colour traits

We used *Genome-wide Efficient Mixed Model Associations* (*GEMMA*) (Zhou and Stephens 2012) to perform genome-wide association of the magenta and yellow box colour scores. We reduced the 1003 samples along the Lower Road transect to 915, that had colour box score information, and were therefore included in the GWAS. 744 of these were from the hybrid core, while rest were from the three demes on either transect (MF1, 2, 3, and YF1, 2, 3).

Linear Models

For all GWA models, the normalised magenta and yellow box scores (bounded between 0 and 1) were used as quantitative traits. SNPs were filtered with a minor allele frequency (MAF) threshold of 0.01 and linkage disequilibrium R^2 threshold of 0.9999 (i.e., essentially using all SNPs) with *PLINK v2* (Chang et al. 2015), leaving a final set of ~15.7m SNPs. We assessed SNP associations using Likelihood Ratio Test (LRT) *P*-values and considered SNPs above a conservative genome-wide threshold of $-\log_{10}P$=8.5 based on Bonferroni correction to be significantly associated. The model fit of each run was checked by visually inspecting quantile-quantile plots for expected and observed $-\log_{10}P$ values.

We first ran a linear model (LM) regressing genotype on colour score using the flag *–lm 4* in *GEMMA* (Genotypes ~ Magenta; Genotypes ~ Yellow), for both sample sets from the Lower Road (that includes the core and flanks, *n*=915) and only the Lower Core (*n*=744) (Fig 3, 4). In a second LM, we added the alternate colour score as fixed effect to account for the correlation between colour scores for the non-hybridised sample (Model II: Genotypes ~ Magenta + FixedEffect: Yellow; Genotypes ~ Yellow + FixedEffect: Magenta) (Fig 5, 6).

Linear Mixed Models and conditional GWAS

Consequently, to further control for population structure and genetic relatedness, we performed a LMM where we added kinship matrix as a random effect. From the full set (~15m), we removed SNPs with MAF<0.05, masked regions around the previously known colour genes (Field et al. 2025; Richardson et al. 2025), and finally subsampled sites with R^2<0.1 in 50-SNP windows with 10-SNP overlap, to have a reduced set of ~600k, that was then used to create the kinship matrix using the flag *–gk 1* in *GEMMA*. The LMM model regressed

genotype on colour score while controlling for the alternate colour as a fixed effect and kinship as a random factor (Genotypes ~ Magenta + FixedEffect: Yellow + RandomEffect: kinship; Genotypes ~ Yellow + FixedEffect: Magenta + RandomEffect: kinship) (Fig 7A, 8A).

Finally, to dissect the independence of these association signals, we performed conditional GWAS by incrementally adding the most strongly associated SNPs as fixed effect covariates in the LMM (Model: Genotypes ~ Magenta + FixedEffect: Yellow + RandomEffect: kinship + FixedEffect: significant SNPs; Genotypes ~ Magenta + FixedEffect: Yellow + RandomEffect: kinship + FixedEffect: significant SNPs) (Fig 7B-E, 8B-G). We also dissected the *ROS/EL* region most associated with the magenta colour, to check if the separate peaks are independent. To do that, we incrementally added each of the five peaks at *ROS1*, *ROS2*, *ROS3*, *MID* and *EL* one at a time (Fig 13).

Bayesian Sparse Linear Mixed Model

We also ran a complementary Bayesian GWAS model— Bayesian sparse linear mixed models (BSLMM) (Zhou et al. 2013)—to broadly summarize the genetic architecture of the flower colour trait and to have an independent layer of evidence for the SNPs that produce significant associations in the LMM. We ran BSLMMs in GEMMA using 10 independent chains, 25m MCMC steps after a burn-in of 5m steps, sampling every 1,000 steps. We then combined results across the independent runs and summarized the genetic architecture of each trait using posterior distributions of three main hyperparameters: the proportion of phenotypic variance explained by all SNP genotypes (PVE), the proportion of PVE explained by SNPs with 'sparse effects' (PGE), and the number of SNPs required to explain the estimated PVE (n SNPs). We further broadly quantified the architecture of each trait by calculating the sum of the posterior inclusion probability (PIP) for SNPs with detectable effects (PIP ≥ 0.01).

6.7.3 | Genome-wide scans for outlier loci

Estimating genomic differentiation, diversity and divergence

We calculated per-site F_{ST} between samples from the MF1/2/3 (*n*=75) and YF1/2/3 (*n*=99) on the full SNP dataset, using the approach described by (Weir and Cockerham 1984), implemented in *vcftools v0.1.16* using the *--weir-fst-pop* flag (Danecek et al. 2011). Site-based estimates were averaged to obtain genome-wide F_{ST} estimates and 99th percentile outliers. We also calculated F_{ST}, D_{XY}, and π_w in 50-SNP non-overlapping windows for the same set of individuals using the script *popgenWindows.py*[4]. Finally, the allele frequency difference between MF and YF was calculated using the *--freq* flag in *vcftools*.

Inference of ARG and topology weighting

We used *Relate* (Speidel et al. 2019) to infer genealogies along the genome for 151 samples from the magenta and yellow flanks, that were sequenced at ≥5x coverage. As above,

[4] https://github.com/simonhmartin/genomics_general

we assumed a mutation rate of 5.7×10^{-9}, a haploid effective population size of 813,388 (derived earlier from $\pi = 4N_e\mu$ where $\pi = 0.009$ is calculated from PoolSeq sequence data in Field et al, 2025) and the recombination map computed using *LDhat* (Auton and McVean 2007). Following the general pipeline recommended in Speidel et al., 2019 and also previously implemented in (Pal et al. 2025), we first inferred genealogies in parallel for each chromosome. We then used the *EstimatePopulationSize* script of *Relate* to jointly infer a time-varying population size history and branch lengths, using *--threshold 0* to ensure all trees are included in the joint fitting and *--num-iter 5*. Finally, we estimated the branch lengths in trees along each chromosome in parallel, using the "*RelateCoalescentRate --mode ReEstimateBranchLengths"* with coalescence rates derived from the MCMC fits in the previous step. Finally, we converted the genealogical trees stored in native *Relate* formats to *newick* format for topology weighting using "*RelateExtract --mode AncToNewick"*.

On the genealogical trees, we first weighted topologies using *TWISST* (Martin and Van Belleghem 2017). Each of the 3 demes from opposing flanks (MF1/2/3 and YF1/2/3) were treated separately, thus resulting in 105 possible unrooted topologies. For each tree, we limited the topological sampling (i.e., picking 1 sample from each of the 6 groups) to 10,000 subtrees using the flag *--method fixed* in *TWISST*. Finally, to simplify interpretation and visualisation, we only considered the 9 topologies where samples were clustered based on their location, i.e., MF1/2/3 were clustered together, and similarly, YF1/2/3 and summed their proportions. Since there is predominantly magenta coloured *pseudomajus* variety in MF and yellow coloured *striatum* variety in YF, we call the sum of these 9 topology weights, the colour topology weight, suggesting separation of haplotypes between the two flanks mostly based the flower colour. Under neutrality and panmixia, one would expect such topologies to be present at a proportion of 9/105=0.086 on average.

Definition of outlier peaks and comparison of outlier windows across methods

To identify genomic outliers from the bottom-up approaches, I applied method-specific thresholds: F_{ST} outliers were defined as 10 Kb windows exceeding the 99th percentile, while topology weight outliers comprised trees with the top 0.1% of Tc weights. Prominent peaks were determined with spatial clustering—at least 5 consecutive F_{ST} outlier windows (≥50 kb) or 50 consecutive outlier trees spanning ≥50 kb.

Cross-method comparison of outlier regions required standardization to a common genomic coordinate system. Since each approach operates at different native scales—GWAS associations at individual SNP positions, F_{ST} estimates within predefined 10 Kb windows, and topology weights calculated for genealogical trees with spans determined by local recombination rates—I converted all outlier positions to the same 10 Kb window framework used in the F_{ST} analysis. Specifically, significant GWAS SNPs and the genomic spans of outlier trees were mapped to their corresponding 10 Kb windows, enabling direct assessment of positional overlap across methodologies.

6.8 | Supplementary Information

Supplementary information related to this chapter is detailed in Appendix E.

Chapter 7
Preliminary genealogical analysis of a genomic island of speciation

Abstract

This chapter describes a preliminary analysis that uses ancestral recombination graph (ARG) inference to shed light on the origin and maintenance of a genomic island of speciation. Past work has shown that the *Rosea/Eluta* region is highly divergent across the *Antirrhinum* hybrid zone at Planoles, but separating the signatures of recent selective sweeps at multiple loci from the long-term barrier effect has been challenging owing to tight linkage within the region. Genealogical trees through the island showed shallow within-population coalescence around *ROS1/2/3, MID,* and *EL* in the yellow *striatum* individuals, consistent with recent sweeps, while the magenta *pseudomajus* individuals showed no comparable reduction. Pairwise cross-coalescence times between the two groups additionally indicated localized barriers to gene flow, strongest at *ROS1/ROS2* and weaker at *ROS3*, *MID* and *EL*, against a well-mixed genomic background. Together, these results show initial evidence of staggered sweeps on the *striatum* background and additionally a persistent barrier at *ROS1/ROS2*, shaping differentiation landscape at this island of speciation while most of the genome remains porous to gene flow across the hybrid zone. Timing estimates are preliminary and sensitive to causal SNP identification and window delimitation, motivating further detailed work in future.

7.1 | Introduction

A key component in the process of speciation is the emergence and accumulation of reproductive barriers, often in the presence of gene flow (Coyne and Orr 1998; Wu 2001; Feder et al. 2012). Studies in the past couple of decades have been devoted to learning the genetic architecture of these reproductive barriers (Ravinet et al. 2017; Wolf and Ellegren 2017). This has led to several genome-wide differentiation landscapes between natural population/species pairs, where islands of differentiation or barrier loci stand out clearly against a genomic background (Martin et al. 2013; Marques et al. 2016; Uy et al. 2016; Knief et al. 2019; Stankowski et al. 2019; Hooper et al. 2024; Nguyen et al. 2024). It has been increasingly clear that these islands of differentiation can stem from a number of differential influences of selection (e.g., selective sweeps, reduced gene flow due to a long-standing barrier, background selection) (Cruickshank and Hahn 2014; Irwin et al. 2016; Vijay et al. 2016), demographic effects (e.g. population bottlenecks, range expansions) (Bierne et al. 2013; Feder et al. 2013), or intrinsic genomic features (e.g., gene density, recombination rate variation) (Noor and Bennett 2009; Burri et al. 2015; Martin et al. 2019).

While much of the research effort has gone towards identifying these barrier loci, comparatively less attention has been towards investigating the relative contribution of different evolutionary processes and the timing of the origin and maintenance of these genomic islands of differentiation (Hejase et al. 2020*b*; Campagna et al. 2022; Wang and Coop 2022; Nguyen et al. 2024; Stankowski et al. 2024). This is key to understanding the process of speciation, because the sequence in which barriers arise and the mechanisms responsible for their establishment can profoundly influence the dynamics of speciation. For example, ecological differences, such as changes in pollinator preference or environmental conditions, can drive initial population differentiation to occupy distinct adaptive peaks separated by fitness valleys, but the extent to which these forces maintain long-term reproductive isolation will depend on the interplay between selection, drift and gene flow (Endler 1977; Felsenstein 1981; Barton and Hewitt 1985). Genetic constraints can also influence the mode and tempo of the speciation process. For example, Dobzhansky-Muller incompatibilities (DMIs), which require changes at multiple loci, often evolve gradually and may be disrupted by early hybridization that breaks down co-adaptive gene combinations (Slatkin 1975; Felsenstein 1981; Barton 1983; Orr and Turelli 2001). While single-locus barriers due to local adaptation, selective sweeps or divergent selection may start the process of early differentiation, higher-order genetic barriers only appear in later stages of speciation after multiple, interacting barriers jointly impede gene flow (Barton and Bengtsson 1986; Yeaman and Whitlock 2011; Feder et al. 2014; Ravinet et al. 2017).

While a lot of theoretical research has addressed this problem, empirical results have been few. The hybrid zone between *Antirrhinum majus* subspecies *majus*—var. *pseudomajus* and var. *striatum*—provide an excellent system to explore the history, timing and mechanism of the accumulation of barriers to gene flow that shape their genomic differentiation. These sister varieties are at an early stage of speciation with low levels of genome-wide

differentiation (F_{ST} ~ 0.003; see Chapters 4-6, Tavares et al., 2018), yet maintain distinct flower colour patterns despite hybridizing in narrow contact zones across the Spanish Pyrenees (Whibley et al. 2006). Several lines of evidence from molecular genetic studies (Schwinn et al. 2006; Whibley et al. 2006; Bradley et al. 2017, 2025; Tavares et al. 2018), F_{ST} scans (Tavares et al. 2018; Pal et al. 2025), genealogical (Pal et al. 2025) or phylogenetic tree-based analyses (Richardson et al. 2025), hybrid zone cline analysis (Field et al. 2025; Surendranadh et al. 2025) and genome-wide association studies (see Chapter 6) have characterised a handful of genomic targets of divergent selection underlying flower colour variation (*Rosea*, *Eluta* and *Rubia* for magenta anthocyanin; and *Cremosa*, *Aurina*, *Flavia* and *Sulfurea* for yellow aurone pigmentation). These flower colour differences represent alternate adaptive solutions to the same functional challenge—creating effective pollinator guides—suggesting they occupy distinct adaptive peaks in a fitness landscape (Whibley et al. 2006; Richardson et al. 2025). The magenta-with-yellow-highlight flower pattern of var. *pseudomajus* and the yellow-with-magenta-veins flower pattern of var. *striatum* both serve to direct bee pollinators to nectar rewards, but through contrasting pigmentation strategies. This system provides a rare opportunity to understand how multiple islands of differentiation accumulate and maintain alternative adaptive peaks and prevent populations from falling into intermediate fitness valleys.

One way to study this evolutionary history is to leverage ancestral recombination graphs (ARG) (Shipilina et al. 2023; Brandt et al. 2024). ARGs capture both the genealogical relationships among sampled individuals and changes to those relationships along the genome due to recombination (Hudson 1990; Wakeley 2009), enabling inference about when and how genetic barrier loci arose and shaped the differentiation landscape through time. By contrasting patterns of genealogical trees around islands of differentiation with neutral genomic regions, we can test if genealogies at flower-colour loci show signals of long-term barriers to gene flow that maintain distinct adaptive peaks or selective sweeps that establish alternative adaptive solutions or an interplay of both. Finally, it also allows us to reconstruct the relative timing of these differentiation peaks in the course of speciation.

Given the low genome-wide differentiation in *Antirrhinum*, genealogies at the neutral regions are expected to show extensive paraphyly across opposite flanks due to ongoing gene flow that homogenises much of the genome. By contrast, peaks of elevated differentiation at flower colour loci can arise through two different evolutionary processes with distinct genealogical signatures (Barton 1998; Hejase et al. 2020*b*; Campagna et al. 2022; Wang and Coop 2022; Nguyen et al. 2024). First, if the loci associated with flower colour are driven to high frequency by a recent selective sweep, genealogies at and around the selected site will exhibit a star-like phylogeny with a 'burst' of coalescence events within the population (or the flank) carrying the swept allele (Maynard Smith and Haigh 1974). This will sharply reduce the time to the most recent common ancestor within the population carrying the swept allele ($TMRCA_w$), and genealogies would tend towards monophyly in the swept population due to lack of time for migration between populations following the sweep. However, $TMRCA_w$ is extremely sensitive to the nature and timing of the sweep, and dominance of the beneficial

allele. Therefore a more robust measure is *TMRCAH* (the *TMRCA* for the first 50% of all lineages from the species) (Rasmussen et al. 2014), or simply, the median pairwise coalescence times within a population (T_w) (Wang and Coop 2022; Pal et al. 2025). Both are expected to be reduced around loci under local sweeps without a necessary increase in cross-coalescence times. Alternatively, selection against an allele on the opposite genetic background reduces effective migration near the locus, creating a localized barrier to gene flow. In such regions, cross-population coalescence events are rarer and shifted deeper in time relative to neutral regions while within-population coalescence is largely unchanged. Consequently, genealogies around such loci should show increased cross-coalescence time between populations (T_b) without a necessary reduction in coalescence rates within population (T_w). Prolonged persistence of the barrier can further elevate pairwise divergence as deeper between-population coalescent times accumulate, but this typically requires a long-standing barrier of gene flow (Charlesworth et al. 1997).

However, in practice, these processes may not always act in isolation. A peak of elevated differentiation may result from both a population specific sweep as well as a barrier to migration. For example, an allele can rapidly sweep through a population (reducing within-coalescence times, T_w) and subsequently be maintained as a spatially balanced polymorphism that prevents introgression (increasing cross-coalescence times, T_b). Conversely, an allele that serves as a barrier to gene flow could have also hypothetically drifted up in frequency or simply be old enough that any trace of a sweep has eroded away. Lastly, It should be noted that these ARG-based quantities map naturally onto conventional statistics (Ralph et al. 2020), but provide greater resolution by leveraging correlated, fully resolved genealogies and the full distribution of coalescent times, while more naturally accommodating background selection and mutation-rate heterogeneity (Hejase et al. 2020*b*)

In this chapter, I present preliminary results from exploring patterns of genealogical trees around one of the flower colour associated regions—*ROS/EL*. This ~650Kb island of divergence (F_{ST} ~ 0.8 between samples from opposite flanks, see Figure S12 in Chapter 6) harbours previously characterised MYB transcription factors, *ROS1*, *ROS2*, *ROS3* and *EL*, as well as a possible candidate (referred to as *MID*) bHLH transcription factor, all of which together form the strongest associated region for magenta pigmentation in *Antirrhinum* (see GWAS results in Chapter 6). Harbouring 5 loci in this narrow region, that have all been shown or proposed to individually affect flower colouration and also shows high differentiation, presents the perfect empirical case to explore how the tension between selection, gene flow and recombination in the early stages of speciation has given rise to these multiple barriers and their relative time of emergence. To this end, I reconstructed the ancestral recombination graph for samples with ≥5x sequence coverage from both magenta (n = 72) and yellow flanks (n = 79) using *Relate* (Speidel et al. 2019) following the pipeline described in Chapters 5 and 6. To focus here specifically on the ROS/EL region, I summarized the minimum, median, and maximum estimates of coalescence times for all sample pairs within each flank ($T_{w,MF}$, $T_{w,YF}$) and between the opposite flanks ($T_{b,MF/YF}$) in each tree within the 1 Mb region centered around ROS/EL (chr 6: 52.45–53.45 Mb). To further establish neutral expectations for the

genomic background, I identified 10 additional 1 Mb regions distributed across the genome, each located at least 5 Mb from both chromosomal ends. These neutral background regions were selected such that they neither showed any association with flower colour nor outliers in F_{ST} and topology weighting analyses (see Chapter 5). I again summarised the coalescence times for trees in these regions in the same way as above, to compare them against those from the *ROS/EL* region.

7.2 | Preliminary Results and Discussion

In line with expectations, the neutral background regions of the genome showed extensive mixing between the magenta and yellow flank, as seen by their very recent minimum pairwise coalescence (minimum $T_{b,MF/YF}$, i.e., time to the first coalescence event between magenta and yellow flank) between the two flanks (83 ± 32 generations ago) and high correlation between the median pairwise coalescence times within each flank (median $T_{w,MF}$ vs. median $T_{w,YF}$, Spearman's *rho* = 0.88) (Fig S1A). This aligns with high levels of gene flow for most parts of the genome across the hybrid zone (Ringbauer et al. 2018; Tavares et al. 2018).

By contrast, there was shallow coalescence within each flank observed in the *ROS/EL* region. Around all 5 loci implicated in controlling magenta pigmentation (*ROS1*, *ROS2*, *ROS3*, *MID*, *EL*), median pairwise coalescence times dipped substantially within samples from the yellow flank ($T_{w,YF}$) compared to those within the magenta flank ($T_{w,MF}$) (Fig 1A, S2A). Median $T_{w,YF}$ was substantially lower around the 5 magenta associated loci compared to the randomly sampled background genomic blocks, that is indicative of selective sweeps. Median $T_{w,MF}$, however, did not show any significant deviation (Fig 2A). Moving beyond the 5 loci, $T_{w,MF}$ and $T_{w,YF}$ in the broader 1Mb region around *ROS/EL* were coincident with each other, and showed levels of paraphyly between the flanks similar to the patterns in the neutral regions (Fig S2A, see trees *i* and *vi* in Fig S3). The minimum pairwise coalescence time between samples from the opposite flanks ($T_{b,MF/YF}$) were surprisingly similar across the *ROS/EL* region (including around the 5 magenta associated loci) compared to the background region (Fig 1A, 2A, S1A, S2A).

Minimum $T_{b,\,MF/YF}$ indicates the time to the first migration event between the magenta and yellow flank in the past. While the median $T_{b,\,MF/YF}$ need not be affected by a selective sweep, minimum $T_{b,\,MF/YF}$ is expected to be elevated since there may not have been enough time for cross-coalescence if the sweep is really recent (Hejase et al. 2020*b*; Wang and Coop 2022). If the locus additionally acts as a long standing barrier, the median $T_{b,\,MF/YF}$ would also be elevated compared to the neutral genomic background (Charlesworth et al. 1997).

Since there is evidence of recent introgression and long distance migration in the *Antirrhinum* hybrid zone (Surendranadh et al. 2025), as well as *ROS* and *EL* both being semidominant (Schwinn et al. 2006; Whibley et al. 2006; Tavares et al. 2018), I regrouped the samples based on their genotypes at the 3 most associated SNP around the loci *ROS3*, *MID* and *EL* (see Chapter 6 for GWAS).

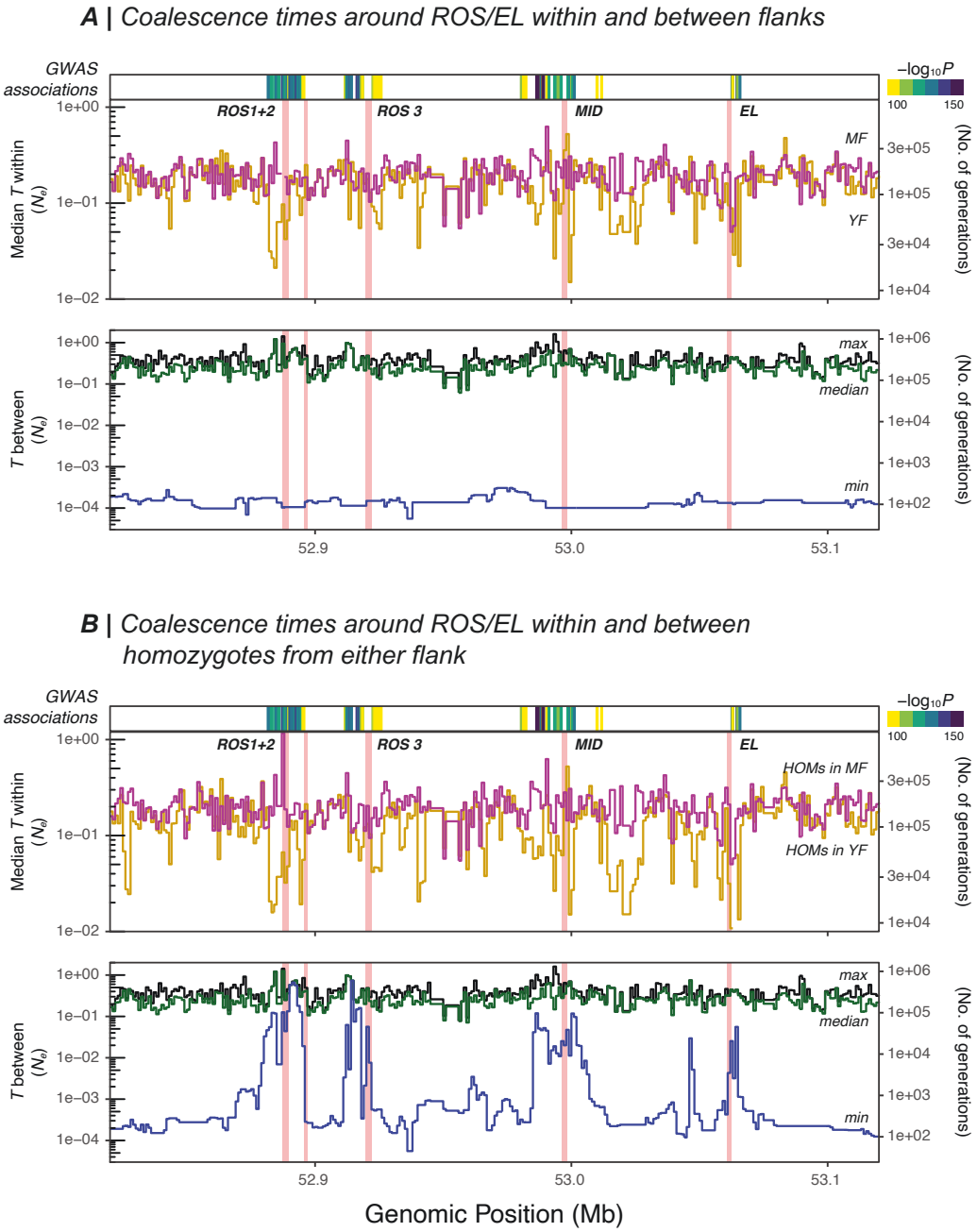

Figure 1. Coalescence history around *ROS/EL* within and between **(A)** all individuals from magenta and yellow flank, and **(B)** individuals homozygote for the *pseudomajus* background from magenta flank and individuals homozygote *striatum* background from yellow flank. Top plot in both **(A)** and **(B)** shows median pairwise coalescence times within respective groups ($T_{w,MF}$ in magenta and $T_{w,YF}$ in yellow in **A**; $T_{w,ps}$ in magenta and $T_{w,st}$ in yellow in **B**), while bottom plots show the minimum (in blue), median (in green) and maximum (in black) pairwise coalescence times between the respective groups ($T_{b,MF/YF}$ in **A**, $T_{b,ps/st}$ in **B**). Coloured bars on top represent $-\log_{10}P$ values for SNPs in GWAS for magenta colouration. For all plots, coalescence times are denoted in terms of effective population size (N_e) in left y-axis and number of generations in right y-axis. MF: magenta flank, YF: yellow flank, *ps*: *pseudomajus* background: $ROS^{ps}MID^{ps}el^{ps}$ / $ROS^{ps}MID^{ps}el^{ps}$, *st*: *striatum* background: $ros^{st}mid^{st}EL^{st}$ / $ros^{st}mid^{st}EL^{st}$.

The two new groups of samples include—$ROS^{ps}MID^{ps}el^{ps}/ROS^{ps}MID^{ps}el^{ps}$ or the *pseudomajus* alleles in the magenta flank and $ros^{st}mid^{st}EL^{st}/ros^{st}mid^{st}EL^{st}$ or the *striatum* alleles in the

yellow flank—allowing the comparison of coalescence rates within and between non-recombinant haplotypes from each flank. Coalescence times within and between the new groups of samples did not change in the genomic background regions, as expected from ongoing gene flow (Fig S1B). However, the median pairwise coalescence times within samples with homozygote *striatum* alleles in the yellow flank ($T_{w,st}$) showed a similar yet larger dip than before, around each of the 5 loci compared to that within the magenta flank as well as the genomic background (Fig 1B, 2B). Samples with the *pseudomajus* background in the magenta flank did not deviate from the genomic background (Fig 1B, 2B). This is indicative of multiple recent selective sweeps on the *striatum* background, visible clearly in genealogical trees with long branches leading to shallow coalescences within homozygotes in the yellow flank (see trees *ii*, *iii*, *v*, *vi* Fig S3). In contrast, lineages ancestral to the homozygotes from magenta flank go much deeper to coalesce to their most recent common ancestor.

To roughly date the sweep on the *striatum* background, I considered marginal trees from around the SNP most associated in GWAS, delineated visually to include trees that contain SNPs significant in GWAS and outliers for both F_{ST} and topology weighting (see Chapter 6). The median pairwise $T_{w,st}$ around *EL* were on average the lowest (~30K generations), followed by the 3 *ROS* loci (57–84K generations) and *MID* (~105K generations) (Table 1, Fig 2). Ideally, for a hard sweep, the $TMRCA_w$ should be considered as the lower bound for a sweep age (Hejase et al. 2020*b*; Wang and Coop 2022). Here, as a preliminary result, I only reported the median pairwise $T_{w,st}$ since $TMRCA_{w,st}$ does not show a substantial decrease around the loci compared to neutral background as is expected in a sweep. This could be due to two reasons. First, the SNPs most associated to flower colour that are chosen to group the samples based on haplotypes linked to colour, may not be causal. Therefore, there is still introgression of lineages into the opposite background, that increases the $TMRCA_{w,st}$, masking the estimate of the lower bound at which the causal allele was swept. Second, more work is needed to robustly delineate the boundaries around each gene/locus that should be considered to estimate the sweep ages. However, a significant decrease in the median pairwise coalescence within a population is suggestive of a sweep, and thus, is reported as a preliminary age. Although this age should be treated with caution, the relative timing of events is more robust and exhibits the heterogeneity in when sweeps occurred at the different loci.

Looking at cross-coalescence between the new groups of samples, the minimum pairwise coalescence times ($T_{b,ps/st}$) were significantly elevated around each of the 5 loci (deeper compared to the genomic background) (Fig 1B, 2B), in contrast to the same statistic measured between all samples from either flank (Fig 1A, 2A). Increase in minimum $T_{b,ps/st}$ may indicate a barrier to gene flow but could also be a consequence of selective sweeps in one or both the populations. A sweep results in rapid shallow coalescence within the population carrying the swept allele (Barton 1998), thus reducing it to a few remaining ancestral lineages that are older than the sweep and are free to cross-coalesce (Hejase et al. 2020*b*; Wang and Coop 2022). To distinguish between genuine barriers to gene flow and a conflated signal from

selective sweeps, Wang and Coop (2022) have suggested focussing on pairs of lineages, one from each population, in order to circumvent the effects of reduced ancestral lineages.

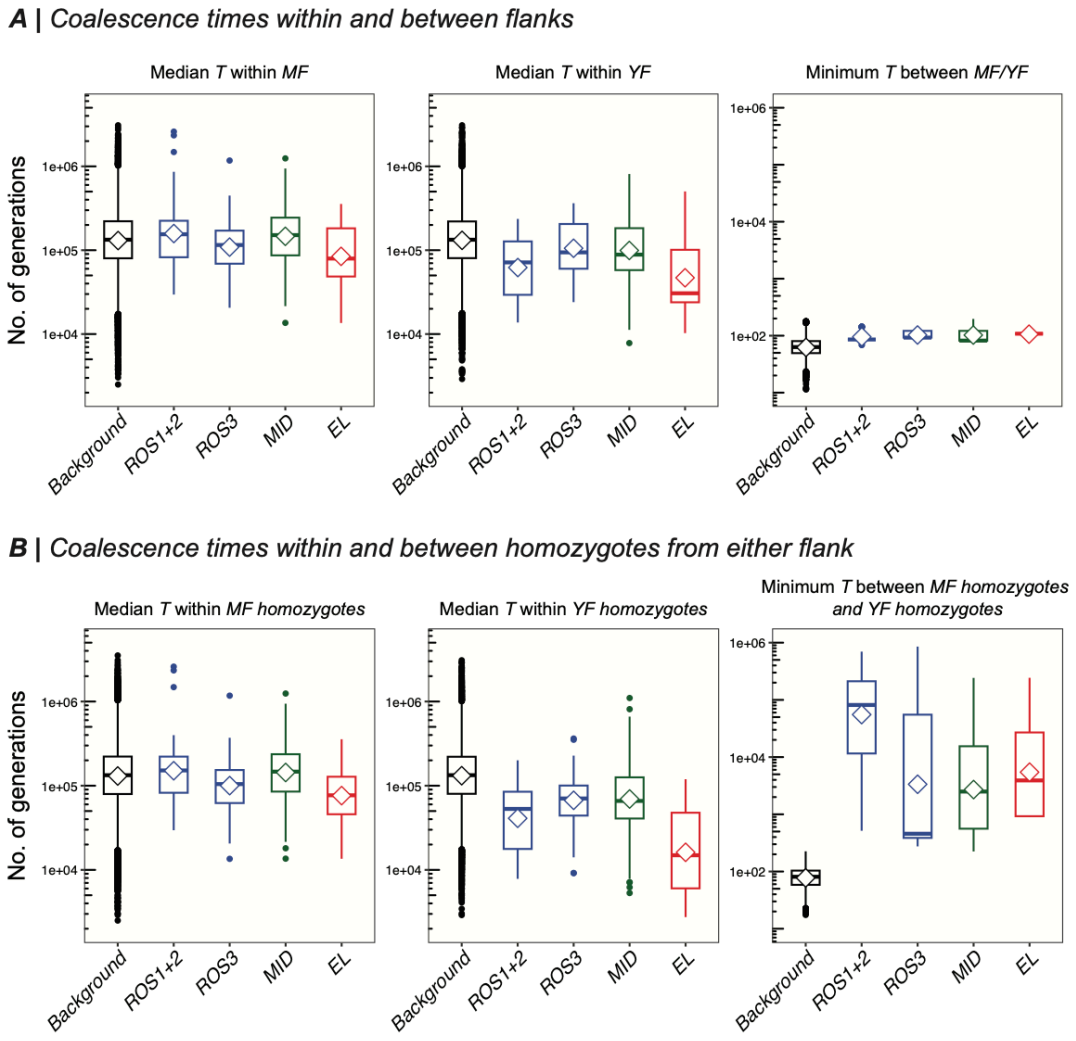

Figure 2. Comparison of **(A)** median pairwise coalescence times within magenta flank ($T_{w,MF}$), within yellow flank ($T_{w,YF}$) and minimum coalescence time between magenta and yellow flank ($T_{b,MF/YF}$), **(B)** median pairwise coalescence times within individuals homozygote for *pseudomajus* background in magenta flank ($T_{w,ps}$), individuals homozygote for *striatum* background in yellow flank ($T_{w,st}$) and minimum coalescence time between the two groups ($T_{b,ps/st}$). Comparisons are shown for the genomic background (10 randomly chosen 1Mb genomic blocks that did not show an outlier in GWAS, F_{ST} or topology weighting), *ROS1/ROS2*, *ROS3*, *MID* and *EL*. MF: magenta flank, YF: yellow flank, *ps*: *pseudomajus* background: $ROS^{ps}MID^{ps}el^{ps}$ / $ROS^{ps}MID^{ps}el^{ps}$, *st*: *striatum* background: $ros^{st}mid^{st}EL^{st}$ / $ros^{st}mid^{st}EL^{st}$.

For computational efficiency, I randomly chose 10 haploid samples from both the *pseudomajus* background in the magenta flank and *striatum* background in the yellow flank, resulting in 100 pairwise comparisons. By tracking the coalescence times of all 100 between-population pairs of lineages, the reduced number of ancestral lineages was clearly visible in the traces of each pairwise $T_{b,ps/st}$ and their lower variance around *ROS/EL* compared to the background (Fig 3). For each marginal tree in *ROS/EL*, I further estimated the proportion of pairwise $T_{b,ps/st}$ that is an outlier compared to the neutral expectation i.e., above the 95th

percentile for all pairwise cross-coalescence times in the neutral background (~350K generations), which should not be affected by selective sweeps (Charlesworth et al. 1997). There was a general enrichment in the proportion of pairwise $T_{b,ps/st}$ that were outliers in the *ROS/EL* region (see coloured bar above Fig 3A), but the strongest signal is found around *ROS1* and *ROS2* where >80% of the pairwise cross-coalescence times (median = ~467K generations) were outliers compared to genomic background (median = ~170K generations) (Table 1, Fig S4), suggesting a barrier around *ROS1* and *2* may have approximately persisted for ~297K generations. *ROS3*, *MID* and *EL* also show elevation in pairwise $T_{b,ps/st}$ (*ROS3:* ~293K, *MID:* ~273K, *EL:* ~323K generations), hinting at weaker and younger barriers that lasted for ~103–153K generations (Table 1, Fig S4, see trees *ii*, *iii*, *v*, *vi* in Fig S3).

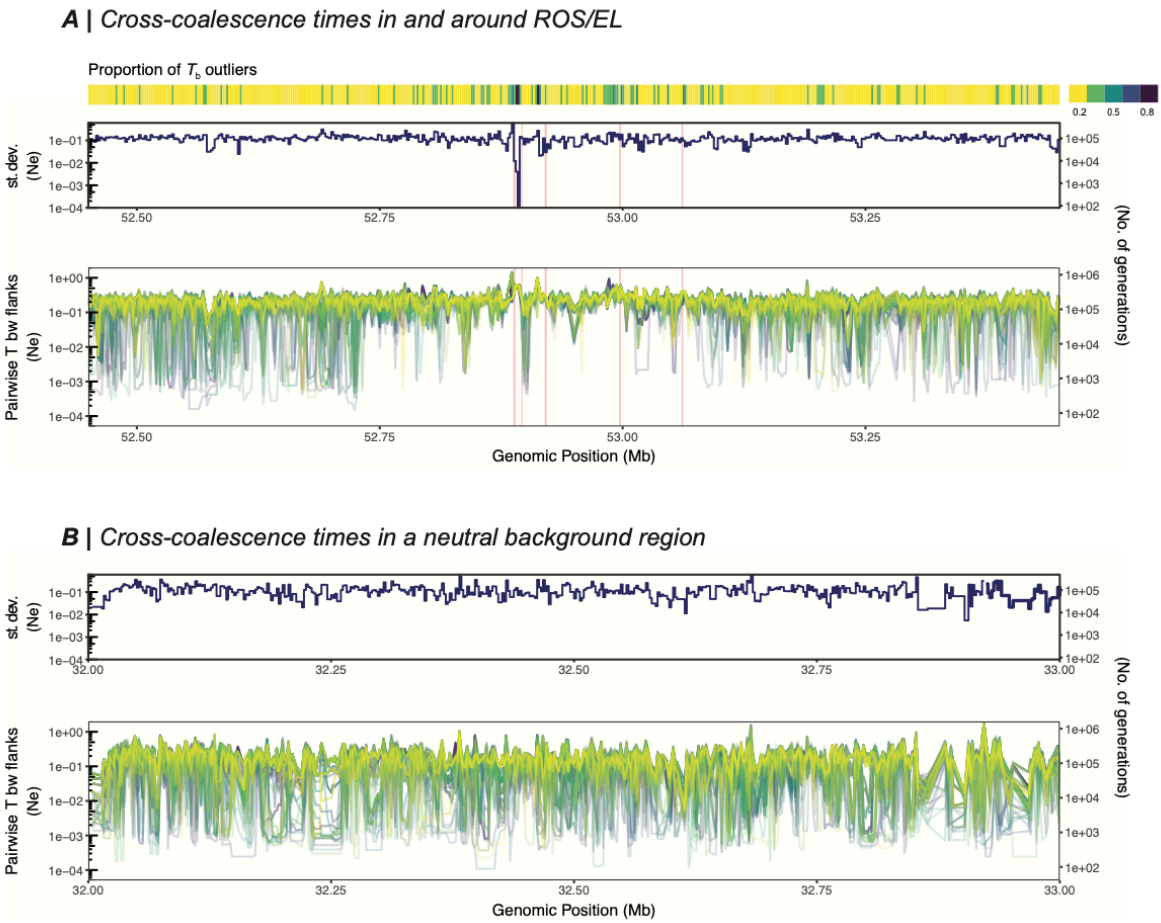

Figure 3. Traces of pairwise cross-coalescence times around **(A)** *ROS/EL* and **(B)** neutral background for a subsample of 10 haploid samples with non-recombinant genomic background (based on alleles at *ROS3/MID/EL*) from either flank. In both **A** and **B**, top plots shows the standard deviation in pairwise cross-coalescence times ($T_{b,ps/st}$) and each line in the bottom plots show the trace of coalescence time between one pair of non-recombinant individual from either flank. Coloured bar above top plot in **A** denotes the proportion of pairwise $T_{b,ps/st}$ around *ROS/EL* that is an outlier based on the 95th percentile of pairwise $T_{b,ps/st}$ in the neutral background region.

Table 1. Summary of median pairwise coalescence times within and between individuals, homozygote for the *pseudomajus* background in magenta flank and individuals homozygote for the *striatum* background in yellow flank. $T_{w,ps}$, $T_{w,st}$ values correspond to Fig 2, and $T_{b,ps/st}$ to Fig S4. *ps*: *pseudomajus* background: $ROS^{ps}MID^{ps}el^{ps}$ / $ROS^{ps}MID^{ps}el^{ps}$, *st*: *striatum* background: $ros^{st}mid^{st}EL^{st}$ / $ros^{st}mid^{st}EL^{st}$.

Genomic Regions	Median pairwise coalescence times		
	$T_{w,ps}$ *(mean, median, range)*	$T_{w,st}$ *(mean, median, range)*	$T_{b,ps/st}$ *(mean, median, range)*
Background	175.7×10^3 133.2×10^3 $2.5 \times 10^3 - 3.52 \times 10^6$	174.2×10^3 133.3×10^3 $(2.9 \times 10^3 - 3.09 \times 10^6)$	169.7×10^3 134.5×10^3 $(2.9 \times 10^3 - 2.65 \times 10^6)$
ROS1 + ROS2	260.6×10^3 151.8×10^3 $(29.5 \times 10^3 - 2.59 \times 10^6)$	57.1×10^3 53.2×10^3 $(7.8 \times 10^3 - 0.20 \times 10^6)$	467.3×10^3 335.3×10^3 $(50.5 \times 10^3 - 2.59 \times 10^6)$
ROS3	133.4×10^3 104.5×10^3 $(13.5 \times 10^3 - 1.17 \times 10^6)$	84.3×10^3 70.5×10^3 $(9.2 \times 10^3 - 0.36 \times 10^6)$	292.5×10^3 201.0×10^3 $(30.6 \times 10^3 - 1.17 \times 10^6)$
MID	183.0×10^3 146.9×10^3 $(13.6 \times 10^3 - 1.24 \times 10^6)$	105.4×10^3 65.9×10^3 $(5.3 \times 10^3 - 1.10 \times 10^6)$	273.4×10^3 203.6×10^3 $(27.9 \times 10^3 - 1.71 \times 10^6)$
EL	101.5×10^3 77.0×10^3 $(13.5 \times 10^3 - 0.36 \times 10^6)$	29.5×10^3 15.5×10^3 $(2.7 \times 10^3 - 0.12 \times 10^6)$	323.2×10^3 299.0×10^3 $(84.8 \times 10^3 - 0.79 \times 10^6)$

7.3 | Conclusions

In this chapter, preliminary genealogical analysis suggests that multiple, staggered selective sweeps and localized barriers to gene flow have shaped the genomic island of differentiation at *ROS/EL* during early *Antirrhinum majus* speciation. All five loci in the region show evidence of sweeps, but their timing is heterogeneous. As discussed earlier, the times reported above are to be treated with caution, while the relative timing of events is more robust. With that caveat in mind, individuals with the *striatum* background seemingly underwent the most recent sweep at *EL* at the least ~30K generations ago ($T_{w,st}$ at *EL*, Fig 2B, Table 1), with older sweeps at *ROS* and *MID* at least 57K generations ago ($T_{w,st}$, Fig 2B. Table 1). *Such patterns invite synthesis in the concluding reflections and in the broader literature that follows.*

Moreover, minimum cross-coalescence between the haplotype backgrounds is elevated around each of the five loci, but tracking of pairwise cross-coalescence times reveals an overrepresentation of unusually deep cross-coalescences near *ROS1/ROS2*, with weaker enrichments at *ROS3*, *MID*, and *EL*, indicating a persistent local barrier that is strongest at *ROS1/ROS2* and plausibly younger or weaker at the other three targets. The barrier around *ROS1* and *ROS2* is at least as old as when the first coalescence event between the two backgrounds can be seen at 81.5K generations ago ($T_{b,ps/st}$ at *ROS1+2*, Fig 2B). In genealogical trees, reciprocal monophyly of haplotypes at *ROS1/ROS2* contrasts with predominantly

striatum-side monophyly at *ROS3/MID/EL*, reinforcing the view that barrier effects are most entrenched near *ROS1/ROS2* and comparatively recent or more permeable at the other peaks (Fig S3).

The evolutionary history of the *ROS/EL* region has been previously studied by comparing simulations with F_{ST} peaks estimated from pooled sequence data in Tavares et al. (2018). It was suggested that the differentiation has been shaped by two processes – (*i*) historic selective sweeps that led to different *ROS* and *EL* alleles becoming fixed in *pseudomajus* and *striatum* varieties, and (*ii*) selection against hybrid genotypes generated when the two varieties met, creating a local barrier to gene flow. Broadly, estimates from coalescence times of the genealogical analysis align with that of the most recent sweep around ~90K generations ago in Tavares et al. (2018). Similarly, the preliminary estimate of the age of the barrier is also consistent, reported to be as old as 100K generations. However, unlike the proposed sweeps in both *pseudomajus* and *striatum* background in Tavares et al. (2018), we found sweep signatures only on the *striatum* background; however, a lack of signal on the *pseudomajus* background signal does not exclude historical sweeps, as recombination could have erased genealogical traces over time.

These preliminary findings suggest a dynamic process of barrier accumulation and sweeps over a prolonged period in the order of 10^5 generations. It also highlights the power of genealogical analysis in disentangling historical processes of selection and gene flow, yet several important questions remain. The interpretations reported here—especially regarding the sweep and barrier timing—are preliminary and would benefit from robust simulations to tease apart confounding effects of recombination, selection, and demography. Furthermore, while the *ROS/EL* region provides a compelling case study, the broader genomic landscape of differentiation awaits investigation, particularly at loci underlying yellow pigmentation and other adaptive traits. With further integration of empirical patterns, simulations, and genome-wide analyses, we hope to reconstruct the sequence and tempo of adaptive events that drove the divergence of ancestral *Antirrhinum* and to clarify the evolutionary paths that maintained these striking flower colour ecotypes in the face of ongoing gene flow. Such a synthesis will illuminate not just when and how barriers arose, but also the genetic and ecological constraints guiding adaptive walks between fitness peaks.

Chapter 8
General Discussion

This thesis investigates the genomic consequences of selection and population structure in natural systems experiencing ongoing gene flow. Below I discuss the broader conclusions, limitations and future directions.

8.1 | From SNPs to haplotypes to genealogies

A central question in evolutionary biology is—how do we infer the origins, mechanisms, and persistence of genomic barriers when populations never fully isolate, and the legacy of demography and selection interact across timescales (Coyne and Orr 1998; Wolf and Ellegren 2017). While much empirical analyses make inferences based on single-site statistics, the prudent path forward is the move beyond SNPs and arbitrary genomic windows to a perspective where haplotypes—the actual, recombined segments of ancestry shared among individuals—become the unit of analysis (Hejase et al. 2020*a*; Shipilina et al. 2023; Lewanski et al. 2024). This thesis first sets out to discuss the fundamental structure of haplotypes—inherited from parents to offspring. By formalizing haplotype blocks as edges within the ancestral recombination graph (ARG), this thesis offers a conceptual bridge between the way genomes evolve (per linked segments) and how they are typically analyzed (per SNPs) (Chapter 2). ARGs encode the full genealogical history of sampled sequences, representing the complex interplay of mutation, recombination, and coalescence as a series of local trees that change topology at recombination breakpoints, a view now recognized as important for disentangling the timing, directionality, and genomic context of divergence (Hudson 1990; Brandt et al. 2024).

Because edges in a sequence of trees overlap, nest, and vary in length (i.e., time-depth), they provide a natural unit for testing whether peaks of differentiation reflect recent sweeps, older barriers, background selection, or some complex mixture—processes that may be difficult to disentangle with site-based outlier scans alone. By demonstrating the richness of information contained within ARGs, Chapter 2 suggests a framework in which the span and depth of haplotype sharing among unique sets of individuals could be a possible structure for empirical analysis. It also presents an empirically tractable way to identify such edges of interest from ARGs, and infer processes, such as selective sweeps. While Chapter 2 lays out the conceptual foundation and an empirical test implementation, the debate about whether we can learn anything more with ARGs is far from over. The rest of the thesis uses the latest methodological advances, both in sequencing techniques and ARG inference, to test the

superiority of ARGs over traditional population genetic analysis, while also disentangling processes that influence speciation.

8.2 | Advances in methods for haplotype reconstruction and ARG inference

A second contribution of this thesis is methodological. Chapters 3–7 show that population-scale ARG inference is now feasible beyond model organisms by pairing linked-read haplotagging (Meier et al. 2021) with a validated pipeline that delivers phased haplotypes at manageable cost, without relying on large genome panels. The pipeline integrates reference-free imputation (Davies et al. 2016), barcode-aware alignment (Shajii et al. 2018*a*), and hybrid molecular and statistical phasing (Edge et al. 2017; Hofmeister et al. 2023) with empirical checks on genotype error and switch error rates. This design is particularly valuable where long-read whole-population sequencing remains impractical, yet accurate phase is essential for inference of genealogies, selection, and recombination landscapes (Ebert et al. 2021; Watowich et al. 2025).

Chapter 4 presents cautious optimism on the benefits of low-coverage linked-read sequencing, by highlighting the limits to genotype imputation and phasing accuracy. We have shown that at the expense of a small (~7%) accuracy loss, we gain a substantial fraction (~36%) of otherwise missed genetic information in >1000 samples (see Table 5 in Chapter 4). This is useful for an analysis, such as GWAS, or even windowed F_{ST} scans. But, researchers may consider the highest quality phaseblocks to be used in genealogical inference, where each phase switch will alter the genealogical tree, and introduce spurious recombination breakpoints (Deng et al. 2021; Brandt et al. 2022). Rare variants that are missed or misimputed in low coverage add limited information in a genealogical analysis; so smaller high-quality subsets can outperform larger error-prone genomic datasets, unless rare-variant association is a primary objective (Zhang et al. 2023). Looking forward, building a species-specific reference panel and integrating haplotype-resolved assemblies with pangenome references can reduce switch errors, improve rare-variant imputation, and capture structural variation and repetitive diversity that confound linear-reference analyses (Igolkina et al. 2025). However, depending on the goal of sequencing, a set of ~100 KASP markers can also serve the purpose (e.g., identifying parent-offspring trios to directly estimate fitness), but we showed that multiple independent runs of imputation can be used to estimate variant-calling and imputation accuracy. Hopefully, the imputation–phasing framework and accuracy metrics established here will provide a template for future sequencing design and analysis in population-scale studies of non-model organisms.

The thesis further benchmarks a suite of genealogy inference methods—including window-based trees, *tsinfer* (Kelleher et al. 2019), *Relate* (Speidel et al. 2019), and *Singer* (Deng et al. 2024) in Chapter 6—showing that while algorithmic choices affect the density and local resolution of trees, they converge on stable, biologically interpretable topology-weights

(Pal et al. 2025). This methodological pragmatism is crucial—it means that scalable, approximate approaches can recover robust signals of genealogical structure, even if fine-scale uncertainty remains. The convergence of different methods on shared topology-weighting profiles—despite their varying assumptions—builds confidence that genealogical summaries such as topology weights and coalescence-time profiles capture biological reality, not just inference artifacts. These advances collectively lower the barrier to ARG-based analysis in natural populations and add to the ever growing empirical work in non-model systems (Hejase et al. 2020*b*; Campagna et al. 2022; Wang and Coop 2022; Nguyen et al. 2024; Stankowski et al. 2024; Pieszko et al. 2025)

8.3 | Empirical case studies highlight the use of ARGs

This thesis uses two empirical study systems as test cases for using ARGs. First, in the evolution of live-bearing in *Littorina* snails, ARG-aware topology weighting reveals a polygenic architecture for this key innovation, with many narrow, reproductive mode associated genomic regions marked by selective sweeps whose ages span approximately 20K to 200K generations (Chapter 3) (Stankowski et al. 2024). These regions are enriched for tissue-specific expression differences and are notably independent of large chromosomal inversions, challenging the expectation that major phenotypic shifts require single, large-effect mutations or structural rearrangements. Instead, the evolutionary history is one of incremental recruitment: novel functions can assemble through the gradual sorting of compatible alleles across the genome. Notably, live-bearer-specific haplotypes cluster by reproductive mode against a background of genome-wide phylogenetic discordance.

Second, the value of genealogical approaches is further underscored by analyses of replicate *Antirrhinum* (snapdragon) hybrid zones (Chapter 5). These hybrid zones show that genome-wide patterns of differentiation and their heterogeneity are strongly contingent on local demographic history—post-contact population size, migration rates, and sampling schemes all leave imprints on the genomic landscape. Yet, despite this noise, shared, trait-coupled haplotypes emerge consistently at known colour loci. For example, the Planoles hybrid zone exhibits largely homogenized genomes punctuated by sharp peaks at colour loci, whereas Avellanet displays higher, more heterogeneous differentiation tied to lower introgression rates. These contrasts highlight how demographic context shapes the genomic background against which barrier peaks are detected. Topology weighting of genome-wide genealogies recovers strong clustering by floral variety at key loci, and reveals asymmetries in the ternary distributions that are difficult to explain by incomplete lineage sorting alone. This is precisely the kind of triangulation envisioned in the introduction: replicate natural contrasts focus attention on repeatable, trait-coupled genealogical features while discounting locus- and location-specific demographic artifacts.

8.4 | Looking into the past

Preliminary coalescence analyses around an island of divergence (*ROS/EL*) illustrate how timing information can discriminate selective sweeps from longer-standing barriers—and sometimes, their combination. Chapter 7 adds to the growing body of coalescence-based work on recreating the history of evolutionary processes (Hejase et al. 2020*b*; Wang and Coop 2022). Reduced within-flank coalescence times near *ROS/EL* implicate recent selection, while elevated cross-flank minimum coalescence times in haplotype–stratified comparisons indicate reduced effective migration consistent with barrier effects. This dual temporal signal clarifies how selection and gene flow jointly maintain sharp phenotypic transitions even on genomically homogenized backgrounds (Barton and Bengtsson 1986). The ability to separate "when lineages coalesce within" from "when lineages coalesce across" populations, and to do so while traversing along the genomes, is precisely the genealogical leverage argued for in Chapters 1–2, now realized in practice.

Chapter 7 only scratches the surface of the evolutionary history of a region that contains multiple loci that are under divergent selection. Looking forward, how do multiple beneficial mutations arise, combine, and get retained en route to alternative fitness peaks? Also, do their time of emergence scale with effect size? Theory predicts two exponential patterns that set clear expectations: among the mutations that actually fix during an adaptive walk, effect sizes are roughly exponentially distributed, with large steps tending to occur early (far from the fitness optimum), and average or smaller steps more common as adaptation nears a phenotypic peak (Orr 1998, 2003). By contrast, the spectrum of fitness effects of new beneficial mutations available at any given step (i.e., not just those fixed) is also predicted to be exponential with the same mean effect, invariant to the starting fitness rank of the wild type (Gillespie 1983, 1992; Orr 2003). This provides concrete, testable nulls for evaluating the timing and effect size of sweeps and barriers at each of the five loci—*ROS1*, *ROS2*, *ROS3*, *MID*, and *EL*. A further locus, *RUBIA*, which has a minor effect and shows epistatic interaction with *ROS*, offers the opportunity to test how unlinked, epistatic loci fit into this adaptive journey. Prior works emphasize that adaptation often mixes a few large- and many small-effect loci, with clustered versus dispersed architectures shaped by pleiotropy, linkage, and demography (Barton and Bengtsson 1986; Yeaman and Whitlock 2011; Burri et al. 2015), implying *ROS/EL* may function as a locally "clustered" module coupling barrier to gene flow (Feder et al. 2014; Ravinet et al. 2017). Differences in coalescence rates in and around *ROS/EL* can now test whether tightly linked targets show earlier or stronger barrier signals than unlinked contributors, as predicted by coupling theory and migration–selection–recombination models (Barton and Bengtsson 1986; Bierne et al. 2011).

Several hypotheses are now testable in snapdragons alongside new questions, given the unique combination of linked and unlinked, major and minor, and interacting loci. *(1)* Effect size ordering: Are the earliest-dated sweeps at *ROS/EL* also those with the largest effect size (Orr 1998, 2002, 2003)? *(2)* Did *RUBIA* arise from de novo mutation or from selection on standing genetic variation—and is it restricted to specific genetic backgrounds, resulting in a

sweep that is visible only on one haplotypic background (e.g., *ROS*^{ps}) while remaining neutral on the other? *(3)* Contrasting Adaptive Walks: Does the adaptive walk of the magenta-flowered *pseudomajus variety*, with linked, major-effect loci show similarities to the yellow-associated loci in *striatum*, where major-effect loci are on different linkage groups?

Together with the classic theory on exponential effect-size spectra and adaptive walks (Gillespie 1983, 1984; Orr 1998, 2003), the snapdragon system presents a unique opportunity to study the sequential recruitment of multiple beneficial mutations and the role of linkage in islands of divergence in maintaining sharp colour boundaries under gene flow . As ARG-based genealogies become available for all loci, we can reconstruct the full history of adaptation—from the first steps of divergence to the maintenance of barriers—revealing not only when and where sweeps and barriers arose, but also how linkage, pleiotropy, and frequency-dependent selection jointly sustain distinct fitness peaks.

8.5 | Final remarks

This thesis studies the process of speciation by bridging genomic landscapes to changing genealogies. Conceptually, it anchors inference to ARG-defined haplotype blocks at the resolution necessary to separate cause from correlation. Methodologically, it demonstrates that linked-read sequencing, validated phasing, and scalable genealogy inference can be combined to obtain robust ancestry reconstructions in large, non-model datasets. Empirically, it shows that a recent key innovation in *Littorina* and replicated floral colour barriers in *Antirrhinum* both leave tractable, genealogical footprints that reveal haplotype reuse, recent sweeps, and sustained barriers to gene flow, even when genome-wide histories are discordant.

The empirical journey of this thesis yields several pragmatic lessons for speciation genomics. Linked-read haplotagging, paired with careful imputation and hybrid phasing, delivers sufficient haplotype continuity for robust genealogy inference in large, non-model datasets—especially when downstream analyses can filter sites by imputation confidence and leverage high-coverage scaffolds. Methods can be chosen pragmatically: fast tree sequence inference (*Relate*, *tsinfer*) suffice to recover stable topology-weighting landscapes and broad time-to-most-recent-common-ancestor (TMRCA) patterns, while more intensive Bayesian ARGs (*Singer*) can refine local structure and uncertainty in targeted regions, like *ROS/EL*. GWAS in hybrid zones must explicitly model phenotypic covariance, relatedness, and conditioning on major loci; otherwise, long-range linkage disequilibrium and structure risk spurious discoveries that do not replicate genealogically. Finally, comparison of coalescence rates provides a tractable path to resolve sweep-versus-barrier questions and possibly retrace the evolutionary timescale of these processes.

The work also helps to clarify limits and open problems outlined at the outset. Genealogies are estimated with uncertainty, especially in low-coverage or high-LD regions; thus, convergence across methods and cross-validation with orthogonal signals (e.g., gene expression, clines) are essential for robust inference. Background selection and mutation-rate

heterogeneity can still mimic or mask barrier signatures, emphasizing the value of replicate contrasts and allele-stratified analyses. Finally, while the haplotype-block definition provides a unifying target for method development, new tools are needed to infer, visualize, and statistically test “edges” and their disjunct spans at scale, ideally with uncertainty quantification that propagates through to ecological and evolutionary parameters. ARGs and tree sequences are very rich structures that are complex and challenging to interpret. However, paired with new linked-read sequencing methods, we think there is tremendous scope for creativity around how we can best visualise local genealogical relationships, account for uncertainty, and identify signatures that are associated with the speciation process.

Bibliography

Al Bkhetan, Z., J. Zobel, A. Kowalczyk, K. Verspoor, and B. Goudey. 2019. Exploring effective approaches for haplotype block phasing. BMC Bioinformatics 20:540.

Albert, N. W., E. Butelli, S. M. A. Moss, P. Piazza, C. N. Waite, K. E. Schwinn, K. M. Davies, et al. 2021. Discrete bHLH transcription factors play functionally overlapping roles in pigmentation patterning in flowers of *Antirrhinum majus*. New Phytologist 231:849–863.

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19:1655–1664.

Allman, E. S., J. D. Mitchell, and J. A. Rhodes. 2022. Gene tree discord, simplex plots, and statistical tests under the coalescent. Systematic Biology 71:929–942.

Arouisse, B., A. Korte, F. Van Eeuwijk, and W. Kruijer. 2020. Imputation of 3 million SNPs in the Arabidopsis regional mapping population. The Plant Journal 102:872–882.

Atwell, S., Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631.

Auton, A., and G. McVean. 2007. Recombination rate estimation in the presence of hotspots. Genome Research 17:1219–1227.

Barrett, R. D. H., and H. E. Hoekstra. 2011. Molecular spandrels: tests of adaptation at the genetic level. Nature Reviews Genetics 12:767–780.

Barton, N., and B. O. Bengtsson. 1986. The barrier to genetic exchange between hybridising populations. Heredity 57:357–376.

Barton, N. H. 1979. The dynamics of hybrid zones. Heredity 43:341–359.

Barton, N. H. 1983. Multilocus clines. Evolution 37:454–471.

Barton, N. H. 1986. The effects of linkage and density-dependent regulation on gene flow. Heredity 57:415–426.

Barton, N. H. 1998. The effect of hitch-hiking on neutral genealogies. Genetical Research 72:123–133.

Barton, N. H., and B. Charlesworth. 1998. Why sex and recombination? Science 281:1986–1990.

Barton, N. H., and K. S. Gale. 1993. Genetic analysis of hybrid zones. Hybrid zones and the evolutionary process. pp. 13–45. Oxford University Press, New York.

Barton, N. H., and G. M. Hewitt. 1985. Analysis of hybrid zones. Annual Review of Ecology and Systematics 16:113–148.

Bastide, H., A. Betancourt, V. Nolte, R. Tobler, P. Stöbe, A. Futschik, and C. Schlötterer. 2013. A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. PLOS Genetics 9:e1003534.

Baum, D. A., and A. Larson. 1991. Adaptation reviewed: a phylogenetic methodology for studying character macroevolution. Systematic Biology 40:1–18.

Baumdicker, F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, et al. 2022. Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220:iyab229.

Beeravolu, C. R., M. J. Hickerson, L. A. F. Frantz, and K. Lohse. 2018. ABLE: blockwise site frequency spectra for inferring complex population histories and recombination. Genome Biology 19:145.

Berg, J. J., A. Harpak, N. Sinnott-Armstrong, A. M. Joergensen, H. Mostafavi, Y. Field, E. A. Boyle, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. eLife 8:e39725.

Bhat, J. A., D. Yu, A. Bohra, S. A. Ganie, and R. K. Varshney. 2021. Features and applications of haplotypes in crop breeding. Communications Biology 4:1266.

Bierne, N. 2010. The Distinctive Footprints Of Local Hitchhiking In A Varied Environment And Global Hitchhiking In A Subdivided Population: The Distinctive Footprints Of Local And Global Hitchhiking. Evolution 64:3254–3272.

Bierne, N., P.-A. Gagnaire, and P. David. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Current Zoology 59:72–86.

Bierne, N., J. Welch, E. Loire, F. Bonhomme, and P. David. 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. Molecular Ecology 20:2044–2072.

Blount, Z. D., J. E. Barrick, C. J. Davidson, and R. E. Lenski. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518.

Bolnick, D. I., A. K. Hund, P. Nosil, F. Peng, M. Ravinet, S. Stankowski, S. Subramanian, et al. 2023. A multivariate view of the speciation continuum. Evolution 77:318–328.

Bomblies, K., and C. L. Peichel. 2022. Genetics of adaptation. Proceedings of the National Academy of Sciences 119:e2122152119.

Bradley, D., L. Boell, D. Richardson, L. Copsey, A. Whibley, T. Xu, Y. Zhang, et al. 2025. Shaping of developmental gradients through selection on multiple loci in *Antirrhinum*. bioRxiv.

Bradley, D., P. Xu, I.-I. Mohorianu, A. Whibley, D. Field, H. Tavares, M. Couchman, et al. 2017. Evolution of flower color pattern through selection on regulatory small RNAs. Science 358:925–928.

Bradshaw, H. D., and D. W. Schemske. 2003. Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. Nature 426:176–178.

Brandt, D. Y. C., C. D. Huber, C. W. K. Chiang, and D. Ortega-Del Vecchyo. 2024. The promise of Inferring the past using the ancestral recombination graph. Genome Biology and Evolution 16:evae005.

Brandt, D. Y. C., X. Wei, Y. Deng, A. H. Vaughn, and R. Nielsen. 2022. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. Genetics 221:iyac044.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796.

Brelsford, A., D. P. L. Toews, and D. E. Irwin. 2017. Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. Proceedings of the Royal Society B: Biological Sciences 284:20171106.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase Inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84:210–223.

Browning, B. L., and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194:459–471.

Browning, S. R., and B. L. Browning. 2011. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics 12:703–714.

Browning, S. R., and B. L. Browning. 2020. Probabilistic estimation of Identity by descent segment endpoints and detection of recent selection. The American Journal of Human Genetics 107:895–910.

Bruders, R., H. Van Hollebeke, E. J. Osborne, Z. Kronenberg, E. Maclary, M. Yandell, and M. D. Shapiro. 2020. A copy number variant is associated with a spectrum of pigmentation patterns in the rock pigeon (*Columba livia*). PLOS Genetics 16:e1008274.

Buerkle, C. A., and C. Lexer. 2008. Admixture as the basis for genetic mapping. Trends in Ecology & Evolution 23:686–694.

Burri, R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. Evolution Letters 1:118–131.

Burri, R., A. Nater, T. Kawakami, C. F. Mugal, P. I. Olason, L. Smeds, A. Suh, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. Genome Research 25:1656–1665.

Burton, P. R., D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, et al. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Campagna, L., Z. Mo, A. Siepel, and J. A. C. Uy. 2022. Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species. PLOS Genetics 18:e1010474.

Campagna, L., M. Repenning, L. F. Silveira, C. S. Fontana, P. L. Tubaro, and I. J. Lovette. 2017. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science Advances 3:e1602404.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender, J. Fitz, D. Koenig, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nature Genetics 43:956–963.

Carmi, S., P. F. Palamara, V. Vacic, T. Lencz, A. Darvasi, and I. Pe'er. 2013. The variance of identity-by-descent sharing in the Wright–Fisher model. Genetics 193:911–928.

Castro, J. P., M. N. Yancoskie, M. Marchini, S. Belohlavy, L. Hiramatsu, M. Kučka, W. H. Beluch, et al. 2019. An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. eLife 8:e42014.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:s13742-015-0047–8.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution 15:538–543.

Charlesworth, B., M. Nordborg, and D. Charlesworth. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research 70:155–174.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang, and H. Li. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18:170–175.

Cheng, X., and M. Steinrucken. 2024. Population genomic scans for natural selection and demography. Annual Review of Genetics 58:319–339.

Clark, A. G. 2004. The role of haplotypes in candidate gene studies. Genetic Epidemiology 27:321–333.

Clauw, P., T. J. Ellis, H.-J. Liu, and E. Sasaki. 2024. Beyond the standard GWAS—a guide for plant biologists. Plant and Cell Physiology pcae079.

Colosimo, P. F., K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, et al. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307:1928–1933.

Coyne, J. A., and H. A. Orr. 1998. The evolutionary genetics of speciation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353:287–305.

Crawford, D. C., and D. A. Nickerson. 2005. Definition and clinical importance of haplotypes. Annual Review of Medicine 56:303–320.

Cruickshank, T. E., and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Molecular Ecology 23:3133–3157.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008.

Davies, R. W., J. Flint, S. Myers, and R. Mott. 2016. Rapid genotype imputation from sequence without reference panels. Nature Genetics 48:965–969.

Davies, R. W., M. Kucka, D. Su, S. Shi, M. Flanagan, C. M. Cunniff, Y. F. Chan, et al. 2021. Rapid genotype imputation from sequence with reference panels. Nature Genetics 53:1104–1111.

De Queiroz, A. 2002. Contingent Predictability in Evolution: Key traits and diversification. Systematic Biology 51:917–929.

Delaneau, O., J.-F. Zagury, M. R. Robinson, J. L. Marchini, and E. T. Dermitzakis. 2019. Accurate, scalable and integrative haplotype estimation. Nature Communications 10:5436.

Deng, Y., R. Nielsen, and Y. S. Song. 2024. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. bioRxiv.

Deng, Y., Y. S. Song, and R. Nielsen. 2021. The distribution of waiting distances in ancestral recombination graphs. Theoretical Population Biology 141:34–43.

Du, H., J. Wu, K.-X. Ji, Q.-Y. Zeng, M.-W. Bhuiya, S. Su, Q.-Y. Shu, et al. 2015. Methylation mediated by an anthocyanin, *O*-methyltransferase, is involved in purple flower coloration in *Paeonia*. Journal of Experimental Botany 66:6563–6577.

Durán-Castillo, M., A. Hudson, Y. Wilson, D. L. Field, and A. D. Twyford. 2022. A phylogeny of *Antirrhinum* reveals parallel evolution of alpine morphology. New Phytologist 233:1426–1439.

Duranton, M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme, and P.-A. Gagnaire. 2018. The origin and remolding of genomic islands of differentiation in the European sea bass. Nature Communications 9:2518.

Ebert, P., P. A. Audano, Q. Zhu, B. Rodriguez-Martin, D. Porubsky, M. J. Bonder, A. Sulovari, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science 372:eabf7117.

Edge, P., V. Bafna, and V. Bansal. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Research 27:801–812.

Eggertsson, H. P., H. Jonsson, S. Kristmundsdottir, E. Hjartarson, B. Kehr, G. Masson, F. Zink, et al. 2017. Graphtyper enables population-scale genotyping using pangenome graphs. Nature Genetics 49:1654–1660.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756–760.

Endler, J. A. 1977. Geographic variation, speciation, and clines. Monographs in population biology. Princeton Univ. Press, Princeton, NJ.

Fan, C., J. L. Cahoon, B. L. Dinh, D. Ortega-Del Vecchyo, C. Huber, M. D. Edge, N. Mancuso, et al. 2023. A likelihood-based framework for demographic inference from genealogical trees. bioRxiv.

Fan, C., N. Mancuso, and C. W. K. Chiang. 2022. A genealogical estimate of genetic relationships. American Journal of Human Genetics 109:812–824.

Faria, R., P. Chaube, H. E. Morales, T. Larsson, A. R. Lemmon, E. M. Lemmon, M. Rafajlović, et al. 2019. Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. Molecular Ecology 28:1375–1393.

Feder, J. L., S. P. Egan, and P. Nosil. 2012. The genomics of speciation-with-gene-flow. Trends in Genetics 28:342–350.

Feder, J. L., S. M. Flaxman, S. P. Egan, A. A. Comeault, and P. Nosil. 2013. Geographic mode of speciation and genomic divergence. Annual Review of Ecology, Evolution, and Systematics 44:73–97.

Feder, J. L., P. Nosil, A. C. Wacholder, S. P. Egan, S. H. Berlocher, and S. M. Flaxman. 2014. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. Journal of Heredity 105:810–820.

Felsenstein, J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. Evolution 35:124–138.

Field, D. L., S. Stankowski, T. Reiter, J. Polechova, D. Bradley, D. Richardson, A. Pal, et al. 2025. Genome-wide cline analysis identifies new locus contributing to a barrier to gene flow across an *Antirrhinum* hybrid zone. bioRxiv.

Fisher, R. A. 1954. A fuller theory of “Junctions” in inbreeding. Heredity 8:187–197.

Flint, J., and T. F. C. Mackay. 2009. Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research 19:723–733.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Garud, N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLOS Genetics 11:e1005004.

Gillespie, J. H. 1983. A simple stochastic gene substitution model. Theoretical Population Biology 23:202–215.

Gillespie, J. H. 1984. Molecular evolution over the mutational landscape. Evolution 38:1116–1129.

Gillespie, J. H. 1992. The causes of molecular evolution. Oxford Series in Ecology and Evolution. Oxford University Press, New York.

Gompert, Z., E. G. Mandeville, and C. A. Buerkle. 2017. Analysis of population genomic data from hybrid zones. Annual Review of Ecology, Evolution, and Systematics 48:207–229.

Gower, G., A. P. Ragsdale, G. Bisschop, R. N. Gutenkunst, M. Hartfield, E. Noskova, S. Schiffels, et al. 2022. Demes: a standard format for demographic models. Genetics 222:iyac131.

Green, R. E., J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, et al. 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444:330–336.

Griffiths, R. C., and P. Marjoram. 1997. An ancestral recombination graph. Progress in Population Genetics and Human Evolution, The IMA Volumes in Mathematics and its Applications (Vol. 87). pp. 257–270. Springer New York, New York, NY.

Grossman, S. R., I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886.

Guerrero, R. F., and M. W. Hahn. 2018. Quantifying the risk of hemiplasy in phylogenetic inference. Proceedings of the National Academy of Sciences 115:12787–12792.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Systematic Biology 59:307–321.

Gutenkunst, R., R. Hernandez, S. Williamson, and C. Bustamante. 2010. Diffusion approximations for demographic inference: DaDi. Nature Precedings 1–1.

Hahn, M. W., and L. Nakhleh. 2016. Irrational exuberance for resolved species trees. Evolution 70:7–17.

Haller, B. C., P. L. Ralph, and P. W. Messer. 2025. SLiM 5: Eco-evolutionary simulations across multiple chromosomes and full genomes. bioRxiv.

Hamilton, N. E., and M. Ferry. 2018. ggtern: Ternary diagrams using ggplot2. Journal of Statistical Software 87.

Harris, K., and R. Nielsen. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLOS Genetics 9:e1003521.

Hartl, D. L., and A. G. Clark. 1997. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

He, C., J. Holme, and J. Anthony. 2014. SNP genotyping: The KASP assay. Crop Breeding, Methods in Molecular Biology (Vol. 1145). pp. 75–86. Springer New York, New York, NY.

Hejase, H. A., N. Dukler, and A. Siepel. 2020*a*. From summary statistics to gene trees: Methods for Inferring positive selection. Trends in Genetics 36:243–258.

Hejase, H. A., Z. Mo, L. Campagna, and A. Siepel. 2022. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. Molecular Biology and Evolution 39:msab332.

Hejase, H. A., A. Salman-Minkov, L. Campagna, M. J. Hubisz, I. J. Lovette, I. Gronau, and A. Siepel. 2020*b*. Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. Proceedings of the National Academy of Sciences 117:30554–30565.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Hickey, G., D. Heller, J. Monlong, J. A. Sibbesen, J. Sirén, J. Eizenga, E. T. Dawson, et al. 2020. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biology 21:35.

Hofmeister, R. J., D. M. Ribeiro, S. Rubinacci, and O. Delaneau. 2023. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. Nature Genetics 55:1243–1249.

Hooper, D. M., C. S. McDiarmid, M. J. Powers, N. M. Justyn, M. Kučka, N. S. Hart, G. E. Hill, et al. 2024. Spread of yellow-bill-color alleles favored by selection in the long-tailed finch hybrid system. Current Biology 34:5444–5456.

Howie, B., J. Marchini, and M. Stephens. 2011. Genotype imputation with thousands of genomes. G3 Genes|Genomes|Genetics 1:457–470.

Hubisz, M. J., A. L. Williams, and A. Siepel. 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. PLOS Genetics 16:e1008895.

Hudson, A., J. Critchley, and Y. Erasmus. 2008. The genus *Antirrhinum* (Snapdragon): A flowering plant model for evolution and development. Cold Spring Harbor Protocols 2008:pdb.emo100.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology 23:183–201.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology 7:44.

Hudson, R. R., and N. L. Kaplan. 1995. Deleterious background selection with recombination. Genetics 141:1605–1617.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23:254–267.

Igolkina, A. A., S. Vorbrugg, F. A. Rabanal, H.-J. Liu, H. Ashkenazy, A. E. Kornienko, J. Fitz, et al. 2025. A comparison of 27 *Arabidopsis thaliana* genomes and the path toward an unbiased characterization of genetic polymorphism. Nature Genetics.

Irwin, D. E., M. Alcaide, K. E. Delmore, J. H. Irwin, and G. L. Owens. 2016. Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. Molecular Ecology 25:4488–4507.

Jain, M., S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology 36:338–345.

Jiggins, C. D., R. E. Naisbit, R. L. Coe, and J. Mallet. 2001. Reproductive isolation caused by colour pattern mimicry. Nature 411:302–305.

Johannesson, K. 2003. Evolution in Littorina: Ecology matters. Journal of Sea Research 49:107–117.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Kong, N. B. Freimer, C. Sabatti, et al. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42:348–354.

Kelleher, J., K. R. Thornton, J. Ashander, and P. L. Ralph. 2018. Efficient pedigree recording for fast population genetics simulation. PLOS Computational Biology 14:e1006581.

Kelleher, J., Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. 2019. Inferring whole-genome histories in large population datasets. Nature Genetics 51:1330–1338.

Kemppainen, P., C. G. Knight, D. K. Sarma, T. Hlaing, A. Prakash, Y. N. Maung Maung, P. Somboon, et al. 2015. Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. Molecular Ecology Resources 15:1031–1045.

Khimoun, A., J. Cornuault, M. Burrus, B. Pujol, C. Thebaud, and C. Andalo. 2013. Ecology predicts parapatric distributions in two closely related *Antirrhinum majus* subspecies. Evolutionary Ecology 27:51–64.

Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and their Applications 13:235–248.

Knief, U., C. M. Bossu, N. Saino, B. Hansson, J. Poelstra, N. Vijay, M. Weissensteiner, et al. 2019. Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. Nature Ecology & Evolution 3:570–576.

Koch, E. L., H. E. Morales, J. Larsson, A. M. Westram, R. Faria, A. R. Lemmon, E. M. Lemmon, et al. 2021. Genetic variation for adaptive traits is associated with polymorphic inversions in *Littorina saxatilis*. Evolution Letters 5:196–213.

Kodama, M., H. Brinch-Pedersen, S. Sharma, I. B. Holme, B. Joernsgaard, T. Dzhanfezova, D. B. Amby, et al. 2018. Identification of transcription factor genes involved in anthocyanin biosynthesis in carrot (*Daucus carota L.*) using RNA-Seq. BMC Genomics 19:811.

Lander, E. S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush. 2012. Inference of population structure using dense haplotype data. PLOS Genetics 8:e1002453.

Le Moan, A., S. Stankowski, M. Rafajlović, O. Ortega-Martinez, R. Faria, R. K. Butlin, and K. Johannesson. 2024. Coupling of twelve putative chromosomal inversions maintains a strong barrier to gene flow between snail ecotypes. Evolution Letters 8:575–586.

Lee, D., Y. Kim, Y. Chung, D. Lee, D. Seo, T. J. Choi, D. Lim, et al. 2021. Accuracy of genotype imputation based on reference population size and marker density in Hanwoo cattle. Journal of Animal Science and Technology 63:1232–1246.

Leitwein, M., M. Duranton, Q. Rougemont, P. Gagnaire, and L. Bernatchez. 2020. Using haplotype information for conservation genomics. Trends in Ecology & Evolution 35:245–258.

Lewanski, A. L., M. C. Grundler, and G. S. Bradburd. 2024. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. PLOS Genetics 20:e1011110.

Lewis, J. J., R. C. Geltman, P. C. Pollak, K. E. Rondem, S. M. Van Belleghem, M. J. Hubisz, P. R. Munn, et al. 2019. Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. Proceedings of the National Academy of Sciences 116:24174–24183.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.

Li, H., and P. Ralph. 2019. Local PCA shows how the effect of population structure differs along the genome. Genetics 211:289–304.

Li, M., D. Zhang, Q. Gao, Y. Luo, H. Zhang, B. Ma, C. Chen, et al. 2019. Genome structure and evolution of *Antirrhinum majus L.* Nature Plants 5:174–183.

Li, N., and M. Stephens. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34:816–834.

Lifchitz, H., S. Vedula, A. Pal, D. Shipilina, and S. Stankowski. 2025. TwisstNTern 2: Ternary analysis of topology weights from tree sequences. EcoEvoRxiv.

Lindtke, D., S. C. González-Martínez, D. Macaya-Sanz, and C. Lexer. 2013. Admixture mapping of quantitative traits in *Populus* hybrid zones: power and limitations. Heredity 111:474–485.

Linnen, C. R., E. P. Kingsley, J. D. Jensen, and H. E. Hoekstra. 2009. On the origin and spread of an adaptive allele in deer mice. Science 325:1095–1098.

Linnen, C. R., Y.-P. Poh, B. K. Peterson, R. D. H. Barrett, J. G. Larson, J. D. Jensen, and H. E. Hoekstra. 2013. Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339:1312–1316.

Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris, Z. Xiong, L. Zhou, et al. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785–794.

Logsdon, G. A., M. R. Vollger, and E. E. Eichler. 2020. Long-read human genome sequencing and its applications. Nature Reviews Genetics 21:597–614.

Lohse, K., M. Chmelik, S. H. Martin, and N. H. Barton. 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. Genetics 202:775–786.

Lotterhos, K. E., and M. C. Whitlock. 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular Ecology 24:1031–1046.

Lou, R. N., A. Jacobs, A. P. Wilder, and N. O. Therkildsen. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology 30:5966–5993.

Lowry, D. B., and J. H. Willis. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLOS Biology 8:e1000500.

Lundberg, M., M. Liedvogel, K. Larson, H. Sigeman, M. Grahn, A. Wright, S. Åkesson, et al. 2017. Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. Evolution Letters 1:155–168.

Mackay, T. F. C., E. A. Stone, and J. F. Ayroles. 2009. The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics 10:565–577.

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

Majidian, S., and F. J. Sedlazeck. 2020. PhaseME: Automatic rapid assessment of phasing quality and phasing improvement. GigaScience 9:giaa078.

Mallarino, R., C. Henegar, M. Mirasierra, M. Manceau, C. Schradin, M. Vallejo, S. Beronja, et al. 2016. Developmental mechanisms of stripe patterns in rodents. Nature 539:518–523.

Manceau, M., V. S. Domingues, C. R. Linnen, E. B. Rosenblum, and H. E. Hoekstra. 2010. Convergence in pigmentation at multiple levels: mutations, genes and function. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2439–2450.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39:906–913.

Marjoram, P., and J. D. Wall. 2006. Fast "coalescent" simulation. BMC Genetics 7:16.

Marques, D. A., K. Lucek, J. I. Meier, S. Mwaiko, C. E. Wagner, L. Excoffier, and O. Seehausen. 2016. Genomics of rapid incipient speciation in sympatric threespine stickleback. PLOS Genetics 12:e1005887.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, et al. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Research 23:1817–1828.

Martin, S. H., J. W. Davey, and C. D. Jiggins. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Molecular Biology and Evolution 32:244–257.

Martin, S. H., J. W. Davey, C. Salazar, and C. D. Jiggins. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLOS Biology 17:e2006288.

Martin, S. H., and S. M. Van Belleghem. 2017. Exploring evolutionary relationships across the genome using topology weighting. Genetics 206:429–438.

Matejovičová, L. 2022. Genetic basis of flower colour as a model for adaptive evolution. Institute of Science and Technology Austria.

Maynard Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23:23–35.

McVean, G. A. T., and N. J. Cardin. 2005. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences 360:1387–1393.

Meier, J. I., P. A. Salazar, M. Kučka, R. W. Davies, A. Dréau, I. Aldás, O. Box Power, et al. 2021. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. Proceedings of the National Academy of Sciences 118:e2015005118.

Messer, P. W., and D. A. Petrov. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28:659–669.

Mészáros, G., M. Milanesi, P. Ajmone-Marsan, and Y. T. Utsunomiya. 2021. Haplotype analysis applied to livestock genomics. Frontiers in Genetics 12:660478.

Meyer, J. R., D. T. Dobias, J. S. Weitz, J. E. Barrick, R. T. Quick, and R. E. Lenski. 2012. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science 335:428–432.

Meyer, L., P. Barry, F. Riquet, A. Foote, C. Der Sarkissian, R. L. Cunha, C. Arbiol, et al. 2024. Divergence and gene flow history at two large chromosomal inversions underlying ecotype differentiation in the long-snouted seahorse. Molecular Ecology 33:e17277.

Miller, A. H., J. T. Stroud, and J. B. Losos. 2023. The ecology and evolution of key innovations. Trends in Ecology & Evolution 38:122–131.

Momigliano, P., A.-B. Florin, and J. Merilä. 2021. Biases in demographic modelling affect our understanding of recent divergence. Molecular Biology and Evolution msab047.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.-Y. Zhong, and S. Myles. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3 Genes|Genomes|Genetics 5:2383–2390.

Morales, H. E., R. Faria, K. Johannesson, T. Larsson, M. Panova, A. M. Westram, and R. K. Butlin. 2019. Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. Science Advances 5:eaav9963.

Nadeau, N. J., M. Ruiz, P. Salazar, B. Counterman, J. A. Medina, H. Ortiz-Zuazaga, A. Morrison, et al. 2014. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Research 24:1316–1333.

Nadeau, N. J., A. Whibley, R. T. Jones, J. W. Davey, K. K. Dasmahapatra, S. W. Baxter, M. A. Quail, et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society B: Biological Sciences 367:343–353.

Naing, A. H., K. I. Park, T. N. Ai, M. Y. Chung, J. S. Han, Y.-W. Kang, K. B. Lim, et al. 2017. Overexpression of snapdragon Delila (*Del*) gene in tobacco enhances anthocyanin accumulation and abiotic stress tolerance. BMC Plant Biology 17:65.

Nguyen, T. N., M. Repenning, C. Suertegaray Fontana, and L. Campagna. 2024. Genomic islands of speciation harbor genes underlying coloration differences in a pair of Neotropical seedeaters. Evolution 78:1161–1173.

Nielsen, R. 2005. Molecular signatures of natural selection. Annual Review of Genetics 39:197–218.

Nielsen, R., A. H. Vaughn, and Y. Deng. 2024. Inference and applications of ancestral recombination graphs. Nature Reviews Genetics 1–12.

Noor, M. A. F., and S. M. Bennett. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103:439–444.

Nosil, P. 2012. Ecological Speciation. Oxford University Press.

Nouhaud, P., S. H. Martin, B. Portinha, V. C. Sousa, and J. Kulmuni. 2022. Rapid and predictable genome evolution across three hybrid ant populations. PLOS Biology 20:e3001914.

Novembre, J., and N. H. Barton. 2018. Tread lightly interpreting polygenic tests of selection. Genetics 208:1351–1355.

Okitsu, N., T. Mizuno, K. Matsui, S. H. Choi, and Y. Tanaka. 2018. Molecular cloning of flavonoid biosynthetic genes and biochemical characterization of anthocyanin *O*-methyltransferase of *Nemophila menziesii* Hook. and Arn. Plant Biotechnology 35:9–16.

Okonechnikov, K., A. Conesa, and F. García-Alcalde. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292–294.

Ono, E., M. Fukuchi-Mizutani, N. Nakamura, Y. Fukui, K. Yonekura-Sakakibara, M. Yamaguchi, T. Nakayama, et al. 2006. Yellow flowers generated by expression of the aurone biosynthetic pathway. Proceedings of the National Academy of Sciences 103:11075–11080.

Orr, H. A. 1998. The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution 52:935–949.

Orr, H. A. 2001. The genetics of species differences. Trends in Ecology & Evolution 16:343–350.

Orr, H. A. 2003. The distribution of fitness effects among beneficial mutations. Genetics 163:1519–1526.

Orr, H. A., and M. Turelli. 2001. The evolution of postzygotic isolation: Accumulating Dobzhansky-Muller incompatibilities. Evolution 55:1085–1094.

Otte, K. A., and C. Schlötterer. 2021. Detecting selected haplotype blocks in evolve and resequence experiments. Molecular Ecology Resources 21:93–109.

Pal, A. 2025. Born to chill in fields; Tracing colours along roads; Speciation's easy. The Promised Haiku.

Pal, A., D. Shipilina, A. Le Moan, A. J. McNairn, J. K. Grenier, M. Kucka, G. Coop, et al. 2025. Genealogical analysis of replicate flower colour hybrid zones in *Antirrhinum*. Molecular Ecology e70067.

Pal, A., and B. Vicoso. 2015. The X chromosome of Hemipteran insects: Conservation, dosage compensation and sex-biased expression. Genome Biology and Evolution 7:3259–3268.

Papa, R., D. D. Kapan, B. A. Counterman, K. Maldonado, D. P. Lindstrom, R. D. Reed, H. F. Nijhout, et al. 2013. Multi-allelic major effect genes interact with minor effect QTLs to control adaptive color pattern variation in *Heliconius erato*. PLOS ONE 8:e57033.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, et al. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Pavlidis, P., and N. Alachiotis. 2017. A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research-Thessaloniki 24:7.

Pei, J., Y. Zhang, R. Nielsen, and Y. Wu. 2020. Inferring the ancestry of parents and grandparents from genetic data. PLOS Computational Biology 16:e1008065.

Pieszko, T., J. Kelleher, C. G. Wilson, and T. G. Barraclough. 2025. Detecting and quantifying rare sex in natural populations. bioRxiv.

Podos, J., and K. M. Schroeder. 2024. Ecological speciation in Darwin’s finches: Ghosts of finches future. Science 386:211–217.

Poelstra, J. W., N. Vijay, C. M. Bossu, H. Lantz, B. Ryll, I. Müller, V. Baglione, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344:1410–1414.

Pokrovac, I., and Ž. Pezer. 2022. Recent advances and current challenges in population genomics of structural variation in animals and plants. Frontiers in Genetics 13:1060898.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38:904–909.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLOS Genetics 5:e1000519.

Ragsdale, A. P., and R. N. Gutenkunst. 2017. Inferring demographic history using two-locus statistics. Genetics 206:1037–1048.

Ralph, P., and G. Coop. 2013. The geography of recent genetic ancestry across Europe. PLOS Biology 11:e1001555.

Ralph, P., K. Thornton, and J. Kelleher. 2020. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. Genetics 215:779–797.

Rancilhac, L., S. G. de Souza, S. M. Lukhele, M. Sebastianelli, B. O. Ogolowa, M. Moysi, C. Nikiforou, et al. 2024. Introgression across narrow contact zones shapes the genomic landscape of phylogenetic variation in an African bird clade. bioRxiv.

Rasmussen, M. D., M. J. Hubisz, I. Gronau, and A. Siepel. 2014. Genome-wide inference of ancestral recombination graphs. PLOS Genetics 10:e1004342.

Ravinet, M., R. Faria, R. K. Butlin, J. Galindo, N. Bierne, M. Rafajlović, M. a. F. Noor, et al. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. Journal of Evolutionary Biology 30:1450–1477.

Raychaudhuri, S., C. Sandor, E. A. Stahl, J. Freudenberg, H.-S. Lee, X. Jia, L. Alfredsson, et al. 2012. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nature Genetics 44:291–296.

Recknagel, H., M. Carruthers, A. A. Yurchenko, M. Nokhbatolfoghahai, N. A. Kamenos, M. M. Bain, and K. R. Elmer. 2021. The functional genetic architecture of egg-laying and live-bearing reproduction in common lizards. Nature Ecology & Evolution 5:1546–1556.

Reeve, J., R. K. Butlin, E. L. Koch, S. Stankowski, and R. Faria. 2024. Chromosomal inversion polymorphisms are widespread across the species ranges of rough periwinkles (*Littorina saxatilis* and *L. arcana*). Molecular Ecology 33:e17160.

Reid, D. G. 1996. Systematics and evolution of *Littorina*. The Ray Society, London.

Reid, D. G., P. Dyal, and S. T. Williams. 2012. A global molecular phylogeny of 147 periwinkle species (Gastropoda, Littorininae). Zoologica Scripta 41:125–136.

Richardson, D. M., D. Bradley, L. Copsey, A. Whibley, M. Burrus, C. Andalo, S. Zhu, et al. 2025. Genomic tree scans identify loci underlying adaptive peaks in *Antirrhinum*. bioRxiv.

Richardson, J. L., S. P. Brady, I. J. Wang, and S. F. Spear. 2016. Navigating the pitfalls and promise of landscape genetics. Molecular Ecology 25:849–863.

Ringbauer, H., A. Kolesnikov, D. L. Field, and N. H. Barton. 2018. Estimating barriers to gene flow from distorted Isolation-by-distance patterns. Genetics 208:1231–1245.

Rockman, M. V. 2012. The QTN program and the alleles that matter for evolution: All that's gold does not glitter. Evolution 66:1–17.

Roze, D. 2021. A simple expression for the strength of selection on recombination generated by interference among mutations. Proceedings of the National Academy of Sciences 118:e2022805118.

Rueda-M, N., C. Pardo-Diaz, G. Montejo-Kovacevich, W. O. McMillan, K. M. Kozak, C. F. Arias, J. Ready, et al. 2024. Genomic evidence reveals three W-autosome fusions in Heliconius butterflies. PLOS Genetics 20:e1011318.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Salojarvi, J., A. Rambani, Z. Yu, R. Guyot, S. Strickler, M. Lepelley, C. Wang, et al. 2024. The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. Nature Genetics 56:721–731.

Savolainen, O., M. Lascoux, and J. Merilä. 2013. Ecological genomics of local adaptation. Nature Reviews Genetics 14:807–820.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Schluter, D., and L. H. Rieseberg. 2022. Three problems in the genetics of speciation by selection. Proceedings of the National Academy of Sciences 119:e2122153119.

Schwartz, R., B. V. Halldórsson, V. Bafna, A. G. Clark, and S. Istrail. 2003. Robustness of inference of haplotype block structure. Journal of Computational Biology 10:13–19.

Schwinn, K., J. Venail, Y. Shang, S. Mackay, V. Alm, E. Butelli, R. Oyama, et al. 2006. A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. The Plant Cell 18:831–851.

Sedghifar, A., Y. Brandvain, and P. Ralph. 2016. Beyond clines: lineages and haplotype blocks in hybrid zones. Molecular Ecology 25:2559–2576.

Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, et al. 2014. Genomics and the origin of species. Nature Reviews Genetics 15:176–192.

Sella, G., and N. H. Barton. 2019. Thinking about the evolution of complex traits in the Era of Genome-Wide Association Studies. Annual Review of Genomics and Human Genetics 20:461–493.

Seshappa, G. 1947. Oviparity in *Littorina saxatilis* (Olivi). Nature 160:335–336.

Shajii, A., I. Numanagić, and B. Berger. 2018*a*. Latent variable model for aligning barcoded short reads improves downstream analyses. RECOMB pp. 280–282.

Shajii, A., I. Numanagić, C. Whelan, and B. Berger. 2018*b*. Statistical binning for barcoded reads improves downstream analyses. Cell Systems 7:219-226.

Sharma, S., I. B. Holme, G. Dionisio, M. Kodama, T. Dzhanfezova, B. Joernsgaard, and H. Brinch-Pedersen. 2020. Cyanidin based anthocyanin biosynthesis in orange carrot is restored by expression of AmRosea1 and AmDelila, MYB and bHLH transcription factors. Plant Molecular Biology 103:443–456.

Shi, Y., K. L. Bouska, G. J. McKinney, W. Dokai, A. Bartels, M. V. McPhee, and W. A. Larson. 2023. Gene flow influences the genomic architecture of local adaptation in six riverine fish species. Molecular Ecology 32:1549–1566.

Shine, R. 1983. Reptilian reproductive modes: The oviparity-viviparity continuum. Herpetologica 39:1–8.

Shipilina, D., A. Pal, S. Stankowski, Y. F. Chan, and N. H. Barton. 2023. On the origin and structure of haplotype blocks. Molecular Ecology 32:1441–1457.

Slatkin, M. 1972. On treating the chromosome as the unit of selection. Genetics 72:157–168.

Slatkin, M. 1975. Gene flow and selection in a two-locus system. Genetics 81:787–802.

Sobel, J. M., and M. A. Streisfeld. 2015. Strong premating reproductive isolation drives incipient speciation in *Mimulus aurantiacus*. Evolution 69:447–461.

Song, M., H. Wang, Z. Wang, H. Huang, S. Chen, and H. Ma. 2021. Genome-wide characterization and Analysis of bHLH transcription factors related to anthocyanin biosynthesis in Fig (*Ficus carica L.*). Frontiers in Plant Science 12:730692.

Sousa, V. C., A. Grelaud, and J. Hey. 2011. On the nonidentifiability of migration time estimates in isolation with migration models. Molecular Ecology 20:3956–3962.

Speidel, L., M. Forest, S. Shi, and S. R. Myers. 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics 51:1321–1329.

Stahl, K., D. Gola, and I. R. König. 2021. Assessment of imputation quality: comparison of phasing and imputation algorithms in real data. Frontiers in Genetics 12:724037.

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

Stankowski, S., M. A. Chase, H. McIntosh, and M. A. Streisfeld. 2023. Integrating top-down and bottom-up approaches to understand the genetic architecture of speciation across a monkeyflower hybrid zone. Molecular Ecology 32:2041–2054.

Stankowski, S., and M. Ravinet. 2021. Defining the speciation continuum. Evolution 75:1256–1273.

Stankowski, S., J. M. Sobel, and M. A. Streisfeld. 2015. The geography of divergence with gene flow facilitates multitrait adaptation and the evolution of pollinator isolation in *Mimulus aurantiacus*. Evolution 69:3054–3068.

Stankowski, S., A. M. Westram, Z. B. Zagrodzka, I. Eyres, T. Broquet, K. Johannesson, and R. K. Butlin. 2020. The evolution of strong reproductive isolation between sympatric intertidal snails. Philosophical Transactions of the Royal Society B: Biological Sciences 375:20190545.

Stankowski, S., Z. B. Zagrodzka, M. D. Garlovsky, A. Pal, D. Shipilina, D. G. Castillo, H. Lifchitz, et al. 2024. The genetic basis of a recent transition to live-bearing in marine snails. Science 383:114–119.

Steinrücken, M., J. Kamm, J. P. Spence, and Y. S. Song. 2019. Inference of complex population histories using whole-genome sequences from multiple populations. Proceedings of the National Academy of Sciences 116:17115–17120.

Steinrücken, M., J. P. Spence, J. A. Kamm, E. Wieczorek, and Y. S. Song. 2018. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Molecular Ecology 27:3873–3888.

Stephens, M., and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in Haplotype inference and missing-data imputation. The American Journal of Human Genetics 76:449–462.

Stern, A. J., P. R. Wilton, and R. Nielsen. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLOS Genetics 15:e1008384.

Stinchcombe, J. R., and H. E. Hoekstra. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100:158–170.

Suda, R. A., S. Kubota, V. Kumar, V. Castric, U. Krämer, S.-I. Morinaga, and T. Tsuchimatsu. 2025. Population genomics reveals demographic history and climate adaptation in Japanese *Arabidopsis halleri*. Plant And Cell Physiology 66:529–541.

Sundquist, A., E. Fratkin, C. B. Do, and S. Batzoglou. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research 18:676–682.

Surendranadh, P., L. Arathoon, C. A. Baskett, D. L. Field, M. Pickup, and N. H. Barton. 2022. Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of *Antirrhinum majus*. Genetics 221:iyac083.

Surendranadh, P., S. Stankowski, D. L. Field, and N. H. Barton. 2025. Extreme long-range migration distorts flower colour clines in an *Antirrhinum* hybrid zone. bioRxiv.

Szpiech, Z. A., and R. D. Hernandez. 2014. selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Molecular Biology and Evolution 31:2824–2827.

Taliun, D., J. Gamper, and C. Pattaro. 2014. Efficient haplotype block recognition of very long and dense genetic sequences. BMC Bioinformatics 15:10.

Tang, K., K. R. Thornton, and M. Stoneking. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. PLOS Biology 5:e171.

Tarasov, A., A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034.

Tavares, H., A. Whibley, D. L. Field, D. Bradley, M. Couchman, L. Copsey, J. Elleouet, et al. 2018. Selection and gene flow shape genomic islands that control floral guides. Proceedings of the National Academy of Sciences 115:11006–11011.

Tewhey, R., V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork. 2011. The importance of phase information for human genomics. Nature Reviews Genetics 12:215–223.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

The International HapMap Consortium, P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Theißen, G. 2009. Saltational evolution: hopeful monsters are here to stay. Theory in Biosciences 128:43–51.

Thompson, E. A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194:301–326.

Todesco, M., G. L. Owens, N. Bercovich, J.-S. Légaré, S. Soudi, D. O. Burge, K. Huang, et al. 2020. Massive haplotypes underlie ecotypic differentiation in sunflowers. Nature 584:602–607.

Turner, I., K. V. Garimella, Z. Iqbal, and G. McVean. 2018. Integrating long-range connectivity information into de Bruijn graphs. Bioinformatics 34:2556–2565.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. De Vries, Y. Okada, A. R. Martin, H. C. Martin, et al. 2021. Genome-wide association studies. Nature Reviews Methods Primers 1:59.

Uy, J. A. C., E. A. Cooper, S. Cutie, M. R. Concannon, J. W. Poelstra, R. G. Moyle, and C. E. Filardi. 2016. Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. Proceedings of the Royal Society B: Biological Sciences 283:20160731.

Vargas, P., J. A. Rosselló, R. Oyama, and J. Güemes. 2004. Molecular evidence for naturalness of genera in the tribe Antirrhineae (Scrophulariaceae) and three independent evolutionary lineages from the New World and the Old. Plant Systematics and Evolution 249:151–172.

Vijay, N., C. M. Bossu, J. W. Poelstra, M. H. Weissensteiner, A. Suh, A. P. Kryukov, and J. B. W. Wolf. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nature Communications 7:13195.

Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 2017. 10 years of GWAS discovery: Biology, function, and translation. The American Journal of Human Genetics 101:5–22.

Wagner, A. 2011. The origins of evolutionary innovations: A theory of transformative change in living systems. Oxford University Press.

Wakeley, J. 2009. Coalescent theory: an introduction. Roberts & Co. Publishers, Greenwood Village, Colo.

Wallbank, R. W. R., S. W. Baxter, C. Pardo-Diaz, J. J. Hanly, S. H. Martin, J. Mallet, K. K. Dasmahapatra, et al. 2016. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. PLOS Biology 14:e1002353.

Wallberg, A., C. Schöning, M. T. Webster, and M. Hasselmann. 2017. Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. PLOS Genetics 13:e1006792.

Wang, O., R. Chin, X. Cheng, M. K. Y. Wu, Q. Mao, J. Tang, Y. Sun, et al. 2019. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Research 29:798–808.

Wang, S., and G. Coop. 2022. A complex evolutionary history of genetic barriers to gene flow in hybridizing warblers. bioRxiv.

Watowich, M. M., K. L. Chiou, B. Graves, M. J. Montague, L. J. N. Brent, J. P. Higham, J. E. Horvath, et al. 2025. Best practices for genotype imputation from low-coverage sequencing data in natural populations. Molecular Ecology Resources 25:e13854.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358.

Weisenfeld, N. I., V. Kumar, P. Shah, D. M. Church, and D. B. Jaffe. 2017. Direct determination of diploid genome sequences. Genome Research 27:757–767.

Wellenreuther, M., and L. Bernatchez. 2018. Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution 33:427–440.

Westram, A. M., R. Faria, K. Johannesson, and R. Butlin. 2021. Using replicate hybrid zones to understand the genomic basis of adaptive divergence. Molecular Ecology 30:3797–3814.

Westram, A. M., M. Rafajlović, P. Chaube, R. Faria, T. Larsson, M. Panova, M. Ravinet, et al. 2018. Clines on the seashore: The genomic architecture underlying rapid divergence in the face of gene flow. Evolution Letters 2:297–309.

Whibley, A. C., N. B. Langlade, C. Andalo, A. I. Hanna, A. Bangham, C. Thébaud, and E. Coen. 2006. Evolutionary paths underlying flower color variation in *Antirrhinum*. Science 313:963–966.

Whiting, J. R., T. R. Booker, C. Rougeux, B. M. Lind, P. Singh, M. Lu, K. Huang, et al. 2024. The genetic architecture of repeated local adaptation to climate in distantly related plants. Nature Ecology & Evolution 8:1933–1947.

Whitlock, M. C., and D. E. McCauley. 1999. Indirect measures of gene flow and migration: FST≠1/(4Nm+1). Heredity 82:117–125.

Whittington, C. M., J. U. Van Dyke, S. Q. T. Liang, S. V. Edwards, R. Shine, M. B. Thompson, and C. E. Grueber. 2022. Understanding the evolution of viviparity using intraspecific variation in reproductive mode and transitional forms of pregnancy. Biological Reviews 97:1179–1192.

Wilding, C. S., R. K. Butlin, and J. Grahame. 2001. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. Journal of Evolutionary Biology 14:611–619.

Wohns, A. W., Y. Wong, B. Jeffery, A. Akbari, S. Mallick, R. Pinhasi, N. Patterson, et al. 2022. A unified genealogy of modern and ancient genomes. Science 375:eabi8264.

Wolf, J. B. W., and H. Ellegren. 2017. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics 18:87–100.

Wu, C. 2001. The genic view of the process of speciation. Journal of Evolutionary Biology 14:851–865.

Xu, W., C. Dubos, and L. Lepiniec. 2015. Transcriptional control of flavonoid biosynthesis by MYB–bHLH–WDR complexes. Trends in Plant Science 20:176–185.

Yeaman, S., and M. C. Whitlock. 2011. The Genetic architecture o adaptation under migration-selection balance: The genetic architecture of local adaptation. Evolution 65:1897–1911.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T. Lam. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28–36.

Zhang, B. C., A. Biddanda, Á. F. Gunnarsson, F. Cooper, and P. F. Palamara. 2023. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nature Genetics 1–9.

Zhang, K., P. Calabrese, M. Nordborg, and F. Sun. 2002. Haplotype block structure and its applications to association studies: Power and study designs. The American Journal of Human Genetics 71:1386–1394.

Zhang, T., W. Peng, H. Xiao, S. Cao, Z. Chen, X. Su, Y. Luo, et al. 2024. Population genomics highlights structural variations in local adaptation to saline coastal environments in woolly grape. Journal of Integrative Plant Biology 66:1408–1426.

Zhou, X., P. Carbonetto, and M. Stephens. 2013. Polygenic modelling with Bayesian sparse linear mixed models. PLOS Genetics 9:e1003264.

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44:821–824.

Zhu, S., Y. Zhang, L. Copsy, Q. Han, D. Zheng, E. Coen, and Y. Xue. 2023. The snapdragon genomes reveal the evolutionary dynamics of the S-Locus supergene. Molecular Biology and Evolution 40:msad080.

Appendix A

Supplementary Information for

On the origin and structure of haplotype blocks

Here, Appendix S2 associated with the original publication (Shipilina et al. 2023) of Chapter 2: "On the origin and structure of haplotype blocks" is detailed below. Appendix S1 associated with the original publication is available online at: https://doi.org/10.1111/mec.16793.

A.1 | Running *ARGweaver*

A.1.1 | Sample information

ARGweaver is applied on a phased SNP dataset of *Heliconius erato* butterflies, sequenced by haplotagging, a technique that produces linked-read sequence data. We used two previously published genomic regions—Herato1801:1362067–1405605 (coincides with the previously identified gene *optix* that has undergone a selective sweep) and Herato1603:3450000–3493538 (a neutral background region) (Meier et al. 2021). 10 individuals of *H. e. lativitta* and *H. e. notabilis* were chosen from opposite ends of the hybrid zone transect for ARG inference.

A.1.2 | Pre-processing files for ARGweaver

To pre-process the input files for *ARGweaver*, we first subset the 20 diploid individuals from full VCF files (see Supplementary Information in Meier et al, 2021), and then convert the SNP information into the *sites* format required by *ARGweaver*. The *sites* format only contains information on the positions that are varying within the 20 individuals that we chose. In the *optix* region (43538 bp long), there are 2812 sites altogether, of which 2426 are variant positions, while 330 and 56 sites are fixed for one or the other allele. Genomic positions that are neither variant nor fixed to one or the other allele (in other words, positions absent in the VCF file) are considered missing information and therefore masked from being used as input data for *ARGweaver*. Altogether, *ARGweaver* uses information of variant alleles and invariant alleles; whereas the rest is masked and treated as missing information. Similarly, the neutral region, Herato1603:3450000-3493538 (43538 bp) has 6407 sites altogether, of which 4926 is variant positions, while 1405 and 76 sites are fixed for either allele.

A.1.3 | Input parameters

In order to run *ARGweaver*, we consider mutation rate, μ = 2.9×10^{-9} per bp per generation, and its ratio to recombination rate, μ/r = 1. We estimate Ne by calculating Tajima's π (= $4N_e r$) from the neutral region. π=0.0225049, N_e=1940078. The total map length of both *optix* and neutral region is 0.01262602 cM.

ARGweaver allows coalescence and recombination events to take place only at discretized time points, defined by the function, $t(i) = \frac{\exp(log(1+\delta t_{max})/K)}{\delta}$; for *K* timepoints and $i \in \{0, 1, \ldots, K\}$. Very small values of δ ($< 1/t_{max}$) will roughly yield a linear distribution of times, whereas larger values of δ will place more time points in recent past and less in deep past. For this analysis, we set *K* = 30, the maximum time for total coalescence (t_{max}), and the shape parameter δ = 0.01 (Fig S1).

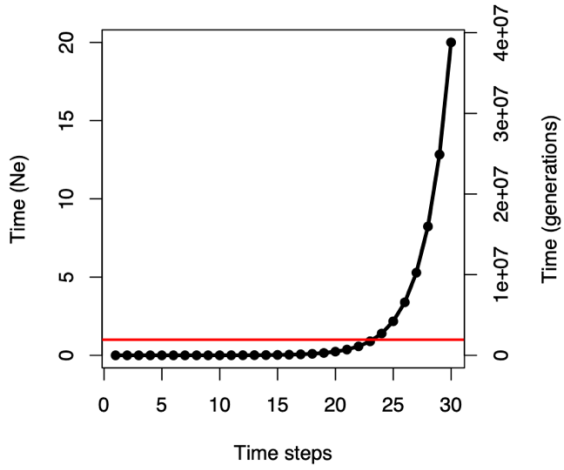

Figure S1. Discrete time points for *ARGweaver* analysis—30 discrete time steps, N_e = 1940078, Left y-axis shows the time points in N_e, while right y-axis in number of generations. Red line denotes 1 N_e.

A.1.4 | Run ARGweaver

ARGweaver is run for 5000 iterations and then resumed for another 5000 iterations for the optix region, whereas only 5000 iterations for the neutral region, with the following command *arg-sample –N 1940078 –m 2.9e-9 –r 2.9e-9 --ntimes 30 --iters 10000 --delta 0.01*.

A.2 | Analysis of *ARGweaver* output

A.2.1 | MCMC summary

After visualising MCMC traces of likelihood, priorand joint probabilities, we decided to set the first 3000 iterations as burnin. Figure S1 shows the *optix* region.

A.2.2 | ARGweaver output

The *ARGweaver* output, *.smc* files are then converted to *.bed* format to extract TMRCAs, trees, recombination breakpoints and total tree branch lengths across all iterations (except burn-in). Two iterations (Iteration 8250 and 9200) are chosen for further analysis and identification of haplotype blocks.

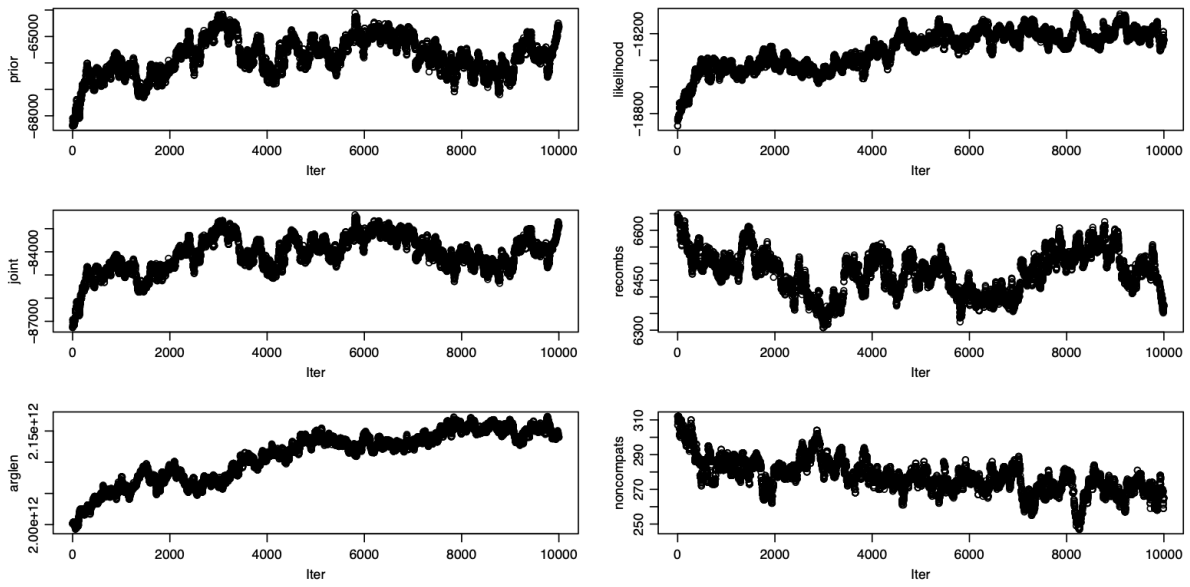

Figure S2. Traces across all MCMC iterations of *prior* (log probability of the sampled ARG given the model), *likelihood* (log probability of the data given the sampled ARG), *joint* (total log probability of the ARG and the data; prior + likelihood), *recombs* (number of recombination events in the sampled ARG), *arglen* (total length of all branches summed across sites) and *noncompats* (number of variant sites that cannot be explained by infinite sites mutation model).

A.2.3 | TMRCA

We first examine the TMRCA of the total tree and the individual populations - *H.e.lativitta* (in red) and *H.e.notabilis* (in yellow) (Fig S3). Unlike the neutral region (scaffold: Herato1603) where all TMRCA estimates are fairly constrained between 1 and 10 Ne, TMRCA traces for individual populations in the optix region exhibit shallow coalescence times at multiple positions throughout the entire ~50kb region. In case of a selective sweeps, where a beneficial mutation sweeps through the population, TMRCA tends to be shallow since all samples in the swept population coalesces quickly to the initial "lucky" ancestor. Moving away from the focal mutations, lineages recombine away from the swept ancestral background. The TMRCA estimates from the *optix* region is consistent with the previously identified selected region in *optix*. To investigate further the haplotype block structures, we focus into a 3kb region at optix:1385966-1388966.

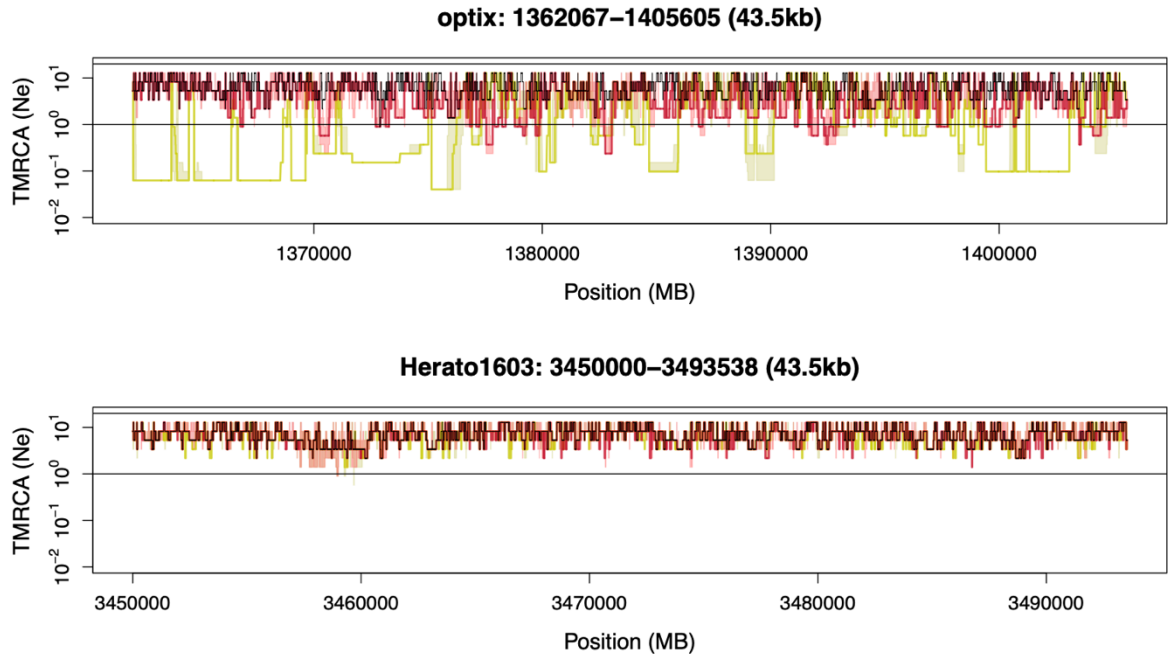

Figure S3. TMRCA (N_e) for each position in the *optix* and the neutral genomic region (scaffold: Herato1603). Black line: total TMRCA, Red: Median TMRCA for *H.e.lativitta*, Yellow: TMRCA for *H.e.notabilis*. Corresponding shaded regions are the 0.05 and 0.95 percentile intervals.

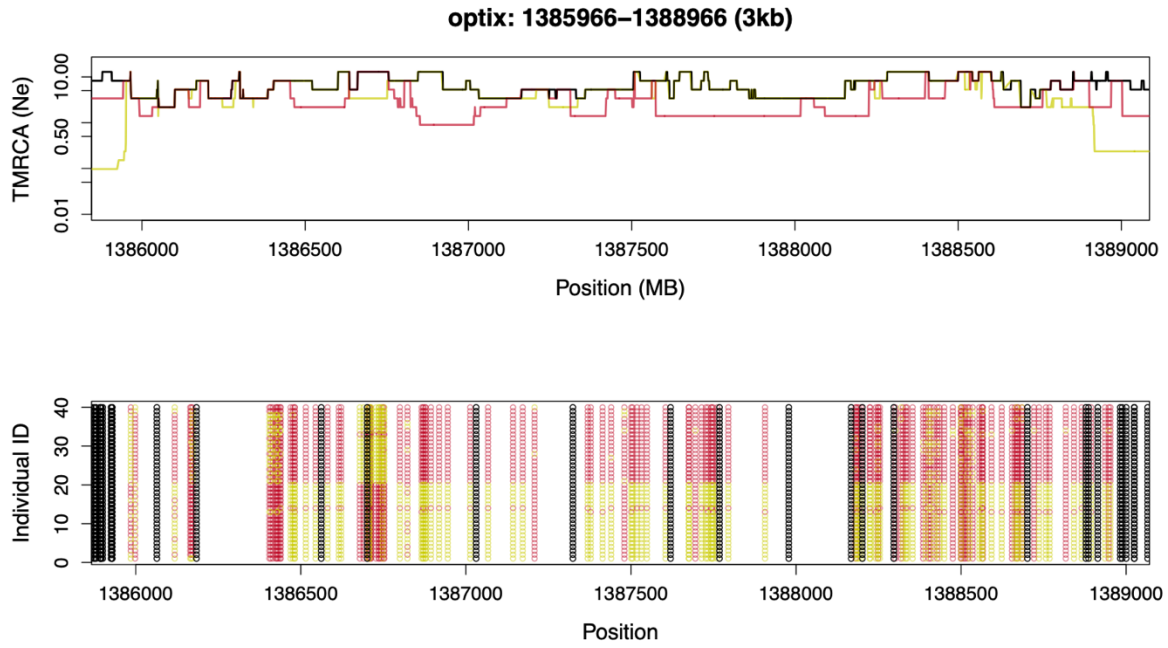

Figure S4. TMRCA for each position in focal genomic region (optix:1385966-1388966). Same colour schemes as above. Top panel: TMRCA plot, Bottom panel: SNPs in both populations; RED and YELLOW alleles are variant positions, coloured according to the respective higher and lower allele counts in the *H.e.lativitta* population. BLACK alleles are fixed (invariant) in both populations. All other positions (in white) are unknown, since those positions do not have SNP information. These unknown positions are masked while running *ARGWeaver*, therefore they are treated as missing information and not as invariant sites.

In the focal region, there are 137 SNPs. SNPs are changed from its nucleotide assignment to 0/1; based on allele frequency within the *H.e.lativitta* samples. (Major allele—1, Minor allele —0). Masking sites can have a critical effect on sampling ARGs, since invariant sites can shift priors towards a recent coalescence (because there has not been enough time yet for a mutation to occur in any of the branches), whereas missing information does not shift priors and therefore, ARGs are sampled neutrally in those regions. In the above region of 3kb, there are 137 SNPs and 16 invariant positions. Out of the 137 SNPs, only 59 SNPs are segregating within the *H.e.lativitta* population in focus (Fig S4) Hereafter, alleles with lower frequency within the *H.e.lativitta* population is referred to as the minor allele, and individuals who share the minor allele is the minor clade. Conversely, individuals that carry that major allele is called major clade. Minor allele is coded as 0 and major as 1.

A.3 | Specific MCMC iteration

A.3.1 | Case 1: Iteration 8250

We extracted ARGs sampled by ARGweaver in its MCMC iteration: 8250, and estimated the TMRCA (total and within population), total branch length of each marginal tree and the recombination breakpoints. Altogether in the ~50kb region, there are 6571 trees (6570 recombination events), of which only 464 trees are present in the focal 3kb region, optix:1385966-1388966.

We investigate the distribution of branch length of trees with respect to their genomic spans (Fig S5). Average tree span in the whole region is ~7 bp. From theory assuming standard coalescence, $P_r(d|\tau) = \frac{\rho}{2} L(\tau)\, e^{-\frac{\rho}{2} L(\tau) d}$ where $L(\tau)$ is in coalescent unit of 2 N_e generations and $\frac{\rho}{2} = 2N_e r$ denotes the population scaled recombination rate per bp. Given the parameters and mean $L(\tau) = 2$, mean $d = 1/(\frac{\rho}{2} L(\tau))$, should be ~45bp. In this region, *ARGweaver* seems to change the tree topology more often than expected. We have not explored closely the cause of this (since it is beyond the main message of this analysis), however, we note that although the ARGs are consistent with the data, we need to take caution in biological inference from these sampled ARGs.

For our analysis, we strictly focus on identifying edges supported by SNPs. In practice, this is done in few steps. First, for each sampled tree along the genome (= 464 trees), we extract the following information—ancestral and descendant node of each edge, edge height (=length), time-point of the descendant node, i.e., when the edge originated (=depth, due to rounding errors in Newick format, we round the depth values to 3 significant digits) and the samples (=tips.from.dec) that each edge is ancestral to. Although we are identifying haplotype blocks in the *H.e.lativitta* (red) population, we use genealogical trees that include both populations. This is done in order to estimate the edge height of the most recent common ancestor to all individuals in the red population. Ideally for making biological inferences from only one population, this need not be done. However in our analysis, in order to illustrate the haplotype block patterns generated in a selected genomic region, we decided to incorporate

both populations to generate trees along the genome. Nevertheless, it is important to note that the identified haplotype blocks will stay the same irrespective of which and how many populations are included in the analysis; however, the edge height will change along the genome.

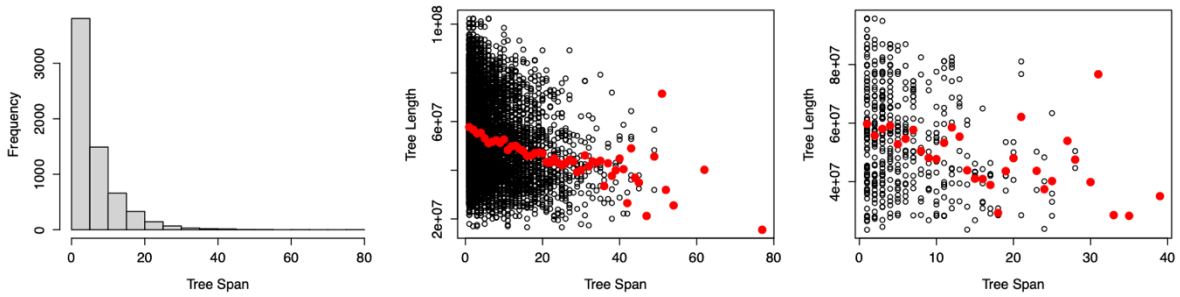

Figure S5. Left: Histogram of tree spans; Center: Branch length vs tree span for all 6571 marginal trees along the genome; Right: Branch length vs tree spans for 464 trees in the focal genomic region. RED points are average tree lengths for each tree span.

Second, for every tree at each SNP position (= 137 SNPs), we identify most recent tree node that is commonly ancestral to all individuals that share the same allele. (Note: this assumes infinite sites mutation model, which is generally the default option in *ARGweaver*). This allows identification of 1 node for the major and 1 for the minor allele at each SNP position, hereafter called major and minor node for each tree.

Third, for each minor node (and major node if the SNP is fixed within the *H.e.lativitta* population) identified from trees at each of the 137 SNPs, we identify all trees along the genome which contains that unique ancestral node. Since we do not know the ancestral reference allele at each SNP position and we assume infinite sites mutation, any minor node that is ancestral exclusively to the minor clade is assumed to contain the causal alternate allele and is considered as a branch on which a mutation has occurred.

For visualization, we choose to only plot the edges that are supported by 3 or more SNPs. Moreover, for SNPs that are fixed in the red population, we only show edges that originate at any time-point below 5Ne. This leaves us with only 8 edges. Now, plotting all the haplotype blocks except singletons (left panel in Fig S7), same as the figure in main text (see Main Text for caption).

A.3.2 | Case 2: Iteration 9200

There are altogether 6457 trees in the ~50kb region, and 425 in the focal genomic region. Figure S7 shows the haplotype blocks and the SNPs that support each block from both iterations.

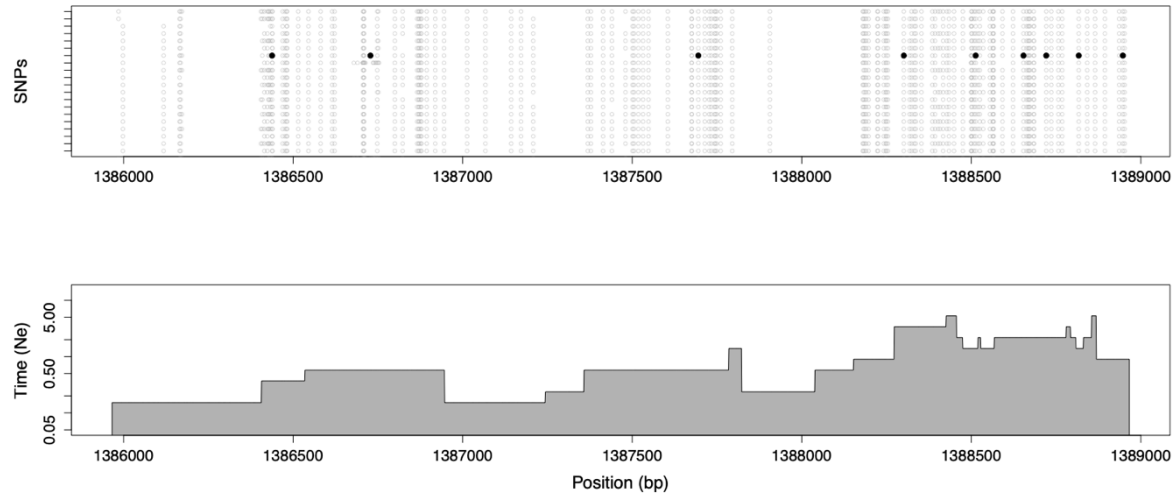

Figure S6: Haplotype block that is supported by singletons. These edges originate directly from the tree tips, i.e., the samples and therefore extends all along the genomic region. Although for biological inference, singletons are often uninformative, this shows the feature of an edge supported by singletons, which extends all along the genome, is normally shallow with certain regions that go are high, where a lot of singleton clusters together. These are normally regions of the genome, which have recombined out and goes all the way back to an ancestor in the deep past. The orange block shows clustering of SNPs in the higher region of the edges, whereas, the green edge shows how SNPs can also occur by chance at other regions.

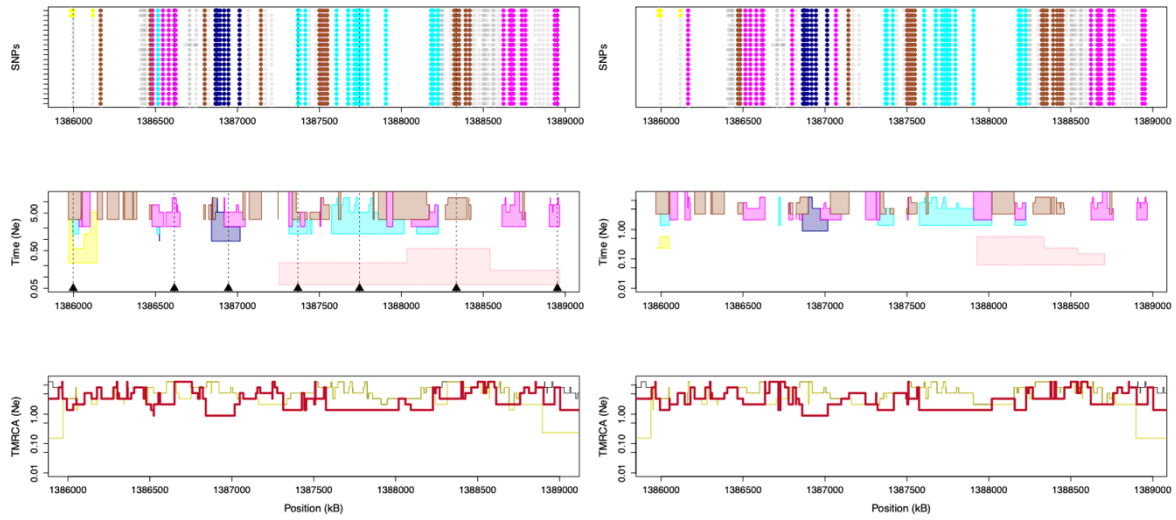

Figure S7:, Main Text for caption; Left panel - Iteration 8250, Right panel - Iteration 9200
Visualisation of haplotype blocks based on application of ARGweaver to the *optix* region of *Helconius erato* butterflies (*2n* = 20). ***Left panel***— MCMC Iteration 8250 (corresponds to Fig 5 in Main Text, Chapter 2), ***Right panel***—MCMC Iteration 9200. ***Top row:*** Genomic location of SNPs for all 20 *H.e.lativitta* haploid samples. Out of a total of 137 SNPs, only those that explain the 6 substantial edges are coloured accordingly. Edges are defined as substantial if 3 or more SNPs occur on them. ***Middle row:*** Visualisation of haplotype blocks as edges similar to Fig 3 (Main Text, Chapter 2) - plotting blocks along the genomic (*x-axis*) and temporal span (*y-axis*). ***Bottom row:*** Median TMRCA estimates along the genome: black: total TMRCA, red: TMRCA within *H.e.lativitta* samples, yellow: TMRCA within *H.e.notabilis* samples.

Appendix B

Supplementary Information for

The genetic basis of a recent transition to live bearing in marine snails

B.1 | Detailed results from the ternary analysis of simulated topology weights

In this section, we interpret the results of ternary analysis of simulated topology weights, focusing on the effects that different parameters have on the extent of lineage sorting (ILS) and asymmetry of the distribution. Each scenario is first discussed separately, and some general conclusions are drawn at the end.

Scenario a: Uniform N_e and even time between splits: This scenario gives a clear picture of how variation in the time between population splits shapes the distribution of topology weights (fig. S10; table S4). With a very small number of generations between splits, most loci are distributed very near to the center of the ternary plot, indicating that we see roughly equal proportions of the three possible topologies from a large sample of subtrees. This is expected, because when the durations between splits are much less than the average time to coalescence (*i.e.*, << $2N_e$ generations; here $2N_e$ = 1000 generations), most gene copies will coalesce in the common ancestor to all of the populations, rather than in the lineage that is ancestral to each spilt (Maddison 1997). In other words, ILS (aka 'deep coalescence'; Hudson, 1990) is extensive and dominates the relationships among the individuals. As the time between splits increases, we begin to see a bias toward the top of the ternary distribution, indicating a tendency of gene trees to show a greater resemblance to the species tree rather than the two alternative trees. However, we only see a strong shift toward the top of the plot as the time between splits approaches $2N_e$ generations. Only long after the average time to coalescence (*i.e.*, >> $2N_e$ generations) do we see perfect concordance between all gene trees and the species tree. This is also expected because the full distribution of coalescence times is expected to be very broad (*e.g.* the standard deviation of times is also $2N_e$ generations). Estimates of D_{LR} for these simulations are all low, and consistent with values expected by chance (p > 0.05). Thus, there is no effect of time between splits on the symmetry of the distribution (table S4).

Scenario b: Varying but even time between T2 and T3, with T1 set to 5k generations: This scenario shows how the effects of the split time play out over longer evolutionary time frames (fig. S11; table S4). The key result here is that the effects of ILS cannot be resolved even if a long period of evolution occurs after the splits. Rather, additional time only reinforces the patterns of discordance, resulting in trees that perfectly fit one of the three alternative topologies. However, the fraction of each topology observed is a function of time between splits (see the pie charts in fig. S11). For example, when the times between splits are short, (*i.e.*, when ILS is extensive) all three trees are observed at similar frequencies. However, as the times between splits increase, the fraction of the two alternative trees decreases, such that an increasing fraction of the loci show the background topology. Estimates of D_{LR} are not significantly different from 0 in any simulation, showing that a period of subsequent evolution has no effect on the symmetry of the distribution (table S4).

Scenario c: Uneven times between splits: This scenario shows how uneven time between splits shapes the distribution of weights. We can see, in both the 'short' and 'long' duration scenarios, that uneven time between splits has a small but pronounced effect on the concordance between gene-trees and the background demography (fig. S12; table S4). Specifically, lineage sorting is more complete when T_2 and T_3 occur deeper back in time. Thus, ILS has a more pronounced impact when a split occurs closer to the present. Estimates of D_{LR} show that the evenness of time between splits has no effect on the symmetry of the distribution (table S4).

Scenario d: Uniform variation in N_e: We know from theory that decreasing N_e has the same effect on lineage sorting as increasing the time between splits (Martin et al. 2015). In other, words the probability of lineage sorting increases with decreasing N_e. This is very clear from this scenario, where simulations were conducted with the same split times, but different values of N_e (fig. S13; table S4). At large values of N_e, the distribution of weights is centered in the middle of the ternary plot, but as the N_e is decreased, the distribution shifts toward the apex of the triangle. As with variation in the time between splits, uniform variation in N_e does not impact the symmetry of the distribution of weights (D_{LR} not significantly different from 0; table S4).

Scenario e: Varying N_e in one population: In this scenario, we alter the N_e only in P3, making it either smaller or larger than the background N_e, which was fixed for all other lineages (N_e = 500) (fig. S14; table S4). This has the effect of increasing or decreasing the probability that gene copies in that population will coalesce before the common ancestor to the ancestral population P123. When the N_e for P3 is larger than the background (*i.e.*, N_e = 5000), we see a stronger effect of ILS. However, when the N_e is smaller than the background, we see little change in the extent of lineage sorting. Rather we tend to see an increase in variance of the topology weights resulting in more extreme values. This makes intuitive sense, because if all gene copies in P3 coalesce before T_3, then all individuals will form a monophyletic group, rather than being randomly scattered across a tree. This will ultimately cause a given gene tree to look more like one of the alternative subtrees, but the overall level of discordance is controlled by the rate of ILS experienced by the other groups. Changing the

N_e of one population has no effect on the symmetry of the distribution (D_{LR} not significantly different from 0; table S4).

Scenarios f & g: Uni- and bi-directional migration between P2 and P3: In these scenarios, we simulated geneflow between P2 and P3 to see how it influences the distribution of weights. In the standard four population model, we generally expect gene copies in P1 and P2 to coalescence first. However, gene flow between non-sister taxa leads to the sharing of haplotypes between them, leading to shallower coalescence between P2 and P3. This effect is very clear in the simulations, as gene flow between P2 and P3 shifts the distribution of weights from the species tree toward the relevant alternative topology where these taxa cluster together (figs. S15 & S16; table S4). This effect is strongest in simulations where lineage sorting is more complete before the onset of gene flow, because, without a dominant topology to begin with, there is no bias in the distribution that gene flow can shift. This is most obvious when the time between splits are very short, and gene flow between P2 and P3 only leads to the exchange of variation that is already shared among all four groups. As expected, the estimates of D_{LR} are substantial and highly significant for most simulations, reflecting the strong asymmetry between the left and right sides of the plots (table S4). In general, the effect of gene flow becomes clearer as the migration rate increases, but only up to a point that is determined by the extent of lineage sorting before the onset of migration. Similarly, gene flow in both directions (*i.e.*, P3 to P2 and P2 to P3) results in more pronounced symmetry at lower migration rates, as there is twice as much migration between the groups.

Scenario h: Unidirectional migration for 10% of the genome: In this scenario, we combine two simulations—one with gene flow and one without—in order to simulate heterogenous gene flow across the genome, as is expected when species boundaries are porous (fig. S17; table S4). The effects of gene flow on the 10% of loci (red points) is clear in all but one of the simulations, where the time between splits is smallest and where the rate of migration is lowest. Even in the context of the symmetrical background distribution consisting of 90% of loci, the asymmetry in the simulations is detectable with the D_{LR} statistic which was highly significant (table S4). This highlights the potential utility of the ternary framework when it comes to detecting gene flow and introgression in genome-wide datasets.

Although neutral gene flow can generate strong asymmetry in some of these simulations, the pattern of asymmetry observed in our empirical data is not recovered in neutral simulations. First, windows that are a perfect fit for the alternative topology never became more abundant than those that perfectly fit the background tree (Table S8). In these simulations, we always observed a ratio of Tr:Tb much less than 1, with a maximum of 0.37 at an extreme migration rate of 50% between P3 and P2. In our observed data, we found a positive ratio, of 1.41, indicating that windows with Tr = 1 are more abundant than those with Tb = 1. Observing this result in a neutral model would require substantial gene flow (> 10%) for more than half of the genome. However, models with migration also fail to produce the otherwise flat-tailed distribution of topology weights for the alternative topology, as seen in Fig. 2E (for Tr and Tc). Rather, the distributions for the alternative tree Tr with migration always had a similar shape to that observed for the background tree (Fig 2E, as for Tb). Thus,

we conclude that realistic levels of neutral gene flow for a small fraction of the genome cannot produce the patterns of asymmetry that we observe in our empirical data (i.e., a strong excess of the alternative tree compared with the background tree, but with an otherwise flat-tailed distribution of weights). We think that selection on a small fraction of shared genome is needed to create this pattern.

Scenario i: Ancestral structure: This scenario, which follows that used Martin et al. (2015), is designed to model a region of the genome undergoing balancing selection or some other process that maintains polymorphism at particular loci in the ancestral lineage P123 (Fig. S18). In this scenario, 10% of loci evolve under a different history, which results in a bias toward the corresponding alternative topology. This is reflected in the non-zero estimates of D_{LR} in all of the simulations.

General conclusions: Based on the results of these simulations, we can make some general conclusions regarding how different factors shape the ternary distribution of topology weights. First, more time between splits and smaller effective population sizes (N_e) increase the probability of lineage sorting. This shifts the distribution of weights toward the top corner of the ternary plot, which corresponds to the topology that matches the background demography. No combination of split times or effective sizes causes an asymmetrical distribution of topology weights between the left and right sides of the ternary plot. These findings are intuitive, and predicted by coalescent theory. For example, in the absence of gene flow, the probability of observing each of the two discordant genealogies is equivalent and approximately $(1/3)e^{-t}$, where t is the length of the internal branch (Kemppainen et al. 2015).

Gene flow and ancestral structure can both produce distributions of weights that are asymmetrical. The ability to detect gene flow with a genome-wide estimate of D_{LR} will depend on the number of loci experiencing gene flow, the rate of migration (including whether it is uni- or bi-directional), and the extent of lineage sorting before gene flow commences. Finally, the pattern of asymmetry that we observed toward Tr is not consistent with what we found in neutral simulations.

B.2 | Supplementary Figures

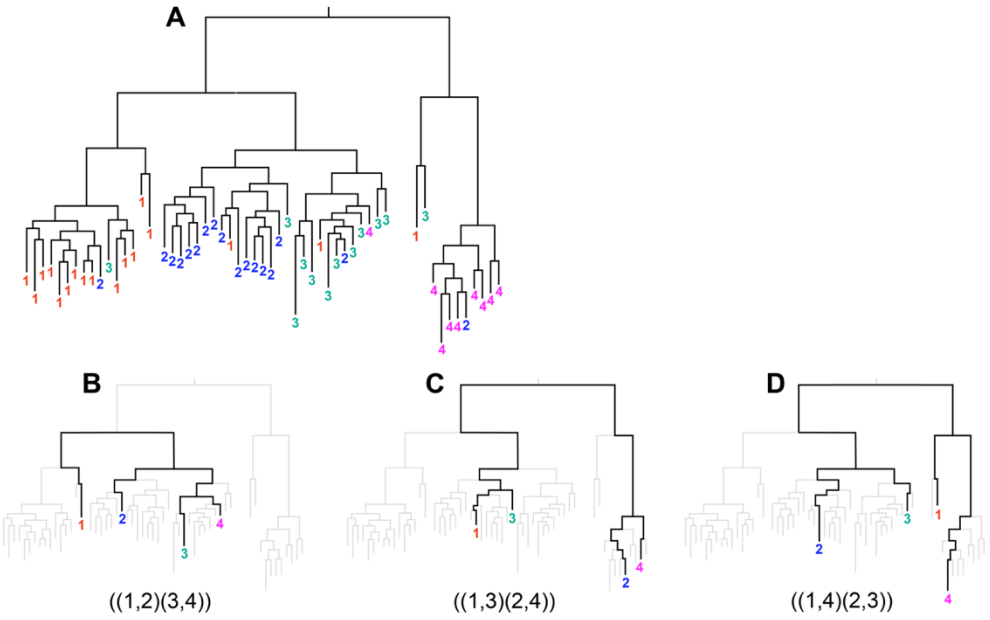

Figure S7. Alternative subtrees sampled from a large genealogy. (A) A hypothetical genealogy for a single non-recombining locus sampled from four species (1,2,3,4). The species do not form monophyletic clades, meaning that relationships between the species vary depending on the samples being considered. (B - D) Three example taxon subtrees sampled from the full tree showing the three possible unrooted topologies that can be observed in a tree with four taxa.

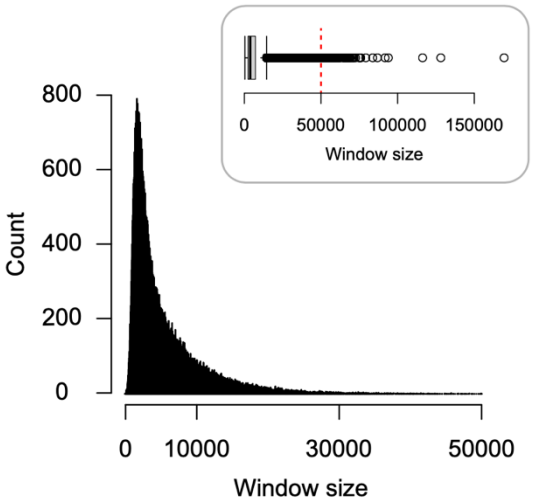

Figure S8. Distribution of physical sizes (kbp) of the 100 SNP windows used for local tree construction and topology weighting. The mean size is 5.8 kb ± s.d. 5.3 kb. The red dashed line in the top plot shows the margin of the main plot.

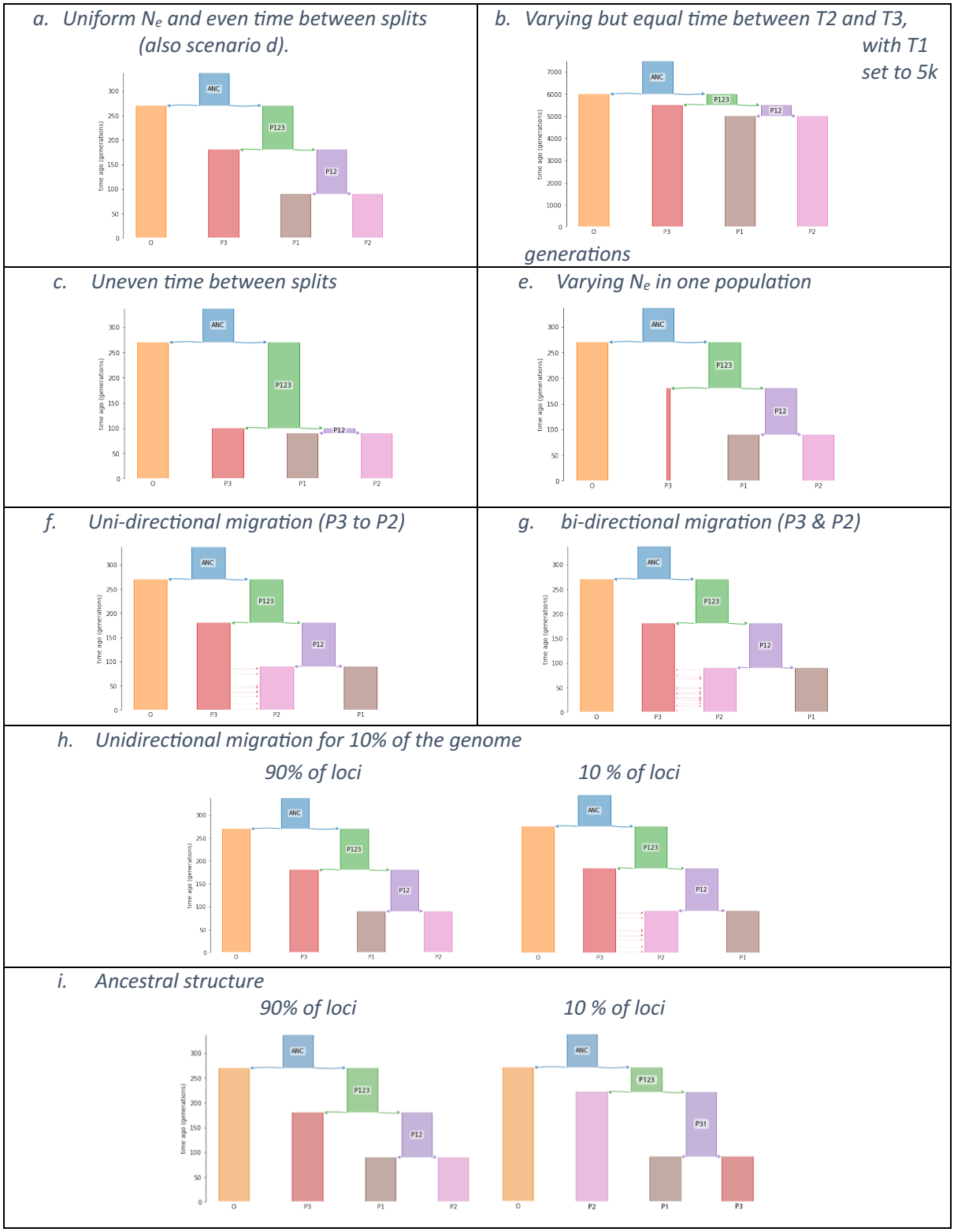

Figure S9. Cartoons of the models used in the simulations. Letters correspond with the scenarios described in the main text. Scenario a and d look the same

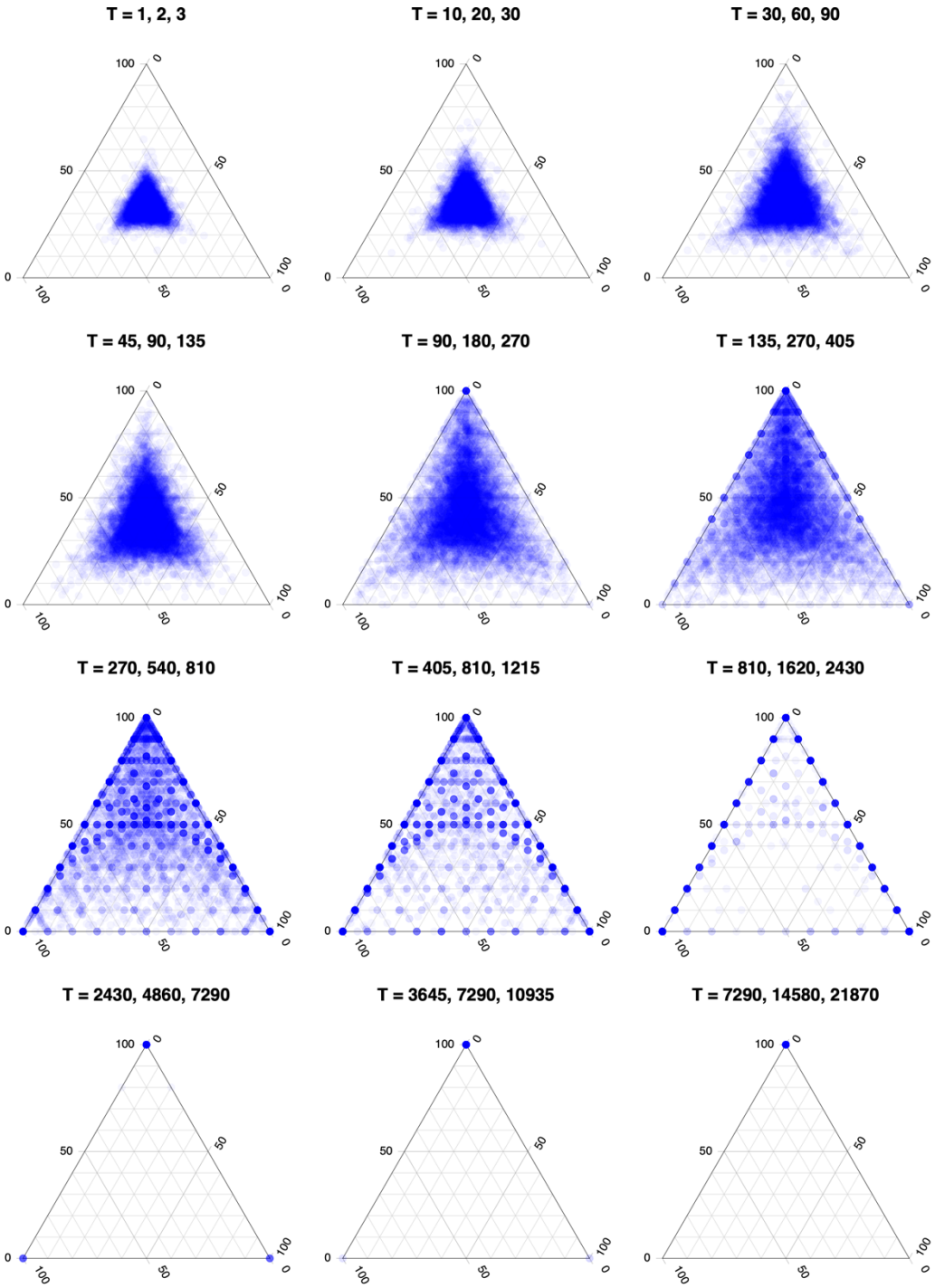

Figure S10. Joint distribution of topology weights for simulations with different but even time between splits. Each point is one of 10000 simulated loci. The time of the three splits (T_1, T_2, T_3) used for each simulation is given above each plot. All lineages have an N_e of 500, with no migration between them.

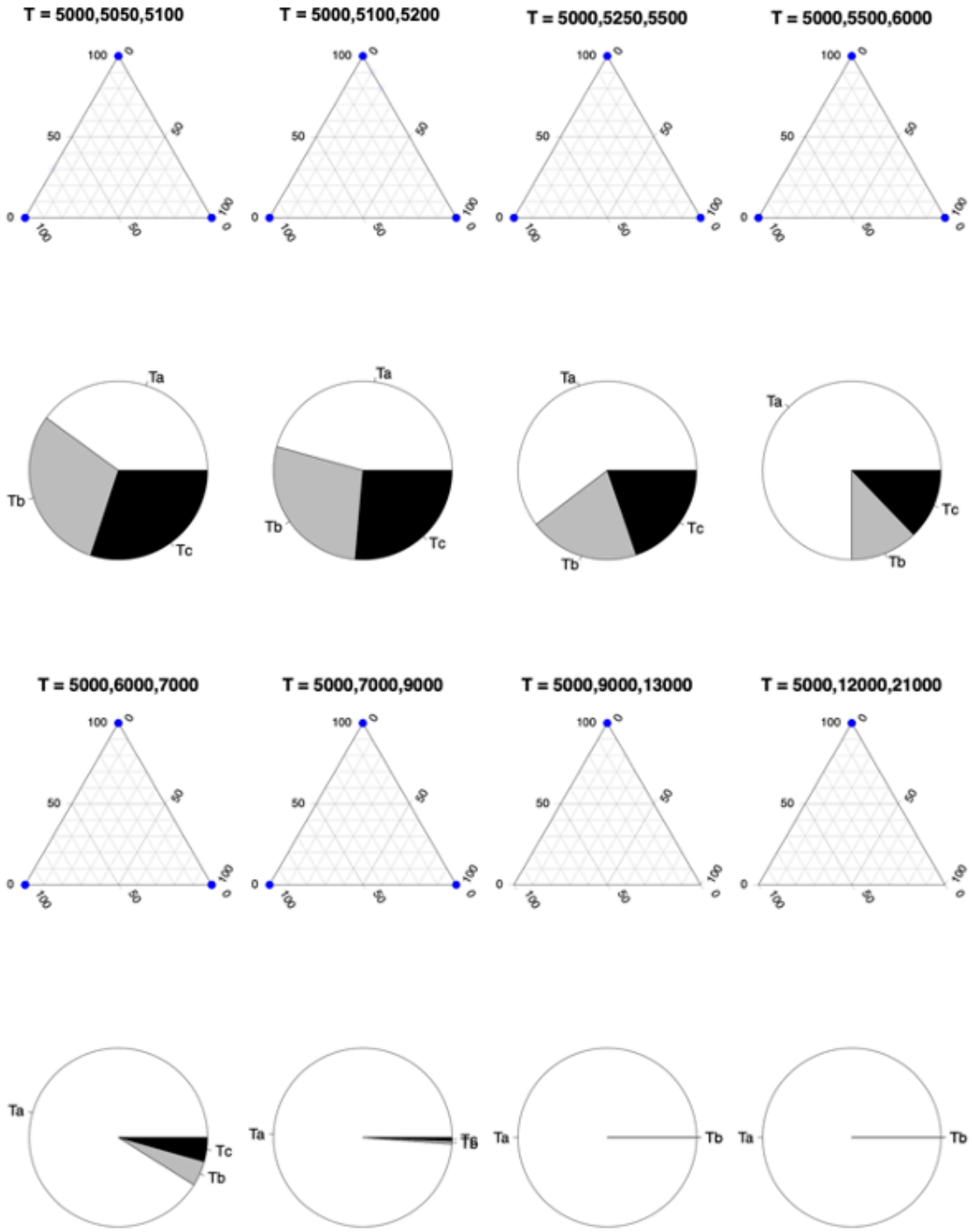

Figure S11. Joint distribution of topology weights for simulations with different but even time between splits, followed by 5000 generations after the final split. The time of the three splits (T_1, T_2, T_3) used for each simulation is given above each plot. All lineages have an N_e of 500, with no migration between them. The pie charts under each triangle show the fraction of trees with a topology that perfectly match one of the three subtrees Ta (apex of the triangle), Tb (left corner of the triangle), or Tc (right corner) for that simulation.

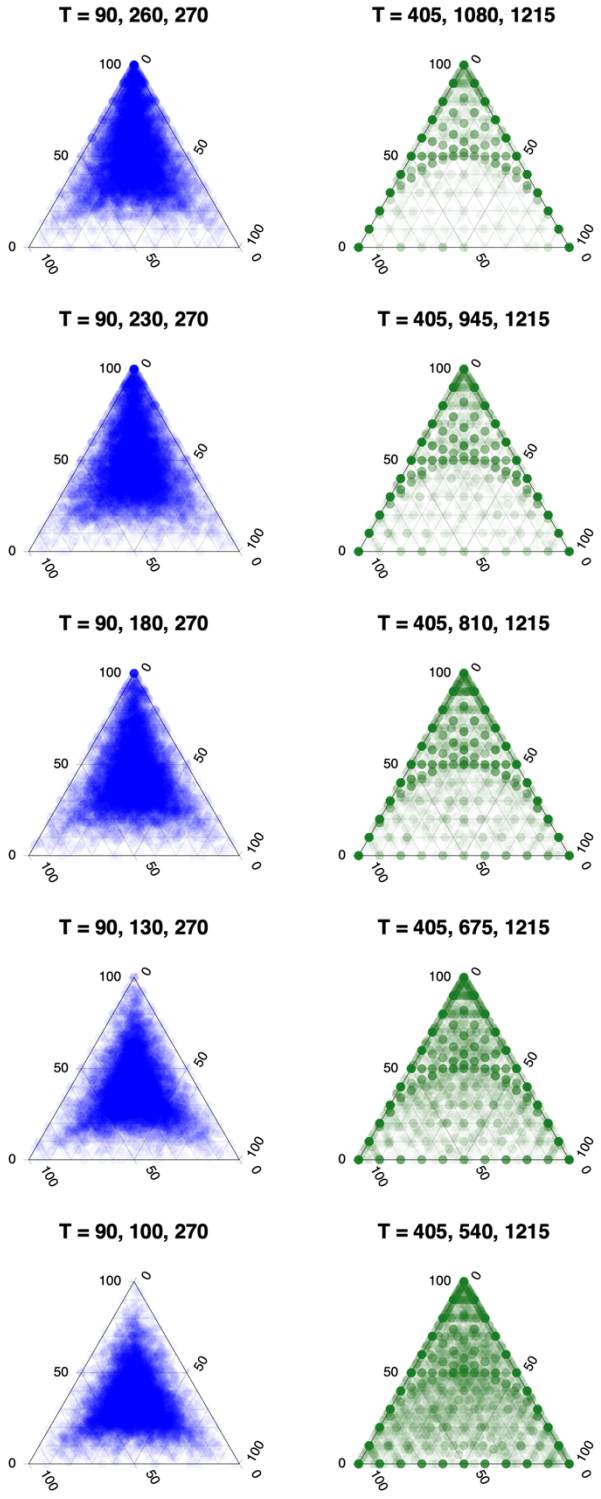

Figure S12. Distribution of topology weights for simulations with uneven time between splits. Each point is one of 10000 simulated loci. The columns show simulations with the same T_1 and T_3 splits, but with the T_2 varying between them. The split times used for all lineages are given above each plot. All simulations have an N_e of 500, with no migration between lineages.

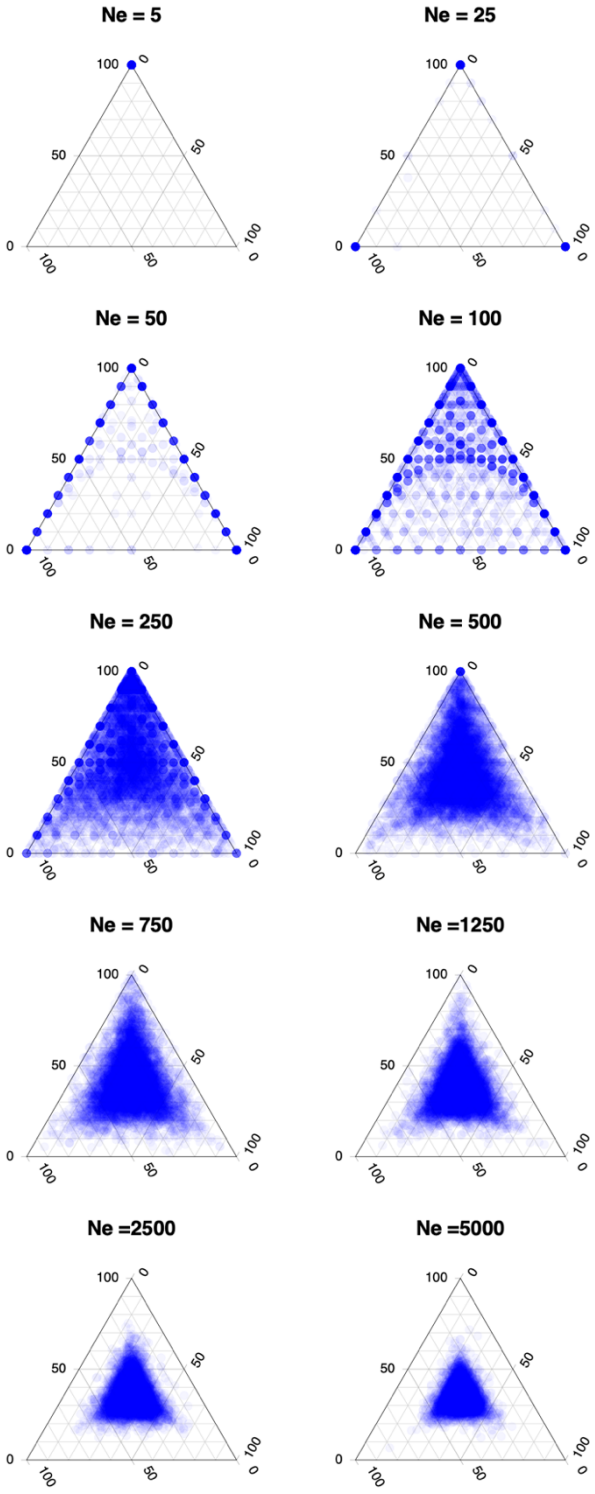

Figure S13. Joint distribution of topology weights for simulations with different effective populations sizes. Each point is one of 10000 simulated loci. The N_e used for all lineages is given above each plot. All simulations have the split times $T_1 = 90$, $T_2 = 180$, $T_3 = 270$, with no migration between lineages.

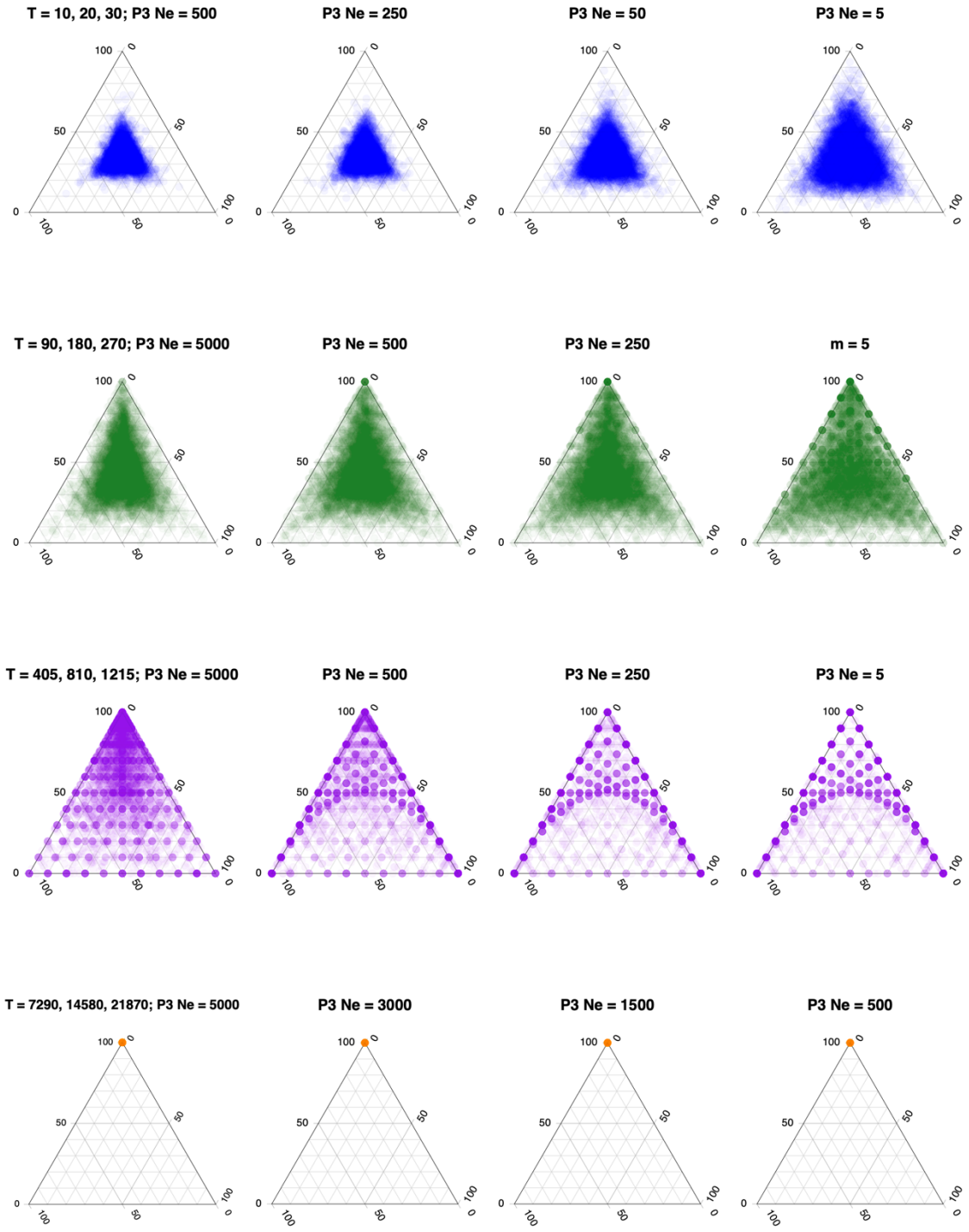

Figure S14. Joint distribution of topology weights for simulations where population P3 has a different effective population size than O, P1, P2. Each row shows four simulations with the same split times (T). The N_e used for lineage P3 is given above each plot. The N_e for all other populations is 500. The split times used for all lineages are given above each plot. No migration was allowed between lineages. Each point is one of 10000 simulated loci.

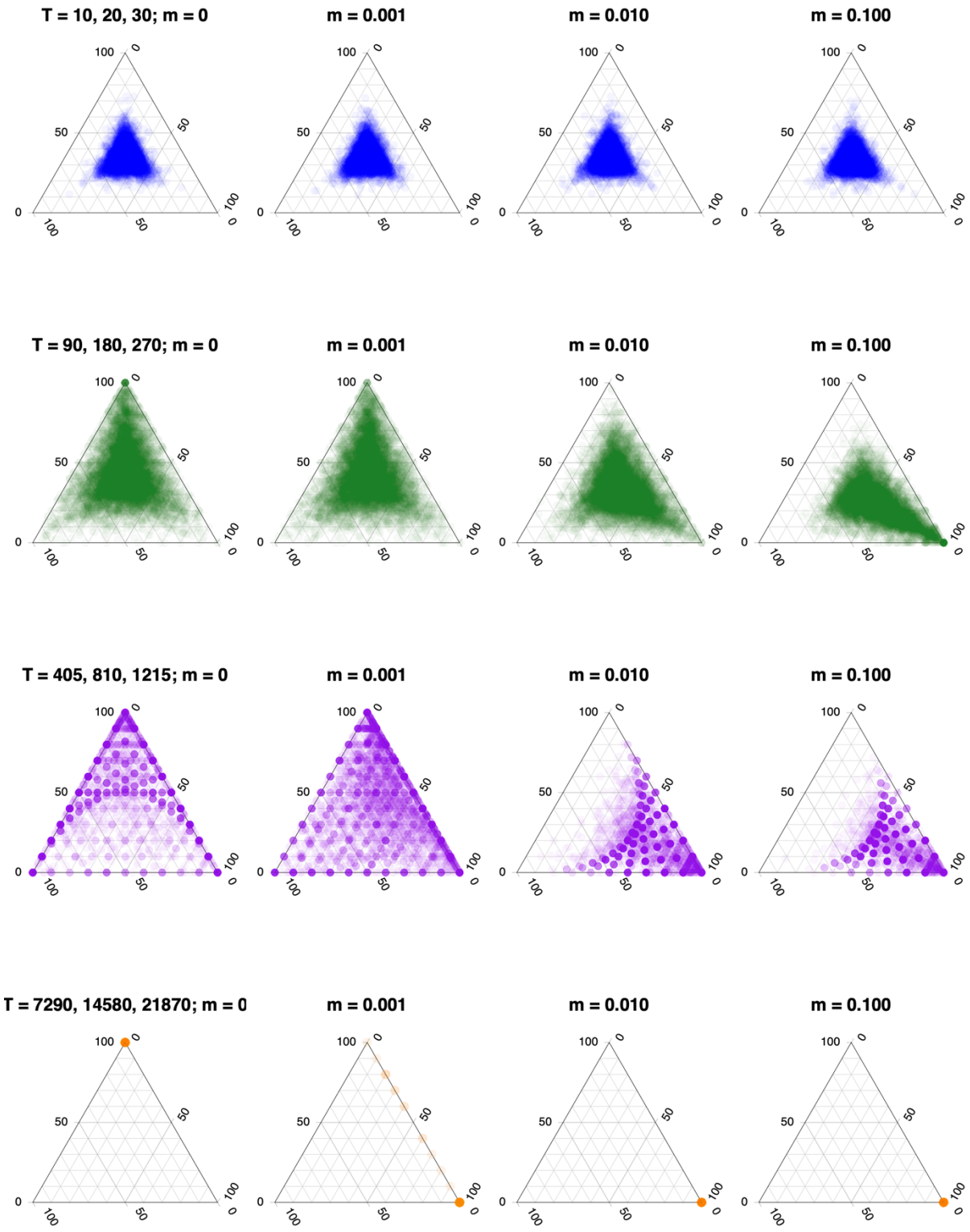

Figure S15. Joint distribution of topology weights for simulations with different rates of unidirectional migration. Each row shows four simulations with the same split times (T), but with different rates of migration from P2 to P3, as indicated above each plot (*m* = 0, 0.001, 0.010 or 0.100). All lineages have an N_e of 500. Each point is one of 10000 simulated loci.

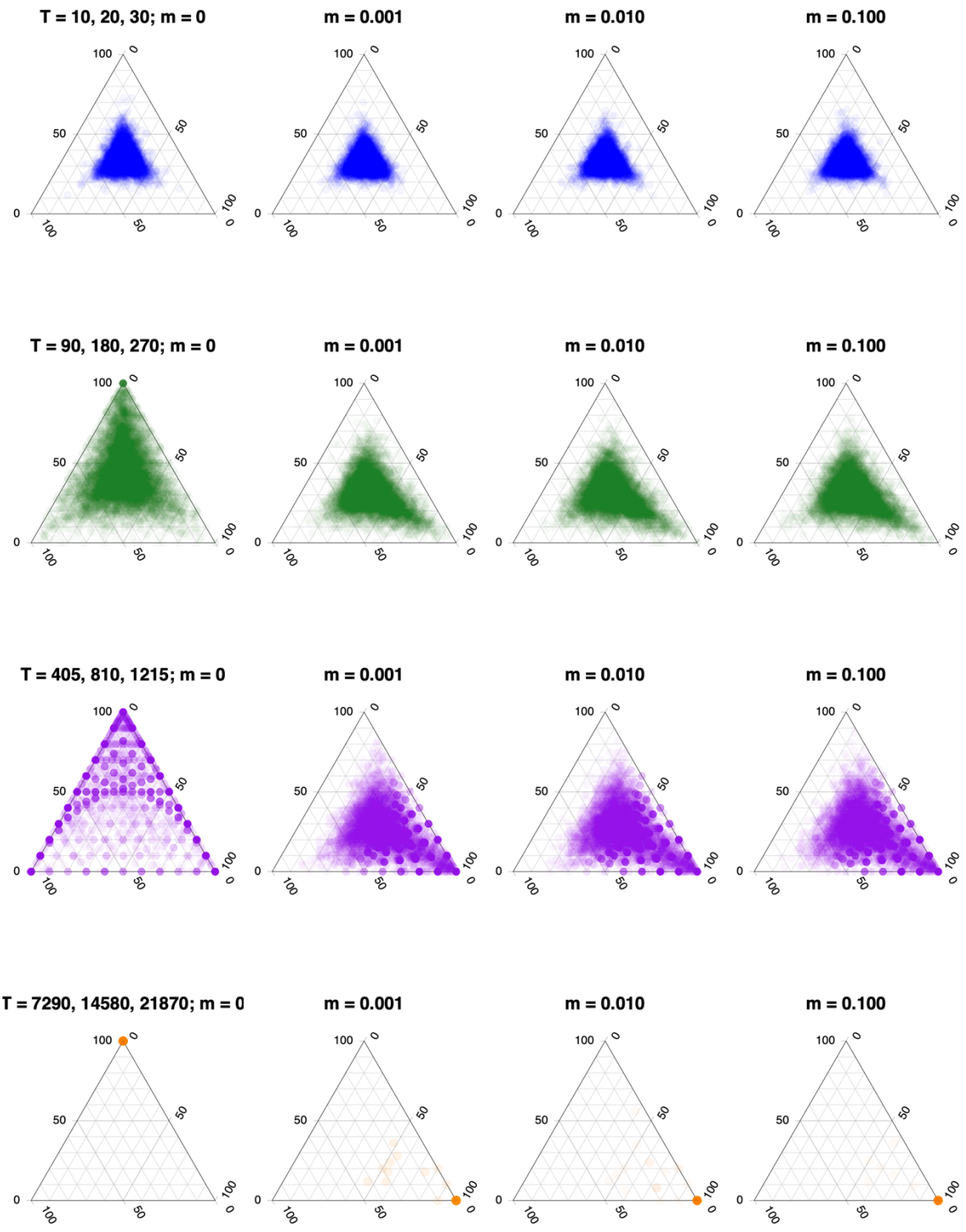

Figure S16. Joint distribution of topology weights for simulations with different rates of bidirectional migration. Each row shows four simulations with the same split times (T), but with different rates of migration from P2 to P3 and P3 to P2, as indicated above each plot (*m* = 0, 0.001, 0.010 or 0.100). All lineages have an N_e of 500. Each point is one of 10000 simulated loci.

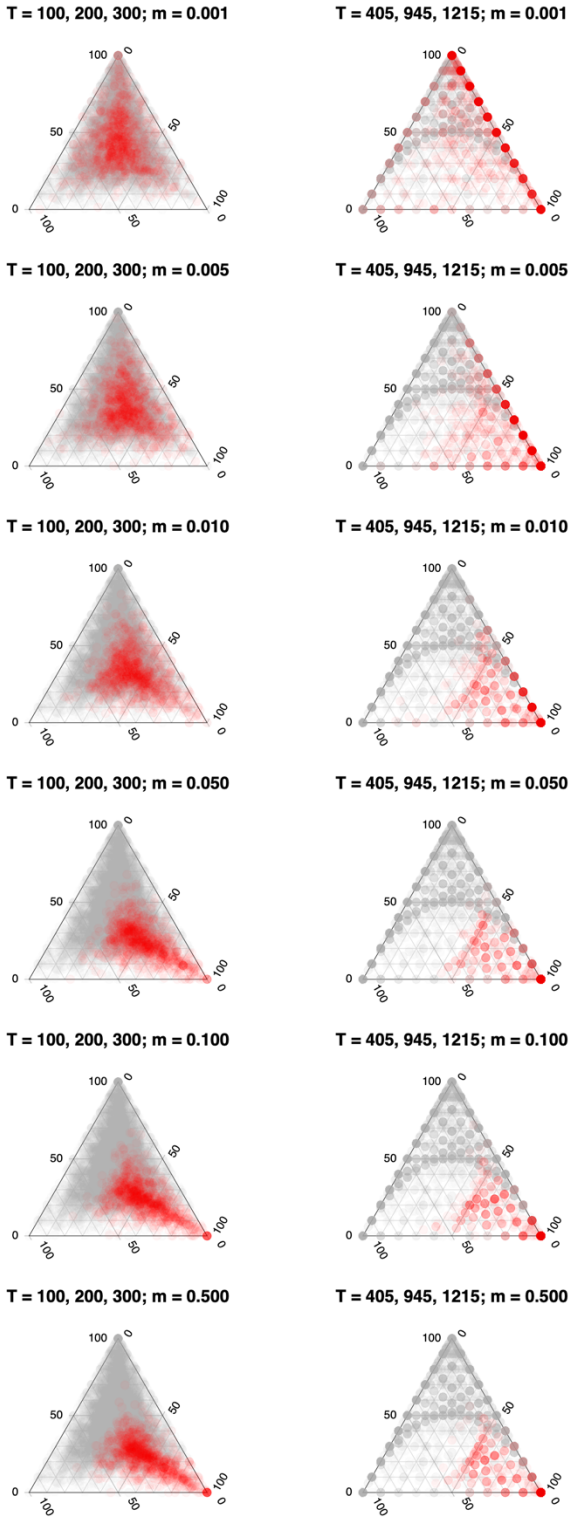

Figure S17. Joint distribution of topology weights for simulations with gene flow for 10% of the genome. Each column shows six simulations with the same split times (T), but with *m* = 0 for 90% of the genome (9000 blue dots) and *m* = 0, 0.001, 0.005, 0.010, 0.050, 0.100, or 0.500 for the 10% of the genome (1000 red dots). All lineages have an N_e of 500.

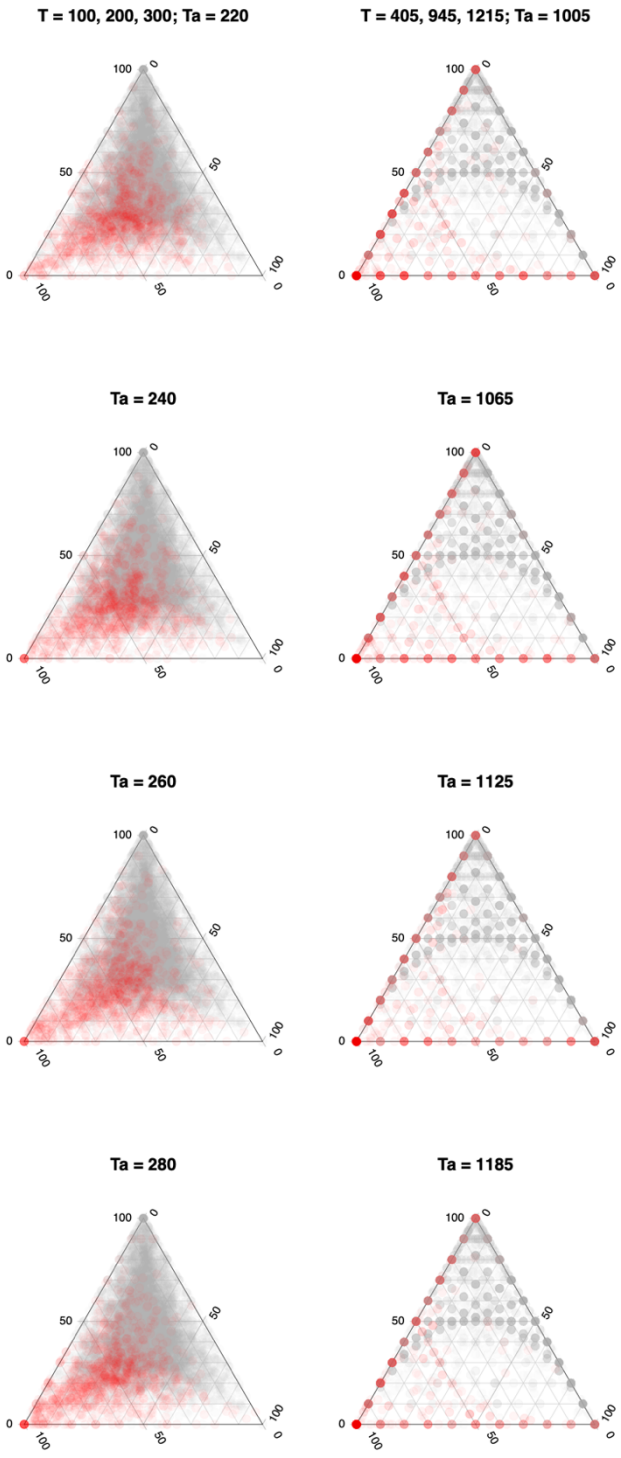

Figure S18. Joint distribution of topology weights for simulations with ancestral structure. Each column shows four simulations with the same split times for 90 % of genomic windows (gray dots). 10 percent of the genome evolves under an alternative demography (O,(P1(P2,P3))), with a higher split time for T2 to mimic a scenario with ancestral structure (red dots). The Ta values used are shown above each plot.

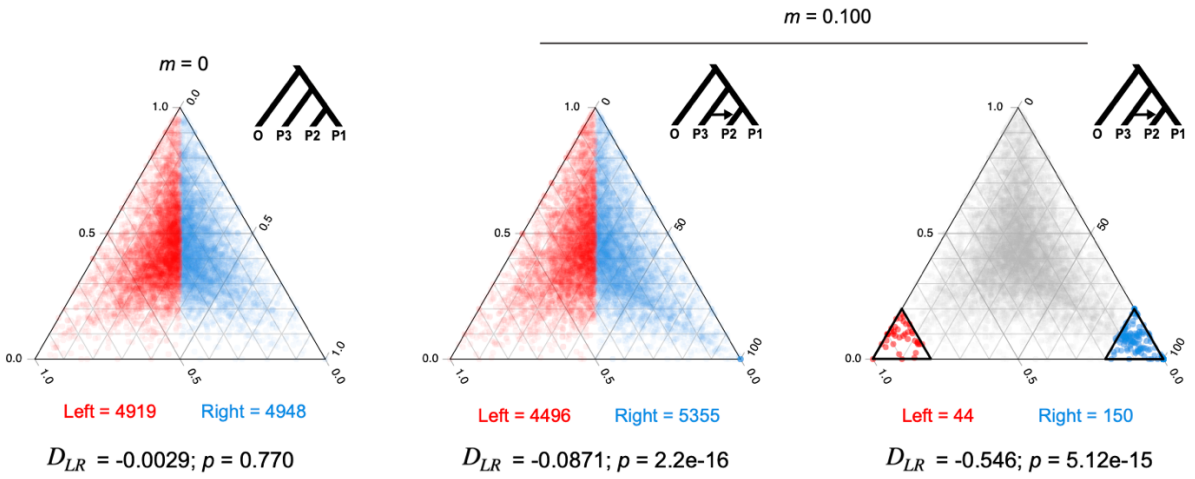

Figure S19. Example of symmetry analysis of simulated topology weights. The D_{LR} statistic calculated for two different simulated distributions, each consisting of 10,000 non-recombining loci. The counts of the left- and right-sided windows, estimate of D_{LR}, and significance of difference from 0 are given below each plot. Left-sided windows are coloured red and right-sided windows are blue. The left plot shows a distribution of topology weights calculated under an ideal four-population model with uniform effective sizes and uniform time between splits (Scenario a; fig. S9). A genome-wide test for asymmetry is conducted by calculating D_{LR} between the full left and right half triangles. The negative estimate of D_{LR} (-0.0029) indicates that there is a small bias toward Tc, with 0.29% fewer windows in the left side of the plot than expected (*i.e.*, There is an equal probability of a point falling on the left or right side). However, the observed bias is not significantly different from the expectation of equality (G-test, p = 0.770). The two plots on the right side show a different distribution of weights, with 90% of the genome simulated under a model without migration but where migration occurs at a rate of 0.1 between P3 and P2 for 10 % of loci (Scenario h; fig. S9). In the middle plot, the genome-wide estimate of D_{LR} (-0.0871) indicates an 8.7% excess of windows on the right side of the plot, which is much greater than expected by chance (G-test, p = 5.12e-15). In the right plot, D_{LR} is estimated between the extreme left- and right-sided sub triangles. In this case, there is very strong and significant asymmetry, with 54% more windows on the right than we would expect by chance (D_{LR} = -0.546, p = 5.12e-15).

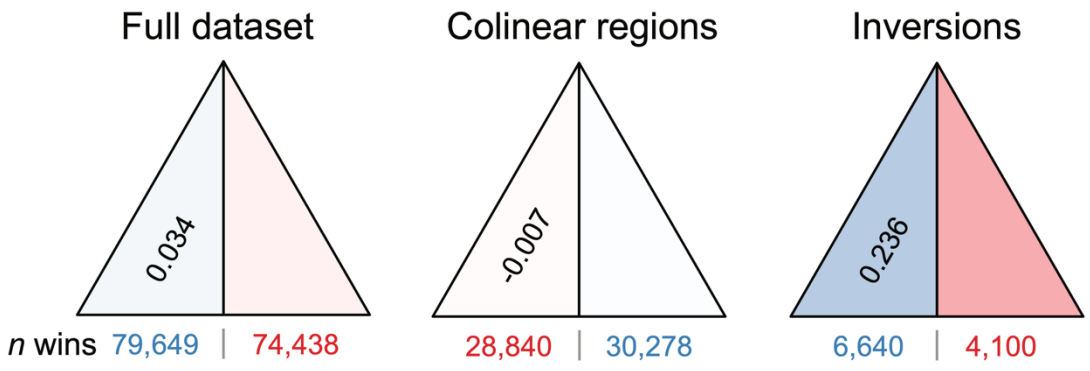

Figure S20. Analysis of symmetry for different partitions of the genome. Results for the full dataset, colinear regions, and chromosomal inversions are show. D_{LR} is provided in the left-hand sub-triangle. The intensity of shading in each side is proportional to the excess (blue) or deficit (red) of windows relative to the expectation. The numbers of windows falling on each side are shown under the plot.

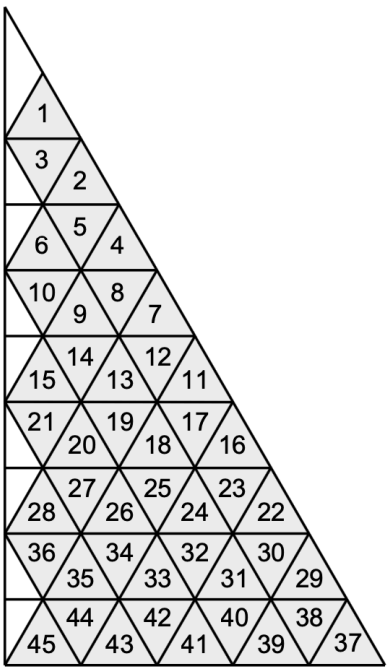

Figure S24. Identities of the sub-triangles analysed in the fine scale symmetry analysis. The numbers correspond to the sub-triangle IDs given in table S6 and main text Fig. 2D. The central sub-triangles of the ternary plot (here, seen as small white right-angled triangles on the left side of the plot) cannot be divided into triangles at the same scale so were not considered in the analysis.

B.3 | Supplementary Tables

Table S3. Parameters of the four-population coalescent model implemented in MSprime. N_e, number of haploid sequences, *m* number of haploid sequences per generation moving from one population to another. T, time in generations.

Parameter	Description
N_eP1	The effective size of population 1
N_eP2	The effective size of population 2
N_eP3	The effective size of population 3
N_eO	The effective size of population O
N_eP12	The effective size of the ancestor of P1,P2
N_eP123	The effective size of the ancestor of P1,P2,P3
N_e*Anc*	The effective size of the ancestor of P1,P2,P3,O
T_1	The time of the split time of P1 and P2
T_2	The split time of P3 and P12
T_3	The split time of O and P123
$m_{2,3}$	The migration rate from P2 to P3
$m_{3,2}$	The migration rate from P3 to P2

Table S4. Parameter values for simulating genealogies in MSprime and estimates of D_{LR} calculated from the resulting topology weights. Each line corresponds to one simulation. Each shaded section corresponds to one of the simulation scenarios described in the text (a - i). SimID: The ID of the simulation, which corresponds with the IDs used in the plotting code and output files on GitHub; Wins: the number of windows simulated for each simulation. *n*, number of individuals simulated for each population; N_e, number of haploid sequences simulated in each of the lineages specified (illustrated in Fig. S9); Splits: the time of each split in generations (Illustrated in fig. S9); Migration: the rates of migration from P3 to P2 and P2 to P3 (Illustrated in fig. S12); D_{LR}, value of D_{LR} calculated for each simulation, with negative values indicating a deficit of windows on the left side. If 'NA', D_{LR} could not be calculated because there were no left or right sided topologies (*i.e.*, lineage sorting was complete). For scenarios *h* and *i*, the values outside parentheses are for the 'alternate' model (10% of simulated loci), while values inside are for the 'background' model (90% of simulated loci); p-value, the p-value for a G-test of the departure of D_{LR} using 0.5 as the expected proportion of left and right sided windows. The p-value is reported as 'NA' for simulations where D_{LR} could not be calculated. The table continues on the next page.

				N_e							splits (Gens.)			Migration			
	SimID	wins	*n*	P1	P2	P3	O	P12	P123	ANC	T1	T2	T3	3>2	2>3	DLR	*p-value*
a) Varying but equal splits	N0	10000	20	500	500	500	500	500	500	500	1	2	3	0	0	0.003	0.734
	N1	10000	20	500	500	500	500	500	500	500	10	20	30	0	0	-0.004	0.674
	N2	10000	20	500	500	500	500	500	500	500	30	60	90	0	0	-0.026	0.008
	N2.5	10000	20	500	500	500	500	500	500	500	45	90	135	0	0	-0.004	0.689
	N3	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.003	0.770
	N3.5	10000	20	500	500	500	500	500	500	500	135	270	405	0	0	-0.002	0.838
	N4	10000	20	500	500	500	500	500	500	500	270	540	810	0	0	0.002	0.890
	N4.5	10000	20	500	500	500	500	500	500	500	405	810	1215	0	0	-0.008	0.560
	N5	10000	20	500	500	500	500	500	500	500	810	1620	2430	0	0	-0.017	0.467
	N6	10000	20	500	500	500	500	500	500	500	2430	4860	7290	0	0	0.043	0.768
	N7	10000	20	500	500	500	500	500	500	500	3645	7290	10935	0	0	-0.143	0.705
	N8	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0	0	NA	NA
b) 5k gens after T1	Y0	10000	20	500	500	500	500	500	500	500	5000	5050	5100	0	0	0.002	0.897
	Y1	10000	20	500	500	500	500	500	500	500	5000	5100	5200	0	0	-0.029	0.034
	Y2	10000	20	500	500	500	500	500	500	500	5000	5250	5500	0	0	-0.005	0.775
	Y3	10000	20	500	500	500	500	500	500	500	5000	5500	6000	0	0	0.024	0.230
	Y4	10000	20	500	500	500	500	500	500	500	5000	6000	7000	0	0	-0.019	0.569
	Y5	10000	20	500	500	500	500	500	500	500	5000	7000	9000	0	0	0.058	0.524
	Y6	10000	20	500	500	500	500	500	500	500	5000	9000	13000	0	0	NA	NA
	Y7	10000	20	500	500	500	500	500	500	500	5000	12000	21000	0	0	NA	NA
c) Uneven splits	U0	10000	20	500	500	500	500	500	500	500	90	260	270	0	0	-0.002	0.878
	U1	10000	20	500	500	500	500	500	500	500	90	230	270	0	0	-0.004	0.670
	U2	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.002	0.879
	U3	10000	20	500	500	500	500	500	500	500	90	130	270	0	0	-0.010	0.326
	U4	10000	20	500	500	500	500	500	500	500	90	100	270	0	0	-0.003	0.756
	U5	10000	20	500	500	500	500	500	500	500	405	540	1215	0	0	-0.002	0.869
	U6	10000	20	500	500	500	500	500	500	500	405	675	1215	0	0	-0.002	0.837
	U7	10000	20	500	500	500	500	500	500	500	405	810	1215	0	0	0.018	0.188
	U8	10000	20	500	500	500	500	500	500	500	405	945	1215	0	0	0.034	0.027
	U9	10000	20	500	500	500	500	500	500	500	405	1080	1215	0	0	-0.005	0.794
d) Variation in Ne	S0	10000	20	5	5	5	5	5	5	5	90	180	270	0	0	NA	NA
	S1	10000	20	25	25	25	25	25	25	25	90	180	270	0	0	NA	NA
	S2	10000	20	50	50	50	50	50	50	50	90	180	270	0	0	-0.008	0.757
	S3	10000	20	100	100	100	100	100	100	100	90	180	270	0	0	-0.001	0.932
	S4	10000	20	250	250	250	250	250	250	250	90	180	270	0	0	-0.004	0.720
	N3	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.003	0.770
	S5	10000	20	750	750	750	750	750	750	750	90	180	270	0	0	0.014	0.176
	S6	10000	20	1250	1250	1250	1250	1250	1250	1250	90	180	270	0	0	-0.001	0.936
	S7	10000	20	2500	2500	2500	2500	2500	2500	2500	90	180	270	0	0	-0.009	0.352
	S8	10000	20	5000	5000	5000	5000	5000	5000	5000	90	180	270	0	0	0.001	0.912

				N_e							splits (Gens.)			Migration			
	SimID	wins	*n*	P1	P2	P3	O	P12	P123	ANC	T1	T2	T3	3>2	2>3	D_{LR}	*p-value*
e) Varying Ne in one population	N1	10000	20	500	500	500	500	500	500	500	10	20	30	0	0	-0.004	0.674
	X1	10000	20	500	500	250	500	500	500	500	10	20	30	0	0	0.002	0.857
	X2	10000	20	500	500	50	500	500	500	500	10	20	30	0	0	-0.017	0.087
	X3	10000	20	500	500	5	500	500	500	500	10	20	30	0	0	-0.004	0.682
	X4	10000	20	500	500	5000	500	500	500	500	90	180	270	0	0	-0.010	0.320
	N3	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.003	0.770
	X5	10000	20	500	500	250	500	500	500	500	90	180	270	0	0	-0.014	0.179
	X6	10000	20	500	500	5	500	500	500	500	90	180	270	0	0	0.004	0.661
	X7	10000	20	500	500	5000	500	500	500	500	405	810	1215	0	0	-0.012	0.334
	N4.5	10000	20	500	500	500	500	500	500	500	405	810	1215	0	0	-0.008	0.560
	X8	10000	20	500	500	250	500	500	500	500	405	810	1215	0	0	0.005	0.719
	X9	10000	20	500	500	5	500	500	500	500	405	810	1215	0	0	-0.007	0.608
	X10	10000	20	500	500	5000	500	500	500	500	7290	14580	21870	0	0	NA	NA
	X11	10000	20	500	500	3000	500	500	500	500	7290	14580	21870	0	0	NA	NA
	X12	10000	20	500	500	1500	500	500	500	500	7290	14580	21870	0	0	NA	NA
	N8	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0	0	NA	NA
f) Uni-directional migration	N1	10000	20	500	500	500	500	500	500	500	10	20	30	0	0	0.003	0.734
	G1	10000	20	500	500	500	500	500	500	500	10	20	30	0.001	0	0.001	0.904
	G2	10000	20	500	500	500	500	500	500	500	10	20	30	0.01	0	-0.020	0.050
	G3	10000	20	500	500	500	500	500	500	500	10	20	30	0.1	0	-0.174	2.20E-16
	N3	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.003	0.770
	G4	10000	20	500	500	500	500	500	500	500	90	180	270	0.001	0	-0.094	2.20E-16
	G5	10000	20	500	500	500	500	500	500	500	90	180	270	0.01	0	-0.510	2.20E-16
	G6	10000	20	500	500	500	500	500	500	500	90	180	270	0.1	0	-0.777	2.20E-16
	N4.5	10000	20	500	500	500	500	500	500	500	405	810	1215	0	0	-0.008	0.560
	G7	10000	20	500	500	500	500	500	500	500	405	810	1215	0.001	0	-0.585	2.20E-16
	G8	10000	20	500	500	500	500	500	500	500	405	810	1215	0.01	0	-0.950	2.20E-16
	G9	10000	20	500	500	500	500	500	500	500	405	810	1215	0.1	0	-0.981	2.20E-16
	N8	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0	0	NA	NA
	G10	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.001	0	-1.000	2.20E-16
	G11	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.01	0	-1.000	2.20E-16
	G12	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.1	0	-1.000	2.20E-16
g) Bi-directional migration	N1	10000	20	500	500	500	500	500	500	500	10	20	30	0	0	-0.004	0.674
	B1	10000	20	500	500	500	500	500	500	500	10	20	30	0.001	0.001	-0.084	2.20E-16
	B2	10000	20	500	500	500	500	500	500	500	10	20	30	0.01	0.01	-0.098	2.20E-16
	B3	10000	20	500	500	500	500	500	500	500	10	20	30	0.1	0.1	-0.095	2.20E-16
	N3	10000	20	500	500	500	500	500	500	500	90	180	270	0	0	-0.003	0.7703
	B4	10000	20	500	500	500	500	500	500	500	90	180	270	0.001	0.001	-0.349	2.20E-16
	B5	10000	20	500	500	500	500	500	500	500	90	180	270	0.01	0.01	-0.346	2.20E-16
	B6	10000	20	500	500	500	500	500	500	500	90	180	270	0.1	0.1	-0.361	2.20E-16
	N4.5	10000	20	500	500	500	500	500	500	500	405	810	1215	0	0	-0.008	0.560
	B7	10000	20	500	500	500	500	500	500	500	405	810	1215	0.001	0.001	-0.735	2.20E-16
	B8	10000	20	500	500	500	500	500	500	500	405	810	1215	0.01	0.01	-0.724	2.20E-16
	B9	10000	20	500	500	500	500	500	500	500	405	810	1215	0.1	0.1	-0.717	2.20E-16
	N8	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0	0	NA	NA
	B10	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.001	0.001	-1.000	2.20E-16
	B11	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.01	0.01	-1.000	2.20E-16
	B12	10000	20	500	500	500	500	500	500	500	7290	14580	21870	0.1	0.1	-1.000	2.20E-16
h) Migration for 10% of the genome	I1	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.001 (0)	0	-0.013	0.200
	I2	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.005 (0)	0	-0.041	4.10E-05
	I3	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.01 (0)	0	-0.057	1.17E-08
	I4	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.05 (0)	0	-0.082	2.20E-16
	I5	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.1 (0)	0	-0.085	2.20E-16
	I6	1000 (9000)	20	500	500	500	500	500	500	500	100	200	300	0.5 (0)	0	-0.087	2.20E-16
	I7	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.001 (0)	0	-0.125	2.20E-16
	I8	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.005 (0)	0	-0.188	2.20E-16
	I9	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.01 (0)	0	-0.197	2.20E-16
	I10	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.05 (0)	0	-0.201	2.20E-16
	I11	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.1 (0)	0	-0.200	2.20E-16
	I12	1000 (9000)	20	500	500	500	500	500	500	500	405	945	1215	0.5 (0)	0	-0.201	2.20E-16
i) Ancestral structure	A1	1000 (9000)	20	500	500	500	500	500	500	500	100	220 (200)	300	0	0	0.055	4.69E-08
	A2	1000 (9000)	20	500	500	500	500	500	500	500	100	240 (200)	300	0	0	0.061	1.56E-09
	A3	1000 (9000)	20	500	500	500	500	500	500	500	100	260 (200)	300	0	0	0.064	2.02E-10
	A4	1000 (9000)	20	500	500	500	500	500	500	500	100	280 (200)	300	0	0	0.071	1.70E-12
	A5	1000 (9000)	20	500	500	500	500	500	500	500	405	1005 (945)	1215	0	0	0.161	2.20E-16
	A6	1000 (9000)	20	500	500	500	500	500	500	500	405	1065 (945)	1215	0	0	0.165	2.20E-16
	A7	1000 (9000)	20	500	500	500	500	500	500	500	405	1125 (945)	1215	0	0	0.175	2.20E-16
	A8	1000 (9000)	20	500	500	500	500	500	500	500	405	1185 (945)	1215	0	0	0.176	2.20E-16

Table S8. Ratio of windows that show perfect concordance with the species tree and the alternative tree Tr in simulations with differing levels of migration. We conducted 12 simulations where migration only occurs for a fraction of the genome. We defined four histories with an N_e of 500 for all populations and equal time between splits, but with the duration between splits varying between the two scenarios (T1 = 100, T2 = 200 T3 = 300; T1 = 405, T2 = 945, T3 = 1215). For each scenario, we modelled heterogenous migration by combining two simulations together. 90% of genealogies were simulated under a 'background' demography without gene flow, and the remaining 10% were simulated under the same model, but one of seven rates of migration from P3 to P2: *m* = 0, 0.001, 0.005, 0.010, 0.05, 0.01, or 0.05. For each simulation we determined the number of windows where T = 1 (*n* Tr = 1) and divided this by the number of windows that perfectly fit the background topology (Tb). In our simulation, we always observed a ratio of Tr:Tb much less than 1, with maximum of 0.37 at an extreme migration rate of 50% between P3 and P2. In our observed data, we observed a positive ratio of 1.41, indicating that windows with Tr = 1 are more abundant than Tb = 1. Thus, our observed result cannot be explained by neutral geneflow.

Split times	$m_{p3>p2}$	*n* Tr = 1	Tr:Tb
T100, 200, 300	0.000	0	0.00000000
T100, 200, 300	0.001	0	0.00000000
T100, 200, 300	0.005	0	0.00000000
T100, 200, 300	0.010	1	0.01538462
T100, 200, 300	0.050	9	0.13846154
T100, 200, 300	0.100	22	0.33846154
T100, 200, 300	0.500	24	0.36923077
T405, 945, 1215	0.000	0	0.00000000
T405, 945, 1215	0.001	54	0.01084337
T405, 945, 1215	0.005	359	0.07208835
T405, 945, 1215	0.010	580	0.11646586
T405, 945, 1215	0.050	715	0.14357430
T405, 945, 1215	0.100	695	0.13955823
T405, 945, 1215	0.500	713	0.14317269

Table S9. The 50 contigs with 1 or more windows where Tr = 1. Contig, the ID of the assembly contig; *n* wins where Tr =1, the number of windows where Tr = 1 on that contig. Start, start of the region where Tr = 1, End, end of the region where Tr = 1. Cumulative length, the sum of lengths of windows of the same contig (bp).

Contig ID	*n* wins where Tr = 1	Start	End	Cumulative length
Contig115331	1	7993	9268	1275
Contig52509	1	28533	30226	1693
Contig1261	1	126474	128742	2268
Contig3970	1	114	2486	2372
Contig116989	1	10306	12707	2401
Contig110028	1	504	3054	2550
Contig217335	1	1777	4453	2676
Contig39155	1	6308	9114	2806
Contig89329	1	6746	9600	2854
Contig7286	1	3835	6853	3018
Contig12403	1	20722	24153	3431
Contig38747	1	33289	36757	3468
Contig9868	1	626	4887	4261
Contig51247	1	433	4747	4314
Contig3201	1	56577	61149	4572
Contig70615	1	50027	54603	4576
Contig43562	1	38079	42770	4691
Contig1808	1	11041	15738	4697
Contig2118	1	102328	107302	4974
Contig67554	1	442	6257	5815
Contig181579	1	2208	8526	6318
Contig66365	1	57216	63741	6525
Contig63723	1	406	6958	6552
Contig86117	1	2927	9608	6681
Contig81547	1	7493	15204	7711
Contig12341	1	104	8623	8519
Contig3802	1	3800	13562	9762
Contig173651	1	983	12024	11041
Contig52055	1	6787	17879	11092
Contig128971	1	4271	18449	14178
Contig47534	1	10949	25863	14914
Contig79517	1	160	15210	15050
Contig40336	1	46155	65216	19061
Contig108528	1	615	29659	29044
Contig48348	2	35646	39610	3964
Contig76820	2	9113	15705	6592
Contig61890	2	9294	16927	7633
Contig41237	2	137578	153114	15536
Contig3968	2	8478	20380	11902
Contig2032	2	42723	56923	14200
Contig61156	2	74402	90066	15664
Contig76495	3	139	17244	17105
Contig47979	3	1979	16551	14572
Contig60406	3	1641	34175	32534
Contig181768	3	96	20699	20603
Contig54667	3	6589	51824	45235
Contig2245	4	3318	61149	57831
Contig59771	6	2623	42145	39522
Contig83530	7	5330	69596	64266
Contig81225	8	2204	79268	77064

Appendix C

Supplementary Information for

Haplotagging pipeline for 1,084 whole-genome sequences in an Antirrhinum hybrid zone

C.1 | Supplementary Tables

Table S1. Design of greenhouse diallel crosses and wild parent-offspring trios from pedigree study in the Planoles hybrid zone. The IDs of the offspring and parents match those in Supplementary Data 1.

Offspring ID	Parent 1 ID	Parent 2 ID	Trio type
D01	406	603	Greenhouse diallel cross
D02	406	608	
D03	603	406	
D04	608	603	
D05	608	406	
D06	603	608	
D07	–	–	Greenhouse diallel cross; parents not sequenced
D08	–	–	
D09	–	–	
D10	–	–	
T0281	P0192	P0247	Wild trios from pedigree study in the Planoles hybrid zone
V0574	P0239	P0578	
T0082	P0278	P0293	
T1352	S5237	S1803	

Table S2. Summary of total number of reads, % mapped and % correct barcodes for each sequencing lane of samples. Each row constitutes a different lane of Illumina Novaseq S4. Batches *trios* and *n-96* consist of 28 and 96 unique samples respectively. Batch n-960 was sequenced twice with the same 960 samples to varied sequencing depths since the Bx-tag (i.e., the molecular barcode) was not sequenced properly in one of the runs. Batch *10x-NEW* consists of 160 samples from the *n-960* batch that were re-sequenced to obtain higher s coverage. For each batch, the total number of paired-end reads, percentage of reads mapped to the *A. majus* genome, and the percentage of reads with correctly sequenced *haplotag* barcode are also provided.

Lane	Batch Name	Sample size	Total Reads	% correct barcodes	% mapped	Full/Part lane
1	trios	28	488,584,359	95.36	96.32	Part lane
2	n-96	96	331,871,048	98.14	95.32	Part lane
3	n-960a	960	5,224,876,464	97.51	97.68	Full lane
4	n-960b *(BX-tag incorrectly sequenced)*	960	5,409,264,651	16.21	95.88	Full lane
5	10x-NEW	160	4,894,636,700	97.34	97.75	Full lane

Appendix D

Supplementary Information for

Genealogical analysis of replicate flower colour hybrid zones in *Antirrhinum*

D.1 | Supplementary Methods

D.1.1 | DNA extraction

DNA was extracted from a 1x0.5 cm^2 piece of dry leaf using a custom protocol optimised for isolating high molecular weight DNA. *(i)* Tissue was first crushed to a fine powder in a 1.5 ml Eppendorf tube using a micropestle. *(ii)* 0.5 ml of lysis buffer was added, containing 400 µl PureLink Genomic Digestion Buffer (Thermofisher: K182301), 40 µl Proteinase K (20mg/ml), 60 µl of 18% Polyvinylpyrrolidone (40 kdaG), and 2% Beta-mercaptoethanol. *(iii)* Samples were incubated for 15-20 mins at 60ºC with mixing at 950 rpm with occasional inverting of the tube, followed by 30 secs on ice. *(iv)* 10 µl of 5 mg/ml RNaseA was added and mixed by inverting the tube several times, and incubated for 5 mins at room temperature. *(v)* 155 µl 5M of potassium acetate was added and immediately mixed by inverting 10-15 times, followed by incubation for 2 mins on ice. *(vi)* The samples were centrifuged at 13k g for 10 mins at 14ºC and 400 µl of lysate was transferred to a new tube using a wide-orifice tip. *(vii)* 200 µl magnetic beads were added and immediately mixed by inverting 10-15 times and incubated for 10 mins at room temperature. *(viii)* Samples were pulse-spun for 2 seconds, and put on a magnet stand for 5 mins and the lysate was discarded. *(ix)* The beads were washed twice for 1 min with 80% EtOH; All EtOH was carefully removed after the second wash and the sample left open to evaporate for 2-4 mins at room temperature. *(x)* Samples were removed from the magnet and 100 µl of 10 mM Tris pH=8, 0.2mM EDTA was added to the tube and incubated for 10 mins at 42ºC to elute the DNA. *(xi)* Samples were mixed gently by inverting to re-suspend the beads and allowed to incubate for 10 mins at 42ºC. Eluted DNA was stored with beads at 4ºC. The concentration of each DNA sample was determined using a Qubit fluorometer or an M200 Pro Tecan plate reader. A rough estimate of size and quality was determined using agarose gel electrophoresis. Samples were diluted to 5 ng/µl using Tris, pH=8, 0.2mM EDTA and stored at 4ºC for future use.

D.1.2 | Polarising allele in A. majus as ancestral or derived

We polarised alleles in *A. majus* as ancestral or derived using high-coverage PoolSeq sequence data (mean coverage = 89.97x) from multiple populations of a closely related outgroup species *A. molle*. For each bi-allellic site in *A. majus*, we determined which alleles were present within *A. molle*. For 83.64% of sites, we observed 3 site-patterns that allowed us to identify the derived allele in *A. majus* (Table S4)

Pattern A: *6,624,396 sites (57.4%) resolved.*

Sites where one of the *A. majus* alleles is fixed in *A. molle*. Out of these 6.6 million sites that constituted 57.44% of all the bi-allelic *A. majus* sites, 5.5 million (47.83%) sites had the *A. majus* major allele fixed in *A. molle*, whereas the remaining 1.1 million (9.61%) sites had the *A. majus* minor allele fixed in *A. molle*. For all such sites, the fixed *A. molle* allele was assigned ancestral, while the other *A. majus* allele derived.

Pattern B: *2,537,001 sites (21.99%) resolved.*

Sites where both *A. majus* alleles are present in *A. molle*, and both species share the same major allele, i.e., allele frequency ≥ 0.5. For such sites, the major allele was assigned ancestral, since the minor allele in either taxa would most likely be derived due to new mutations being relatively rare.

Pattern C: *484,585 (4.21%) resolved.*

Sites where both species are polymorphic, but only share 1 allele. In this case, the shared allele is considered ancestral, while the unique allele is considered derived.

The remaining 16.36% of bi-allelic sites in *A. majus* had no logical basis for assigning ancestral or derived alleles based on the information from *A. molle* (Patterns D–F in Table S4). The most unresolvable site-pattern (Pattern D: 1,418,581 or 12.30% of all bi-allelic sites), involved both species sharing the same 2 alleles, but the minor allele in one species was the major allele in the other. For 437,801 sites (3.80%**)**, there was simply no sequence information available for *A. molle* (Pattern E in Table S4). Finally for the remaining 30,666 sites (0.26%), neither of the *A. majus* alleles was present in *A. molle* (Pattern F in Table S4), suggesting both alleles being derived in an unknown order in the ancestor to *A. majus*. For all the 1,887,048 unresolved bi-allelic sites (16.36%), we assumed the minor allele (< 0.5 allele frequency) in *A. majus* to be the most likely derived allele.

D.2 | Supplementary Figures

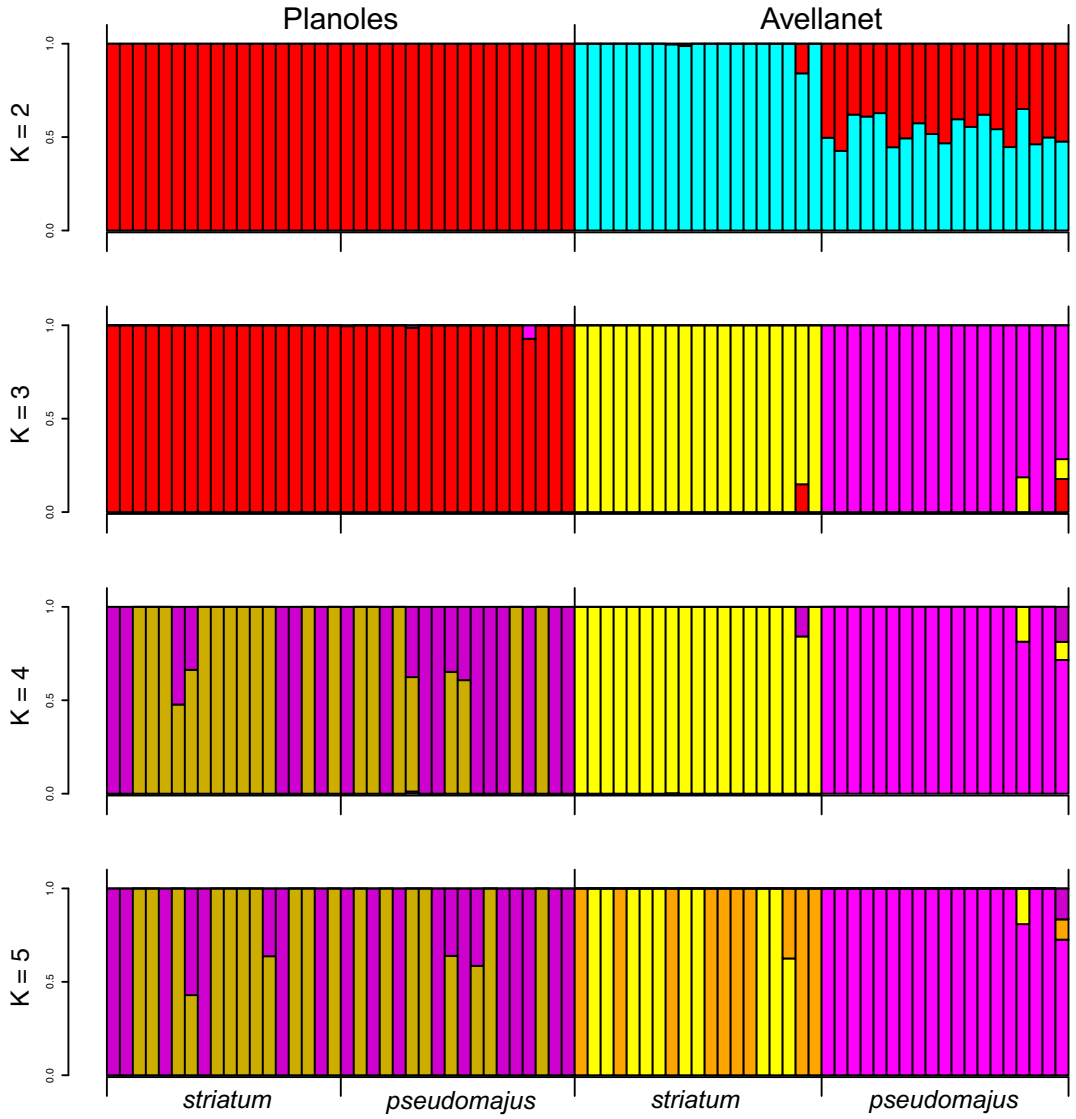

Figure S1. Admixture (Alexander et al. 2009) analysis of all 74 samples on 1,710,010 independent SNPs with *K* ranging from 2 to 5. *K*=3 describes the clearest ancestral grouping pattern, separating Avellanet into the 2 varieties, but grouping all Planoles samples together. None of the *K* values uniquely separate the var. *pseudomajus* and var. *striatum* samples at Planoles.

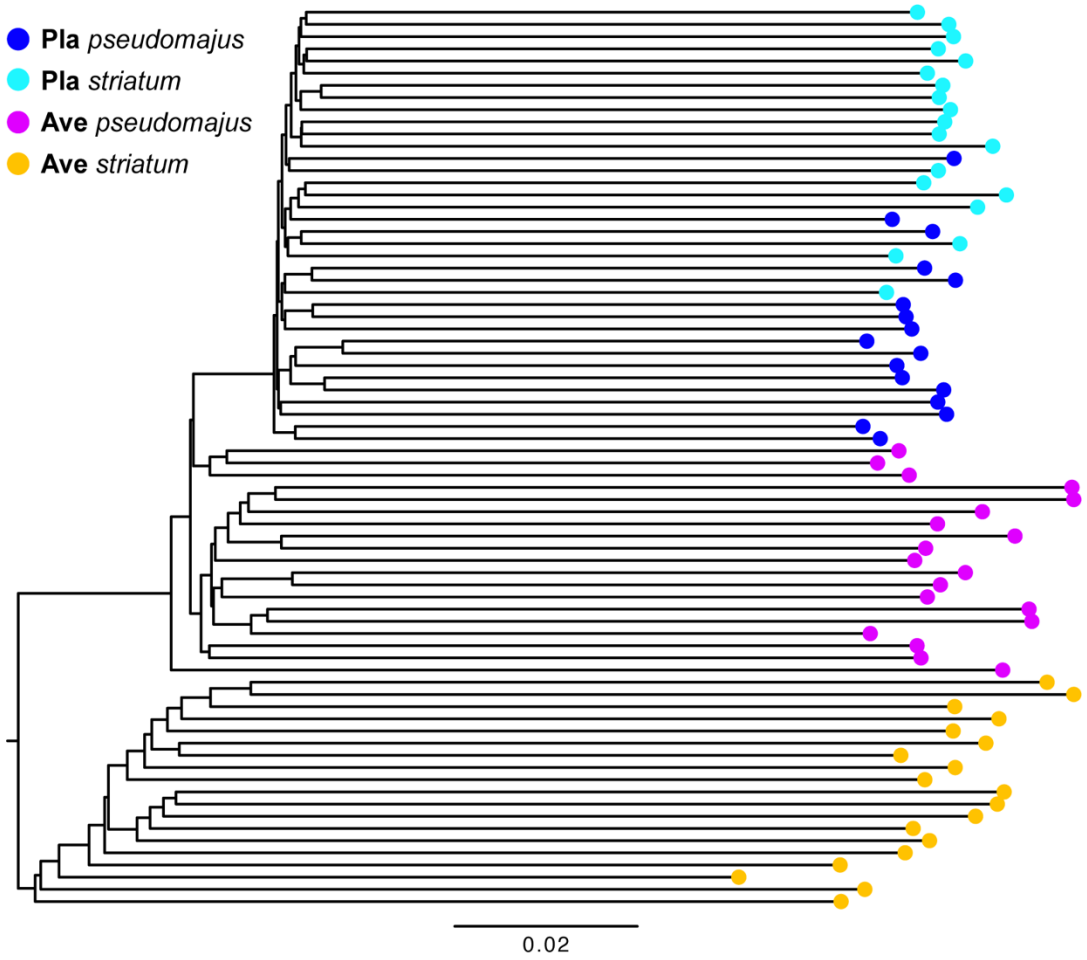

Figure S2. Neighbour-Joining tree inferred from 1,710,010 independent SNPs, which is used to calculate topology weighting for the whole genome.

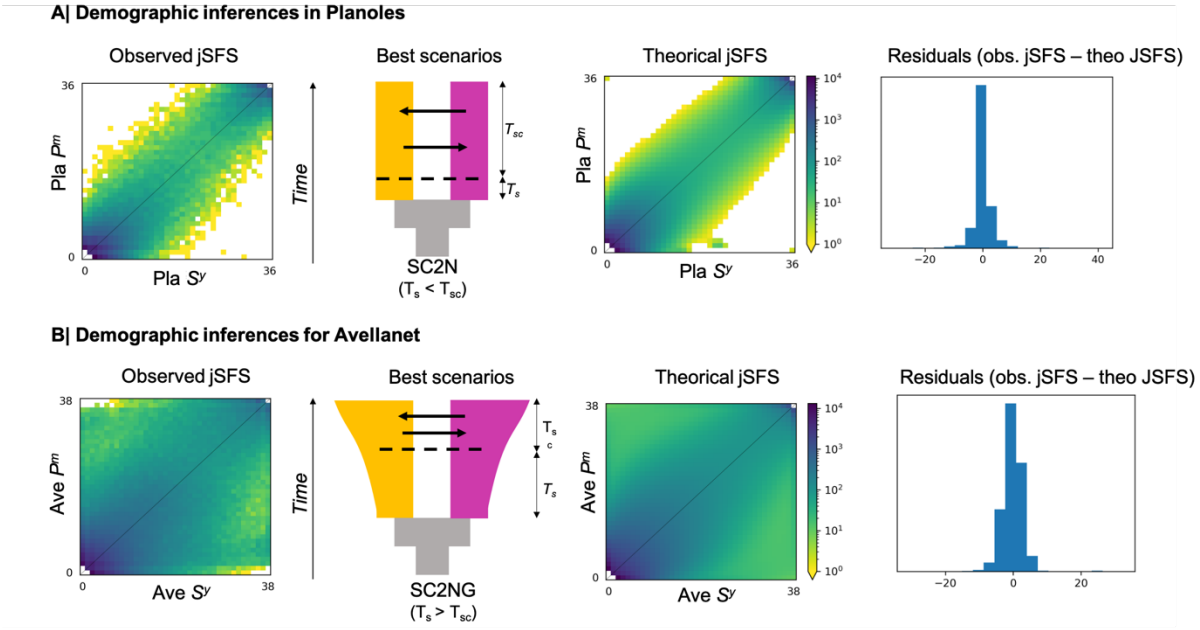

Figure S3. Results of the best demographic scenario inferred from *δaδi* analyses in **(A)** Planoles and **(B)** Avellanet. From left to right: the observed joint Site Frequency Spectrum ($jSFS_{obs}$), a schematic representation of the best-fitted demographic scenario, the theoretical spectrum ($jSFS_{theo}$) inferred from the best demographic scenario, and the distribution of residuals ($jSFS_{obs}$ – $jSFS_{theo}$). For Planoles, the best-fitted scenario is a Secondary Contact with heterogeneous effective population size across the genome (SC2N). For Avellanet, the best-fitted scenario is a Secondary Contact with heterogeneous effective population size across the genome and exponential growth of the lineages (SC2NG).

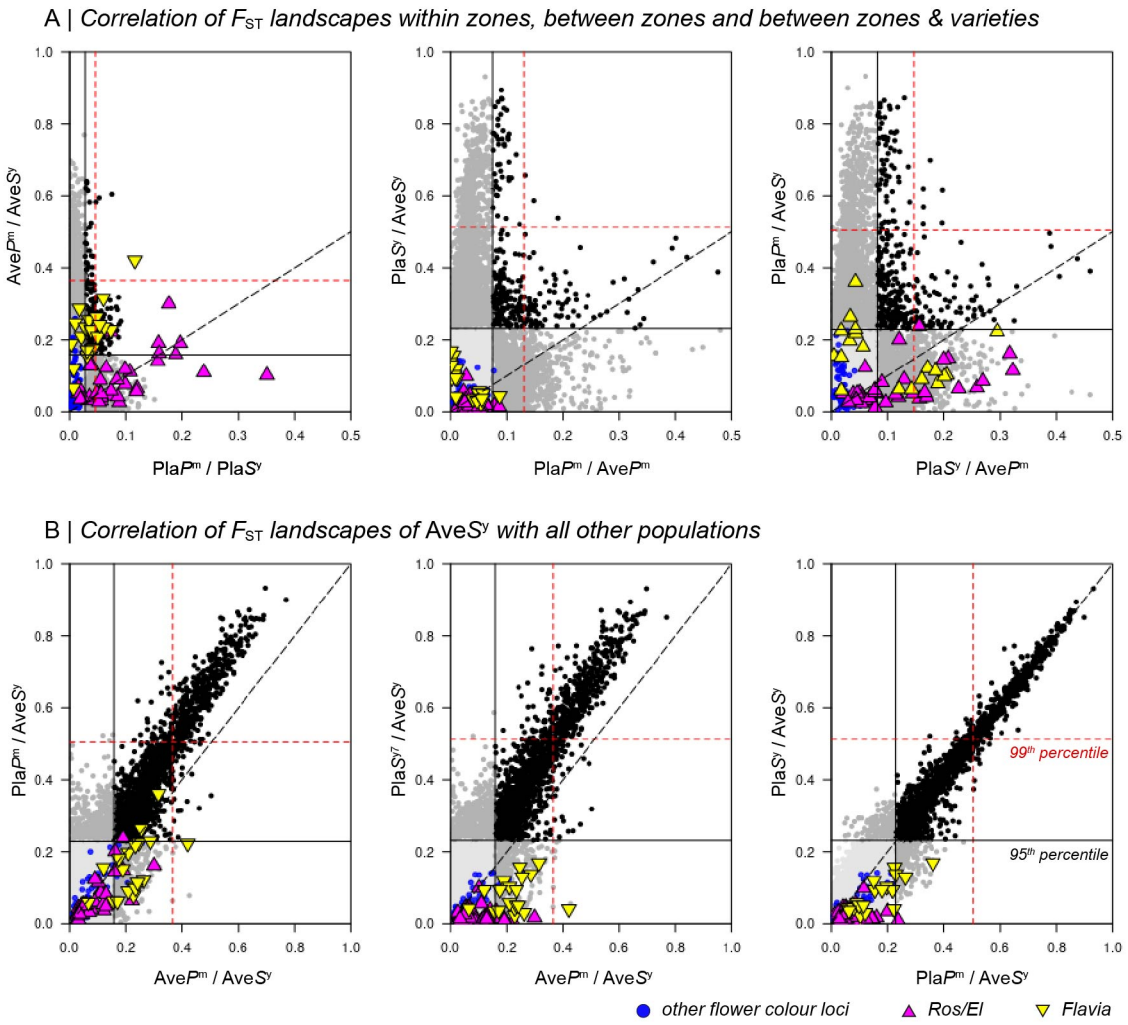

Figure S4. Correlations of pairwise Weir & Cockerham F_{ST} in 10 Kbp non-overlapping windows. **(A)** The landscapes chosen for pairwise comparisons follow the same categories from top to bottom in Fig. 2 (Main Text) – i.e., within zones (PlaP^m/PlaS^y vs. AveP^m/AveS^y), between zones (PlaP^m/AveP^m vs. PlaS^y/AveS^y) and between zones & varieties (PlaS^y/AveP^m vs. PlaP^m/AveS^y). **(B)** Correlations of F_{ST} landscapes between AveS^y and all the other populations (AveP^m/AveS^y, PlaP^m/AveS^y, PlaS^y/AveS^y). 95th and 99th percentile thresholds are drawn in black and red within each plot, and black dots represent windows that are outliers (using 99th percentile) in both F_{ST} landscapes. Magenta, yellow and blue dots represent windows that carry *Flavia*, *Ros/El* and other loci associated with flower colour. Pla: Planoles, Ave: Avellanet, P^m: magenta-coloured var. *pseudomajus*, S^y: yellow-coloured var. *striatum*.

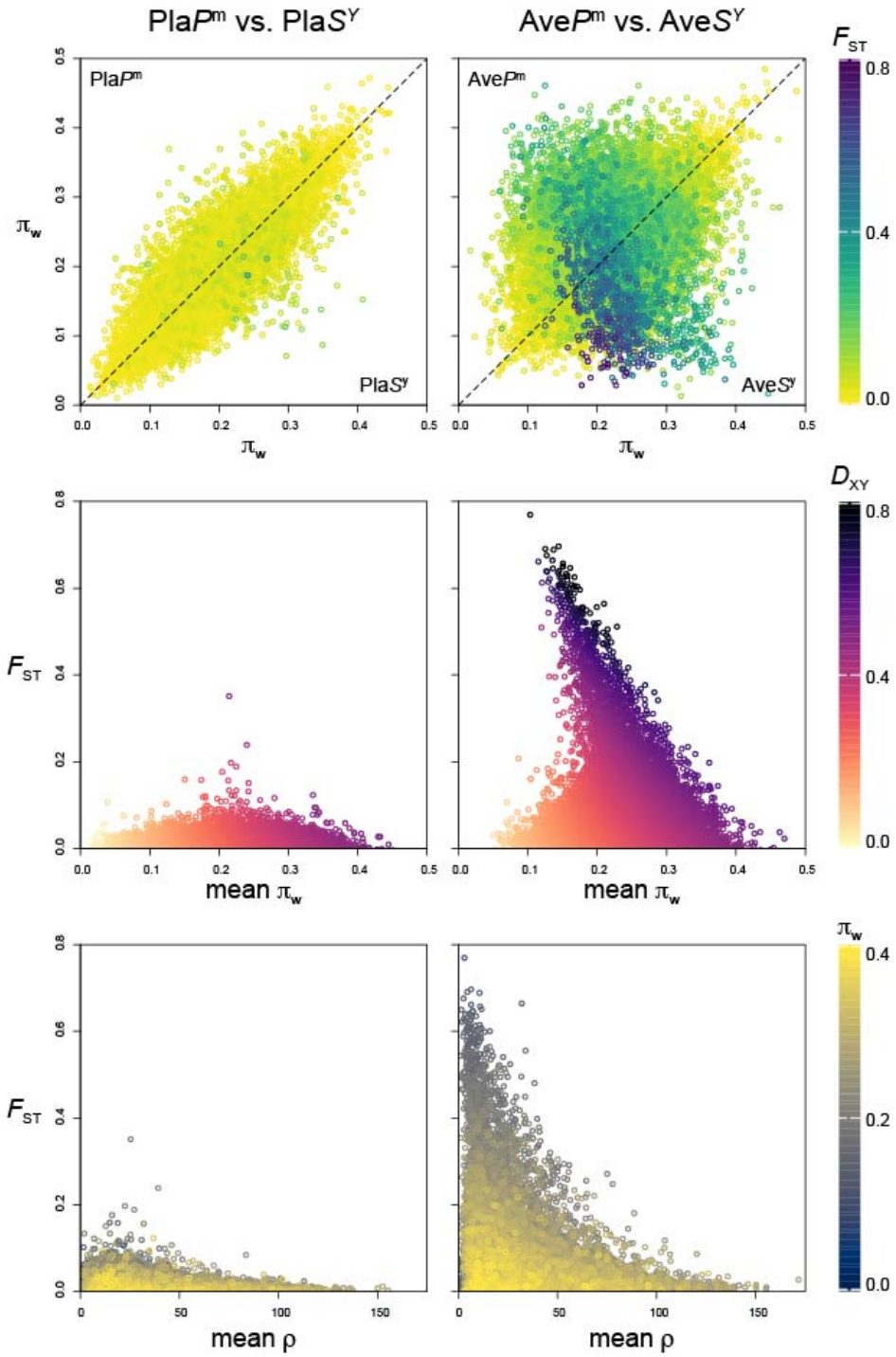

Figure S5: Relationships between population genetic estimates in 10 Kbp non-overlapping windows between both varieties at each hybrid zone. Top panel: π_w between the var. *pseudomajus* and var. *striatum*; coloured by F_{ST}. Middle panel: F_{ST} vs. mean π_w ; coloured by D_{XY}. Bottom panel: F_{ST} vs. mean ρ (population recombination rate); coloured by mean π_w. Pla: Planoles, Ave: Avellanet, P^m: magenta-coloured var. *pseudomajus*, S^y: yellow-coloured var. *striatum*.

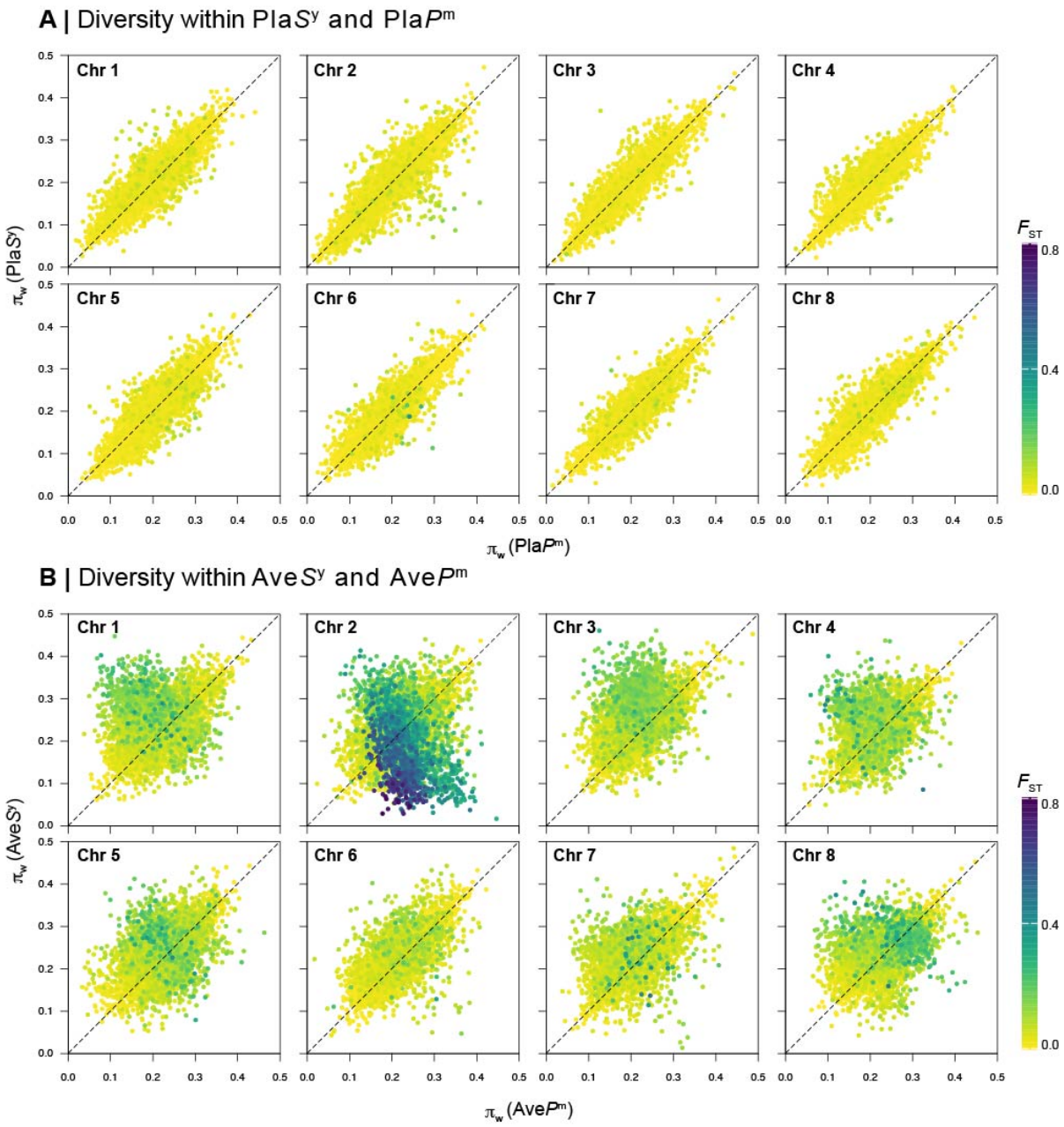

Figure S6. Relationship of diversity (π_w) within var. *pseudomajus* and var. *striatum* in 10 Kbp non-overlapping windows at **(A)** Planoles and **(B)** Avellanet, shown for each chromosome. Dots are coloured by F_{ST}. Pla: Planoles, Ave: Avellanet, P^m: magenta-coloured var. pseudomajus, S^y: yellow-coloured var. *striatum*.

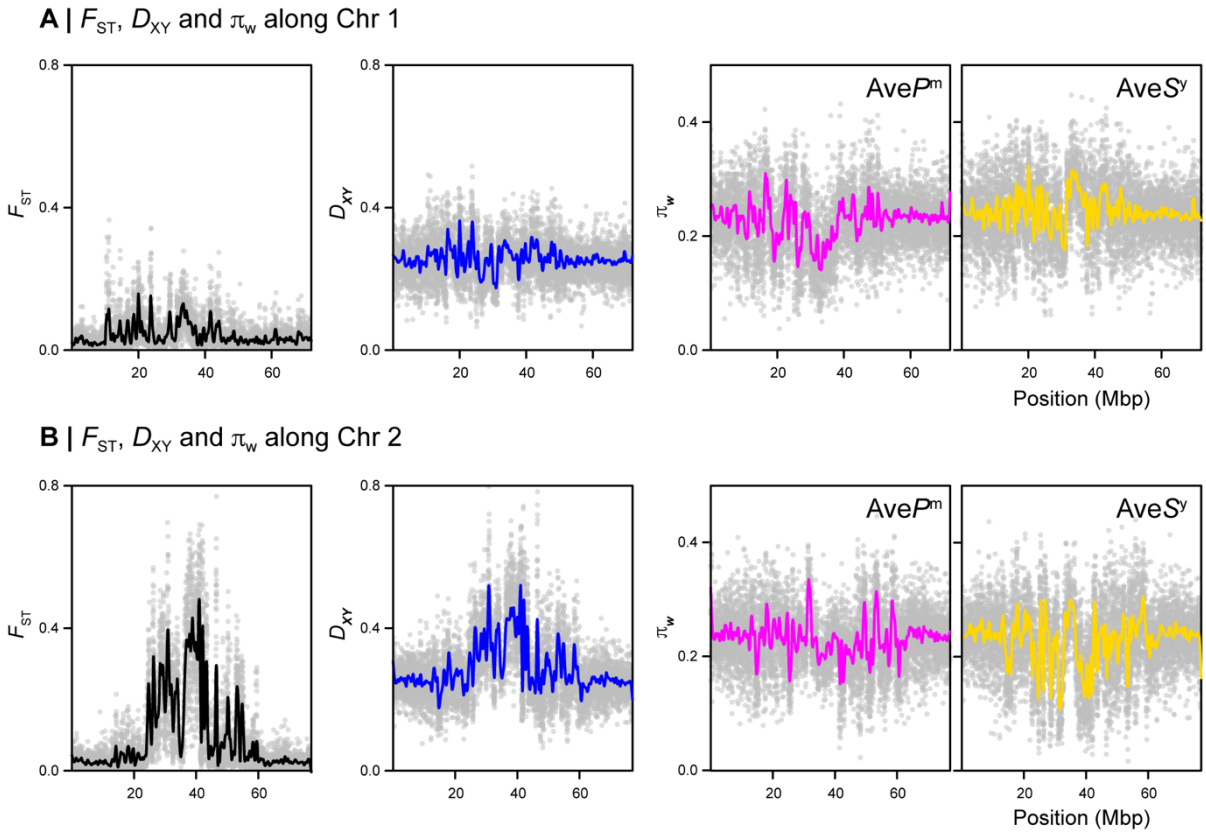

Figure S7. Population genetic parameters F_{ST}, D_{XY} and π_w along **(A)** Chr 1 and **(B)** Chr 2 for both varieties at Avellanet. Grey dots show estimates in 10 Kbp non-overlapping windows, while solid coloured lines show loess smoothed (span = 50 Kbp) estimates. Ave: Avellanet, P^m: magenta-coloured var. *pseudomajus*, S^y: yellow-coloured var. *striatum*.

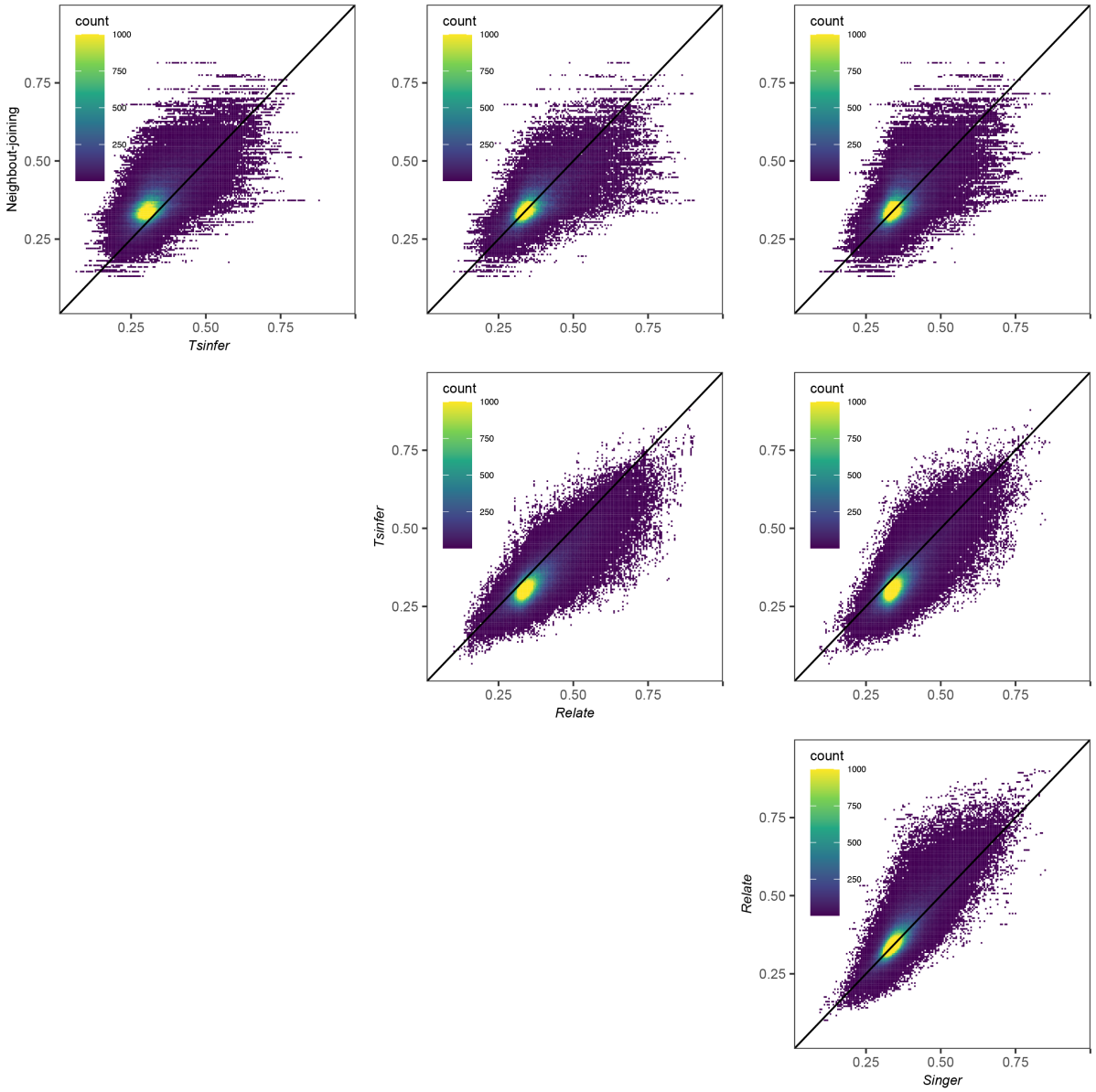

Figure S8. Correlation of T_{geo} (clustered by geography) topology weights of trees inferred by different tree inference methods. Each square tile represents the number of trees in it.

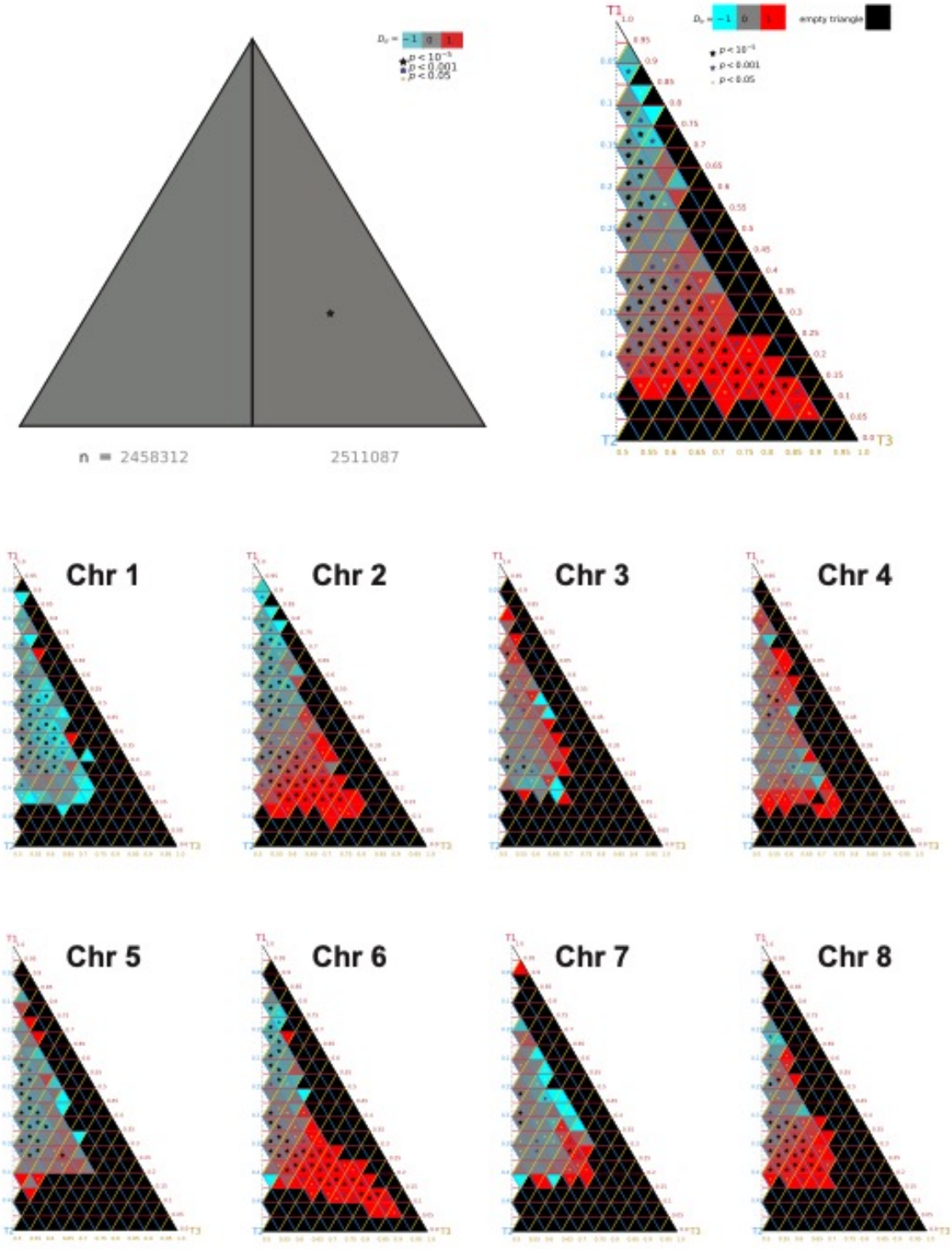

Figure S9. Left-right asymmetry from *TwisstNTern* analysis. Counts of trees in the left and right half-triangles, with asymmetry quantified using D_{LR}. Asterisks indicate significant asymmetry between corresponding left- and right-sided sub-triangles, suggesting excess of T_{var} or T_{alt} topologies.

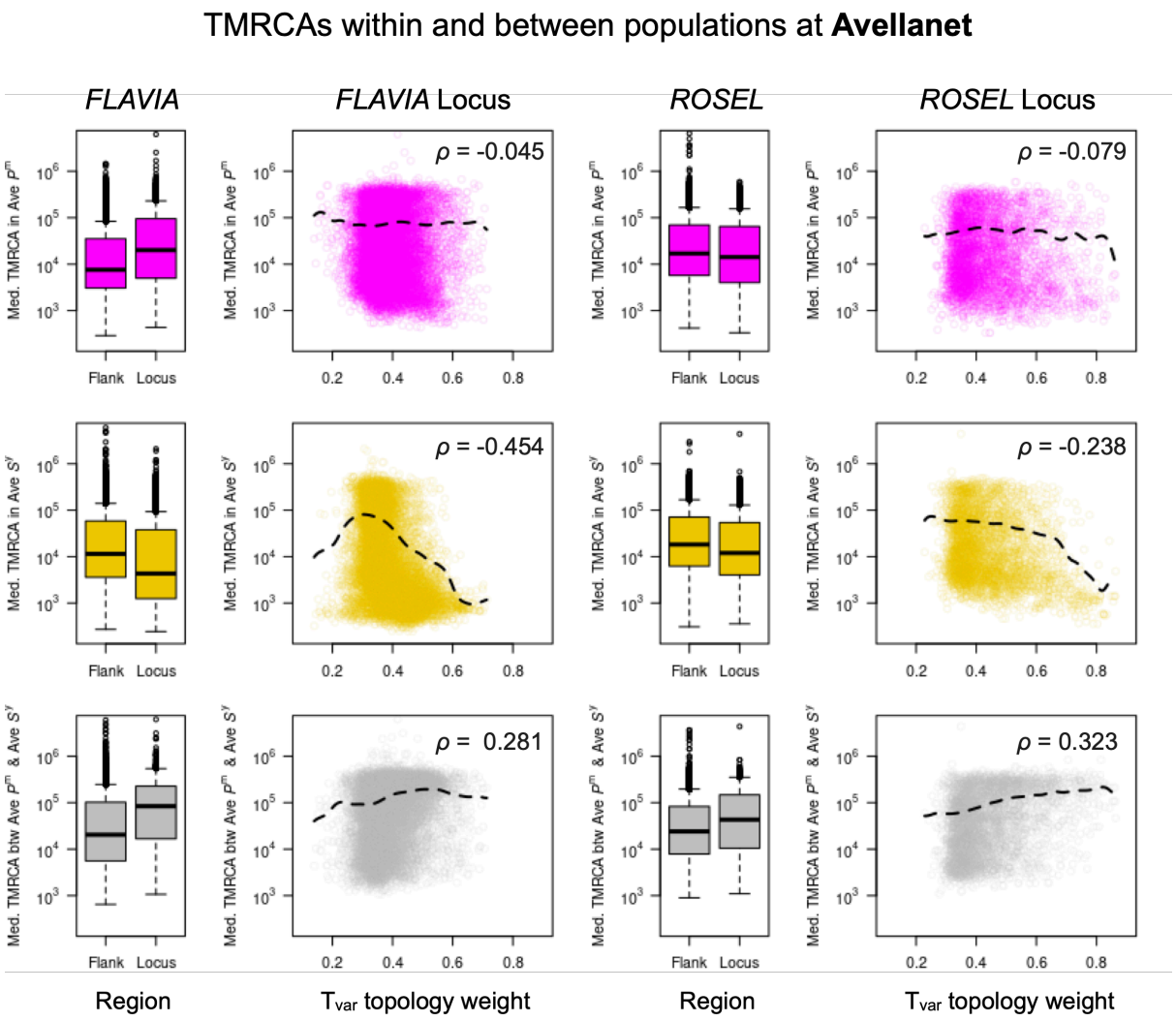

Figure S10. Coalescence times at *FLAVIA* and *ROS/EL* at the Avellanet populations. Boxplots for *FLAVIA* and *ROS/EL* show the median time to most recent common ancestor (TMRCA) within var. *pseudomajus* (top row), within var. *striatum* (middle row), and between var. *pseudomajus* and var. *striatum* (bottom row), all on a log scale. The left boxes ('Flank') show the TMRCAs in the flanking regions around the locus, while the right boxes ('Locus') show the values inside the locus. The scatterplots and dashed black lines show the full distribution and the overall smoothed trend between the T_{var} weight and median TMRCA within the locus, (row are as indicated for the boxplots). *ρ* is the correlation coefficient from a Spearman's rank correlation.

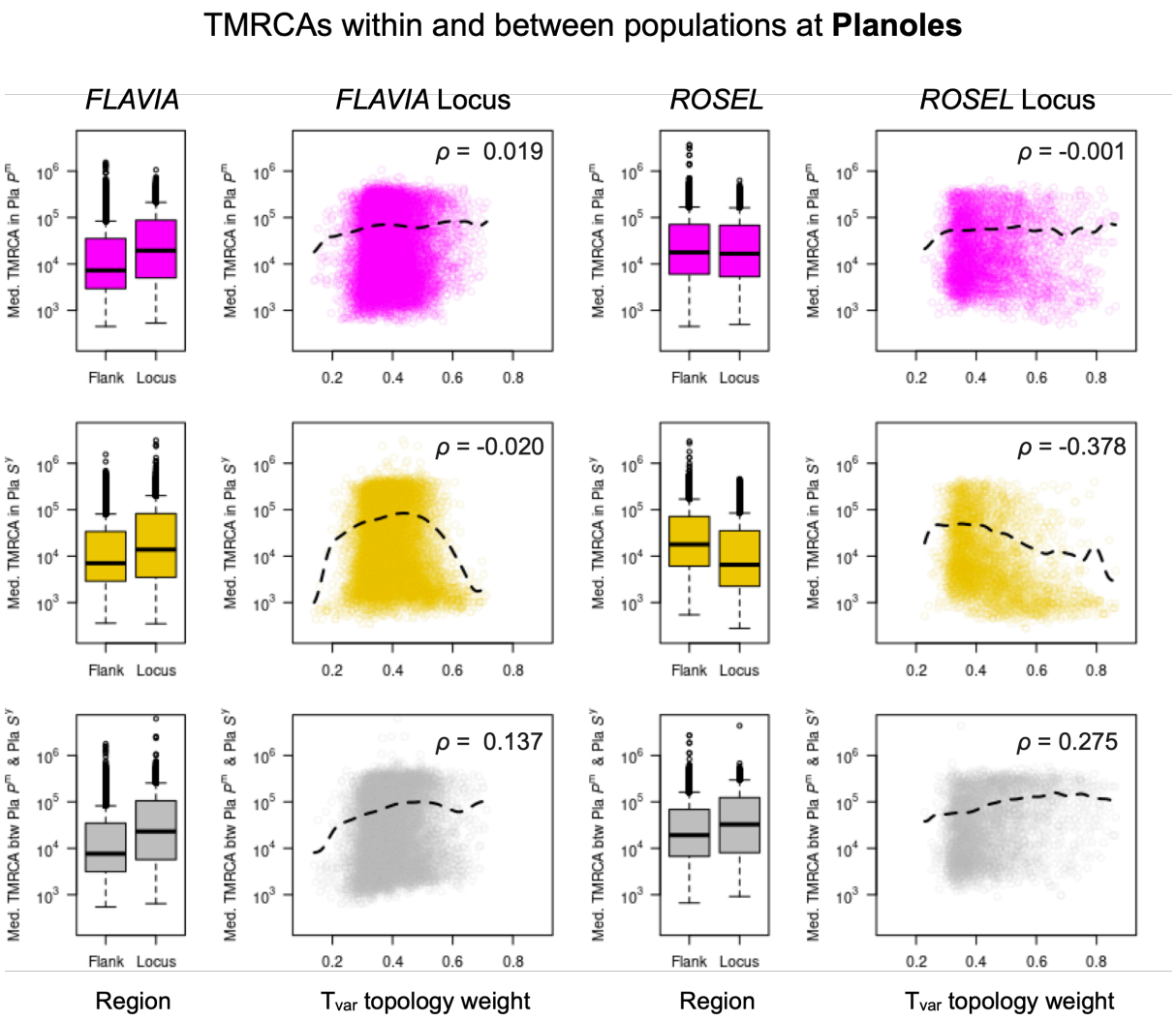

Figure S11. Coalescence times at *FLAVIA* and *ROS/EL* at the Planoles populations. Boxplots for*FLAVIA* and *ROS/EL* show the median time to most recent common ancestor (TMRCA) within var. *pseudomajus* (top row), within var. *striatum* (middle row), and between var. *pseudomajus* and var. *striatum* (bottom row), all on a log scale. The left boxes ('Flank') show the TMRCAs in the flanking regions around the locus, while the right boxes ('Locus') show the values inside the locus. The scatterplots and dashed black lines show the full distribution and the overall smoothed trend between the T_{var} weight and median TMRCA within the locus, (row are as indicated for the boxplots). *ρ* is the correlation coefficient from a Spearman's rank correlation.

D.3 | Supplementary Tables

Table S1. Locations and sequencing parameters for all samples. Pla: Planoles, Ave: Avellanet, *P* : *pseudomajus* , *S* : *striatum*.

Plant ID	Population		Latitude	Longitude	Altitude	Total reads	# Reads with correct barcodes	# Total reads mapped	Coverage		Mapping Quality
	HZ	var.							mean	sd	
x3318	Ave	*P*	42.3265041	1.34596441	1180.397	31,88,122	31,29,590	31,05,569	0.79	5.3	43.86
x3327	Ave	*P*	42.3253905	1.34795383	1176.437	23,71,661	23,29,207	22,97,793	0.58	3.97	43.79
x3333	Ave	*P*	42.3254111	1.34793968	1175.417	39,81,670	39,22,672	38,70,878	0.98	5.78	43.51
x3394	Ave	*P*	42.3265476	1.34605381	1182.238	52,86,233	51,98,347	51,08,647	1.3	7.35	43.64
x4101	Ave	*P*	42.3254284	1.3479035	1173.368	49,97,162	48,98,390	48,35,944	1.23	6.37	43.47
x4102	Ave	*P*	42.3253405	1.34797537	1174.698	40,02,073	39,35,551	38,92,352	0.99	7.08	43.11
x4116	Ave	*P*	42.3253172	1.34792778	1173.042	38,73,173	37,80,189	37,55,484	0.96	5.67	43.98
x4128	Ave	*S*	42.352512	1.32733675	1306.395	37,42,683	36,86,947	36,20,227	0.92	4.62	43.8
x4129	Ave	*S*	42.3516395	1.32255535	1294.688	43,07,949	42,44,685	42,03,622	1.07	6.3	43.27
x4134	Ave	*S*	42.3562788	1.31035377	1305.424	59,14,001	58,13,443	57,52,068	1.46	8.94	43.06
x4137	Ave	*S*	42.3562711	1.31059675	1309.896	48,63,149	47,45,309	46,99,409	1.2	6.04	43.67
x4140	Ave	*S*	42.3518215	1.32795229	1306.514	38,91,753	38,18,083	37,36,714	0.95	5.81	43.23
x4148	Ave	*S*	42.3498082	1.32983188	1298.141	47,21,946	46,35,050	45,90,030	1.17	5.58	43.69
x4149	Ave	*S*	42.3529544	1.31888622	1297.671	54,35,863	53,13,479	52,64,802	1.34	5.99	43.1
x4150	Ave	*S*	42.3563588	1.31021196	1310.725	37,02,921	36,36,255	36,13,515	0.92	6.29	44.29
x4153	Ave	*S*	42.3563674	1.31020257	1309.695	38,92,121	38,31,083	37,89,206	0.96	6.91	44.06
x4155	Ave	*S*	42.3563344	1.31034666	1310.153	35,19,212	34,29,232	34,09,840	0.87	4.86	43.96
x4158	Ave	*S*	42.3517021	1.32282928	1286.822	53,27,782	52,36,008	51,69,177	1.32	8.49	43.87
x4159	Ave	*S*	42.3562196	1.3108351	1368.269	44,28,404	43,58,874	42,27,330	1.07	8.93	43.72
x4161	Ave	*P*	42.3253486	1.34791663	1173.946	42,38,342	41,67,294	40,72,708	1.04	6.11	43.38
x4178	Ave	*P*	42.3253335	1.34789948	1172.338	37,75,442	37,06,434	36,57,954	0.93	6.72	44.25
x4359	Ave	*P*	42.3257335	1.3478534	1178.421	35,23,972	34,72,840	34,28,504	0.87	5.57	43.46
x4392	Ave	*S*	42.3527129	1.31938189	1292.184	38,88,237	38,15,603	37,80,500	0.96	5.38	42.93
x4429	Ave	*P*	42.3262125	1.34754946	1182.735	37,54,580	36,98,112	36,52,600	0.93	4.5	43.14
x4459	Ave	*P*	42.3260837	1.34722933	1171.97	50,96,342	50,17,740	49,68,912	1.27	7.98	44.11
x4463	Ave	*S*	42.3563577	1.31069132	1308.321	47,29,151	46,54,861	46,23,876	1.18	8.56	44.2
x4477	Ave	*S*	42.3525811	1.32008953	1290.641	41,87,543	41,09,349	40,77,855	1.04	8.96	43.86
x4491	Ave	*S*	42.3570279	1.30991165	1334.262	55,44,770	54,57,620	52,74,002	1.34	12.48	43.93

Table S1. *continued*

Plant ID	Population		Latitude	Longitude	Altitude	Total reads	# Reads with correct barcodes	# Total reads mapped	Coverage		Mapping Quality
	HZ	var.							mean	sd	
x4570	Ave	*S*	42.3515774	1.32177665	1294.526	34,63,415	34,09,673	33,71,245	0.86	5.21	42.92
x4579	Ave	*S*	42.3561238	1.31298015	1302.808	29,32,332	28,76,876	27,90,603	0.71	4.41	42.79
x4580	Ave	*S*	42.3561336	1.3130905	1303.809	27,94,530	27,52,510	27,10,713	0.69	4.72	43.09
x4585	Ave	*P*	42.3255331	1.34790478	1178.988	41,37,568	40,73,764	39,86,251	1.02	5.2	42.94
x4593	Ave	*P*	42.3255351	1.34790367	1179.321	36,24,245	35,64,875	33,71,160	0.84	8.32	43.91
x4608	Ave	*P*	42.3265208	1.34601267	1181.009	34,47,603	33,53,019	33,25,831	0.84	6.45	43.72
x4624	Ave	*P*	42.326054	1.34714137	1171.926	52,76,887	51,70,031	51,05,650	1.29	10.77	43.93
x4786	Ave	*P*	42.3263883	1.34558351	1183.084	31,24,533	30,72,873	30,25,727	0.77	5.37	44.02
x4837	Ave	*P*	42.3250915	1.34802625	1176.246	30,06,238	29,32,758	28,57,202	0.72	5.89	43.77
x4867	Ave	*P*	42.3263835	1.34565351	1181.582	20,81,727	20,47,913	20,09,962	0.51	2.38	42.61
z0169	Pla	*S*	42.32448609	2.06488582	1225.909	33,31,472	32,57,556	31,36,601	0.8	6.43	44.77
z0347	Pla	*S*	42.32502574	2.062415859	1230.095	26,00,606	25,62,974	24,77,371	0.63	4.24	44.03
z0704	Pla	*P*	42.3236173	2.08200593	1184.011	22,65,609	22,26,435	21,98,175	0.56	3.02	43.52
z0713	Pla	*P*	42.32365681	2.081934964	1182.894	34,02,360	33,36,764	33,02,472	0.84	5.82	43.32
z0847	Pla	*S*	42.32463343	2.06416347	1229.266	26,62,730	26,20,318	25,86,796	0.66	5.02	45
z0852	Pla	*S*	42.32446099	2.065055654	1222.145	35,22,296	34,60,620	33,50,803	0.85	7.37	44.35
z0854	Pla	*S*	42.3246058	2.06428122	1229.051	33,98,868	33,43,220	32,27,563	0.82	6.5	44.71
z1778	Pla	*P*	42.32322484	2.082552166	1177.311	40,23,628	38,90,344	38,53,827	0.98	8.99	44.83
z1802	Pla	*S*	42.32451883	2.064910802	1224.598	38,04,550	37,45,410	36,73,814	0.94	7.19	45.03
z2008	Pla	*S*	42.3250673	2.062047337	1238.035	33,00,071	32,44,531	31,48,951	0.8	6.35	44.8
z2018	Pla	*S*	42.32451109	2.064902788	1225.374	45,36,347	44,59,591	43,28,480	1.1	7.73	44.36
z2020	Pla	*S*	42.32503701	2.06272073	1237.244	36,93,881	36,26,029	34,42,237	0.87	5.91	43.75
z2046	Pla	*P*	42.32322195	2.082707483	1195.883	28,56,165	27,86,839	27,38,491	0.7	4.79	45.05
z2281	Pla	*P*	42.32368318	2.081933918	1159.164	22,78,684	22,35,650	21,67,369	0.55	2.94	44.38
z2283	Pla	*P*	42.3235961	2.081999912	1182.457	27,53,499	27,14,511	26,39,088	0.67	3.12	43.28
z2286	Pla	*P*	42.323607	2.081954368	1185.514	23,69,109	23,04,211	22,08,590	0.56	2.82	43.27
z2287	Pla	*P*	42.32360853	2.081924361	1180.977	23,58,238	23,04,338	22,02,698	0.56	2.84	43.82

Table S1. *continued*

Plant ID	Population		Latitude	Longitude	Altitude	Total reads	# Reads with correct barcodes	# Total reads mapped	Coverage		Mapping Quality
	HZ	var.							mean	sd	
z2290	Pla	*P*	42.32361059	2.081978279	1184.116	31,35,858	30,71,886	29,85,428	0.76	4.75	44.93
z2292	Pla	*P*	42.32359454	2.081963851	1184.393	32,48,824	31,82,852	30,54,138	0.77	4.88	44.92
z2293	Pla	*P*	42.32360997	2.081978164	1182.295	21,29,244	20,92,966	20,43,766	0.52	3.51	44.68
z2294	Pla	*P*	42.32359948	2.081961153	1183.949	32,45,003	31,77,589	30,85,916	0.78	4.67	44.18
z2295	Pla	*P*	42.32360819	2.081973936	1183.365	33,43,971	32,93,189	32,07,905	0.81	4.69	44.42
z2296	Pla	*P*	42.32360588	2.081925299	1181.766	23,50,694	22,97,670	22,02,817	0.56	5.46	44.12
z2298	Pla	*P*	42.32361395	2.081979136	1181.428	39,02,911	38,40,117	37,67,598	0.96	8.04	44.67
z2300	Pla	*P*	42.32360116	2.081929359	1183.732	31,75,981	31,05,901	29,97,729	0.76	3.42	42.94
z2706	Pla	*S*	42.32458183	2.064503601	1233.894	32,40,057	31,85,969	31,20,068	0.8	3.25	44.17
z2711	Pla	*S*	42.32504359	2.061967047	1235.348	50,45,346	49,64,032	48,67,405	1.24	8.07	44.45
z3891	Pla	*S*	42.32512037	2.062697473	1467.332	38,59,686	37,98,512	37,48,425	0.95	5.75	44.61
z4710	Pla	*P*	42.32364551	2.081975459	1177.4	23,61,709	23,22,695	22,03,640	0.56	3.64	44.18
z4712	Pla	*P*	42.32364214	2.081965884	1176.806	28,02,140	27,49,078	26,09,075	0.66	4	44.29
z4758	Pla	*S*	42.3248285	2.063343366	1300.989	53,01,316	51,89,766	51,32,422	1.31	10.83	44.73
z4783	Pla	*S*	42.32459417	2.064393839	1228.686	32,20,924	31,65,208	31,23,531	0.8	5.27	44.34
z4793	Pla	*S*	42.32491715	2.063297364	1222.045	22,57,174	22,20,722	21,78,325	0.55	4.28	44.72
z4797	Pla	*S*	42.32166099	2.070664005	1203.821	25,68,966	24,87,622	24,47,585	0.62	4.29	44.8
z5444	Pla	*S*	42.32488623	2.063280532	1220.282	32,09,147	31,52,037	31,28,019	0.8	5.27	44.79
z5451	Pla	*S*	42.3245906	2.06439466	1213.522	30,30,447	29,49,741	29,16,628	0.74	3.73	44.63

Table S2: Results of the demographic modelling for the five best replicate (lower AIC value) of each the eight tested scenarios. The table shows, in order of appearance, the focal studied population, the model tested, its AIC, the population mutation rate of the ancestral population (Θ), and the following parameter estimated by the model that are scaled to Θ: the change in ancestral population size (Na), the population size of the first ($Ne_1 = S^y$) and second ($Ne_2 = P^m$) population after the split, the exponential population growth parameter in S^y (b_1) and P^m (b_2), the Hill-Robertson effect (hrf) corresponding to the fraction by which Ne is reduced in region strongly affected by selection at linked site, the time at which ancestral variation in effective size occurred (T_a), the time at which derived population have split (T_s), the time of secondary contact (T_{sc}), and the fraction of the genome affected by reduced Ne (Q), and the fraction of derived mutation miss oriented in the unfolded spectrum (O). The best model is highlighted in bold.

Population	Model	AIC	Theta	Na	Ne_1	Ne_2	b1	b2	hrf	$M_{1>2}$	$M_{2>1}$	T_a	T_s	T_{sc}	Q	O	Ts/Tsc
Avellanet	**SCA2N**	**17221**	**8505**	**35.46**	**2.62**	**1.67**	**0.02**	**-**	**-**	**1**	**8.44**	**1.65**	**0**	**0.4**	**0.01**	**0.93**	**3365.85**
	SCA2NG	17462	6920	33.96	3.92	1.68	0.83	1.48	0.16	0.84	5.75	2.19	0.03	0.59	0.01	0.93	18.08
	SCAG	18414	9715	17.06	6.74	2.72	0.09	0.28	-	7.09	11.9	1.11	0.33	0.19	-	0.93	0.56
	SCA	18995	6114	11.02	2.14	2.74	-	-	-	0.92	1.66	3.48	0.02	0.06	-	0.93	2.41
	SIA2NG	18396	7129	9.71	2.34	2.04	2.53	5.43	0.02	-	-	2.92	0	-	0.94	0.93	-
	SIA2N	24261	14828	4.6	8.18	3.64	-	-	0.83	-	-	0.09	0.18	-	0.26	0.95	-
	SIAG	18444	4826	14.04	0.16	0.96	11.14	0.46	-	-	-	4.75	0.01	-	-	0.92	-
	SIA	18418	6987	9.67	0	0.01	-	-	-	-	-	3.08	0	-	-	0.92	-
Planoles	SCA2N	26078	25952	15	0.96	1.39	-	-	0.2	4.71	3.05	4.76	0.28	0.26	0.37	0.97	0.93
	SCA2NG	**26036**	**15638**	**15.16**	**1.35**	**2.07**	**1.53**	**0.61**	**0.15**	**2.33**	**2.67**	**8.55**	**0.41**	**0.25**	**0.18**	**0.96**	**0.6**
	SCA	49246	48581	42.57	0.41	0.41	5.63	6.07				1.43	0.01	0.29		0.97	38.41
	SCAG	48740	10656	17.15	2.45	2.4	0.71	0.84	1.41	1.36		13.77	0.65	0.44		0.97	0.68
	SIA2N	49784	102652	0.4	0.07	0.07	-	-	0.01	-	-	0	0.11	-	0.19	1	
	SIA2NG	47075	104479	0.02	0.01	0.01	2.42	2.18	0.2	-	-	0	0	-	0.18	1	
	SIA	69406	104153	0.33	0.05	0.05						0.01	0			1	
	SIAG	53104	195956	0.5	0.09	0.07	0.04	0.05				6.44	0			1	

Table S3. Overview of number of sites at the end of step of variant filtering by *bcftools*, along with the representative codes to implement each step. The final 11,574,426 sites were used as candidate positions for imputation and inference of SNPs by *STITCH*.

Step	Dataset	Representative Code	# sites
1	All sites	*bcftools mpileup --annotate AD,ADF,ADR,DP,QS,SP -d 500 |* *bcftools call --annotate GQ,GP -m*	495,577,994
2	Variant sites within 5bp of INDELs removed and only bi-allelic SNPs retained.	*bcftools view -m2 -M2 -v snps -e "AC==0 || AC==AN"*	17,107,390
3	SNPs removed based on depth and quality.	*bcftools filter -e "INFO/DP>130 | QUAL<20 | MQ<20"*	12,727,484
4	SNPs with ≤ 0.8 missing fraction retained.	*bcftools filter -e "F_MISSING>0.8"*	11,574,426

Table S4. Overview of number of sites before and after calling and imputation of SNPs by *STITCH*. INFO score, computed by *STITCH*, reflects confidence in imputation accuracy. Generally, INFO ≥ 0.8 is considered high confidence (93.9% of all sites).

Step	Types of sites	# sites (%)	
Before *STITCH* imputation	Initial candidate variant sites (inferred from *bcftools*)	11,574,426	–
After *STITCH* imputation	Invariant sites	41,396	(0.4%)
	Variant sites (all)	11,533,030	(99.6%)
	Variant sites (INFO ≥ 0.8)	10,873,003	(93.94%)
	Variant sites (INFO ≥ 0.6)	11,443,277	(98.9%)
	Variant sites (INFO ≥ 0.4)	11,526,221	(99.6%)

Table S5. Overview of number of sites where *A. majus* alleles could be polarised into ancestral or derived, based on alleles present in *Antirrhinum molle*. For unresolved sites, major allele in *A.majus* was considered ancestral. *unres.*: Unresolved

Type	Site Description	Ancestral Allele	Derived Allele	# sites for each type	# Total sites	
A	One of the *A.majus* allele is fixed in *A molle*.	FIXED	Other	6,624,396 (57.4%)	9,645,982 (83.6%)	Resolved sites
B	Both *A.majus* alleles present in *A.molle* and they have the same major allele (frequency ≥ 0.5).	MAJOR	MINOR	2,537,001 (22.0%)		
C	Both *A.majus* and *A.molle* are polymorphic, but only share 1 allele.	SHARED	Other	484,585 (4.2%)		
D	Both *A.majus* allel es present in *A.molle*, but they do **not** have the same major allele (frequency ≥ 0.5).	*unres.*	*unres.*	1,418,581 (12.3%)	1,887,048 (16.4%)	Unresolved sites
E	No genotype information for *A.molle*.	*unres.*	*unres.*	437,801 (3.8%)		
F	No shared allele between A.majus and *A.molle*.	*unres.*	*unres.*	30,666 (0.3%)		

Table S6. Correlation (Spearman's rho) between population genetic parameters in 10Kb non-overlapping windows. Values ≥ 0.60 are highlighted in green. Pla: Planoles, Ave: Avellanet, P: magenta-coloured var. *pseudomajus* , S: yellow-coloured var. *striatum* .

π_w				D_{XY}						F_{ST}							
AveP^m	AveS^y	PlaP^m	PlaS^y	AveP^m – AveS^y	AveP^m – PlaP^m	AveP^m – PlaS^y	AveS^y – PlaP^m	AveS^y – PlaS^y	PlaP^m – PlaS^y	AveP^m – AveS^y	AveP^m – PlaP^m	AveP^m – PlaS^y	AveS^y – PlaP^m	AveS^y – PlaS^y	PlaP^m – PlaS^y	mean recomb. rate	
0.26	0.13	0.38	0.39	0.07	0.3	0.29	0.07	0.07	0.39	-0.3	-0.23	-0.27	-0.37	-0.37	-0.08		
	0.47	0.66	0.64	0.73	0.89	0.89	0.63	0.63	0.67	-0.07	0.02	0.04	0	0	-0.01	AveP^m	π_w
		0.32	0.31	0.77	0.42	0.41	0.73	0.73	0.32	0.1	-0.03	-0.01	0.15	0.16	-0.04	AveS^y	
			0.89	0.42	0.88	0.84	0.53	0.5	0.97	-0.17	-0.17	-0.12	-0.28	-0.24	-0.02	PlaP^m	
				0.39	0.82	0.87	0.47	0.52	0.97	-0.19	-0.13	-0.17	-0.26	-0.29	-0.01	PlaS^y	
					0.62	0.61	0.92	0.91	0.42	0.36	0.07	0.11	0.35	0.35	-0.02	AveP^m – AveS^y	D_{XY}
						0.97	0.64	0.62	0.87	-0.11	0.08	0.08	-0.1	-0.09	0.01	AveP^m – PlaP^m	
							0.61	0.64	0.88	-0.11	0.07	0.1	-0.09	-0.09	0.02	AveP^m – PlaS^y	
								0.97	0.52	0.33	0.06	0.1	0.41	0.4	-0.02	AveS^y – PlaP^m	
									0.52	0.32	0.06	0.11	0.39	0.41	-0.01	AveS^y – PlaS^y	
										-0.18	-0.15	-0.13	-0.27	-0.26	0.05	PlaP^m – PlaS^y	
											0.19	0.24	0.64	0.63	0.04	AveP^m – AveS^y	F_{ST}
												0.67	0.31	0.24	0.11	AveP^m – PlaP^m	
													0.29	0.38	0.16	AveP^m – PlaS^y	
														0.86	0.07	AveS^y – PlaP^m	
															0.07	AveS^y – PlaS^y	
																PlaP^m – PlaS^y	

Table S7. Correlation (Spearman's rho) between topology weights inferred from different tree inference methods. Values ≥ 0.60 are highlighted in green.

Tg - Geography Topology				
NJ	*tsinfer*	*Relate*	*Singer*	
	0.57	0.65	0.59	*NJ*
		0.72	0.67	*tsinfer*
			0.77	*Relate*
				Singer
Tv- Variety Topology				
NJ	*tsinfer*	*Relate*	*Singer*	
	0.57	0.65	0.59	*NJ*
		0.66	0.61	*tsinfer*
			0.73	*Relate*
				Singer
Ta - Alternate Topology				
NJ	*tsinfer*	*Relate*	*Singer*	
	0.45	0.56	0.50	*NJ*
		0.62	0.57	*tsinfer*
			0.71	*Relate*
				Singer

Appendix E

Supplementary Information for

Dissecting the genetic basis of flower colour in a hybrid zone: Integrating top-down and bottom-up approaches

E.1 | Supplementary Figures

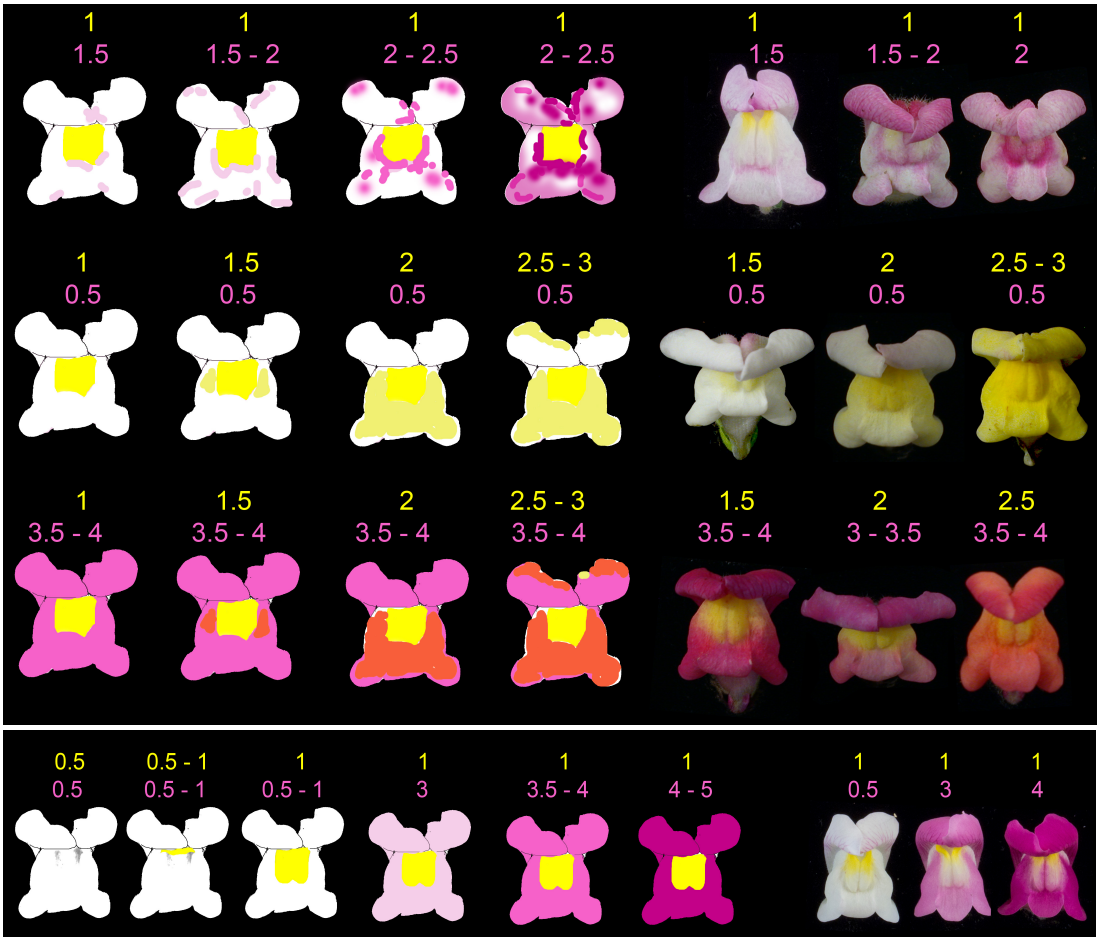

Figure S1. Colour scoring chart for field colour scores. Each flower in the field was visually scored on a scale of 0.5-5 for magenta and 0.5-3 for yellow.

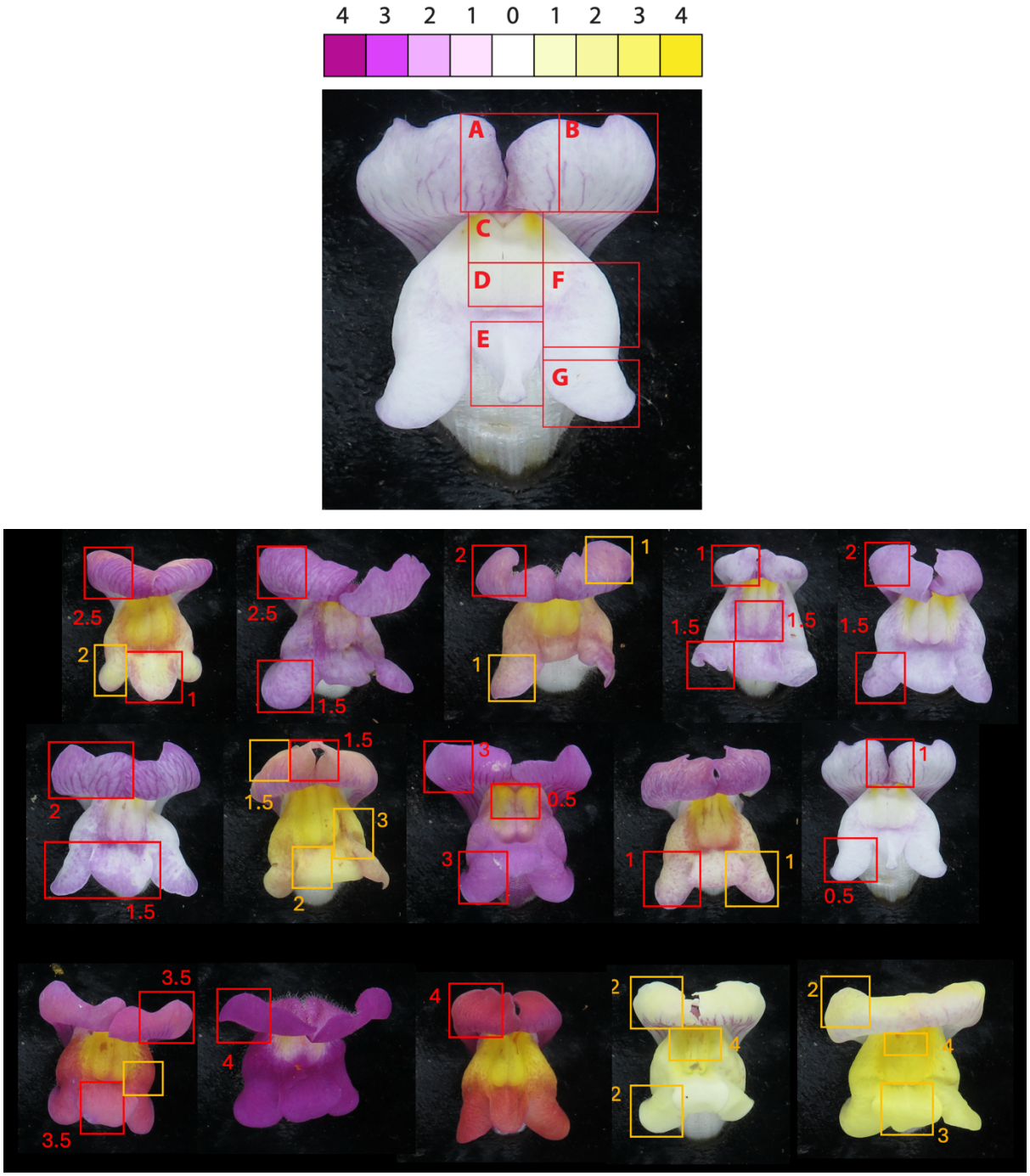

Figure S2. Colour scoring chart for box colour scores. Each flower was scored manually by 2 scorers from the images taken in the field. For each box A-G on each flower **(Top)**, scores of 0-4 were given separately for magenta and yellow. Examples of magenta and yellow scores for each box showing the distribution of colours across the flower petals **(Bottom)**.

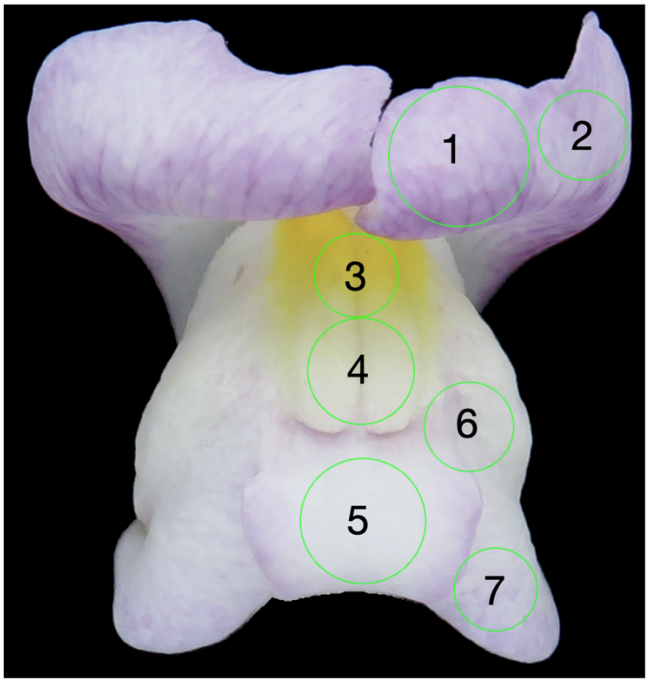

Figure S3. Circles used from different parts of the flower petals for the digital scoring method.

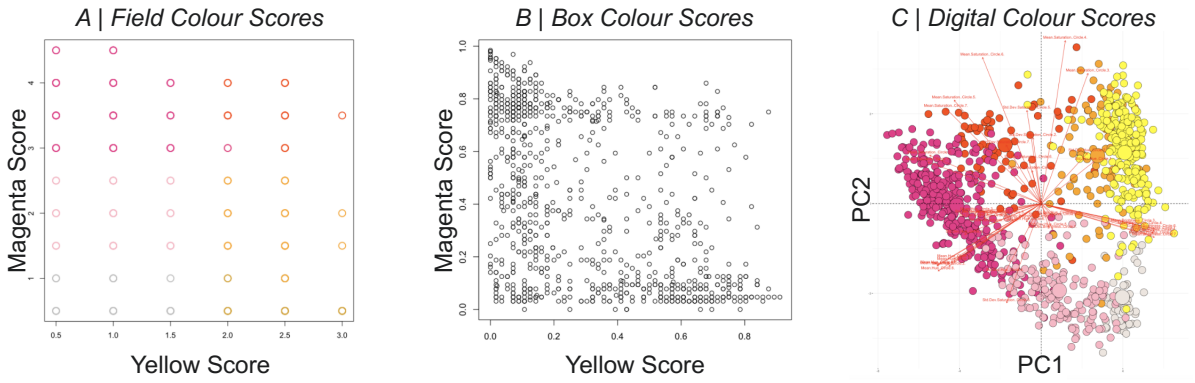

Figure S4. Correlation between **(A)** magenta and yellow field scores, **(B)** magenta and yellow box scores, and **(C)** PC1 and PC2 from digital scores.

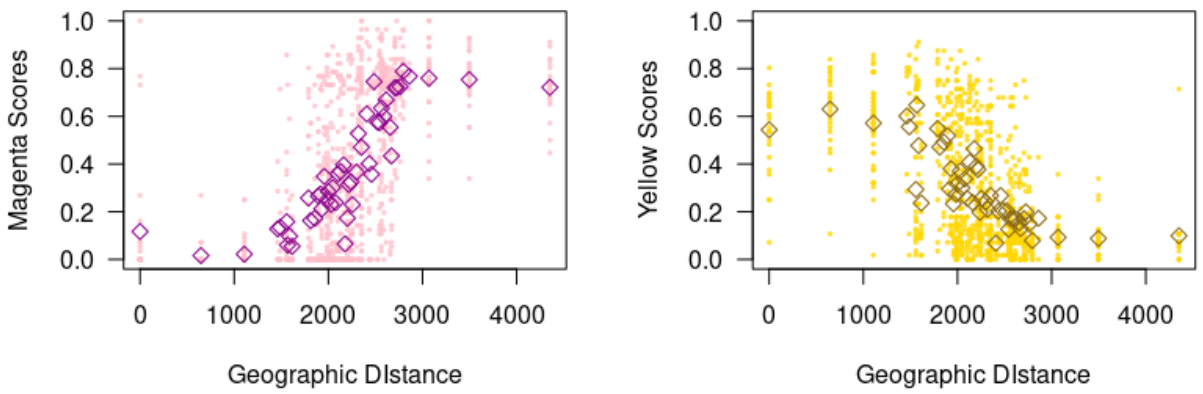

Figure S5. Raw magenta and yellow scores along the transect. Each coloured dot represents individual flowers, while diamonds represent mean colour scores in each deme.

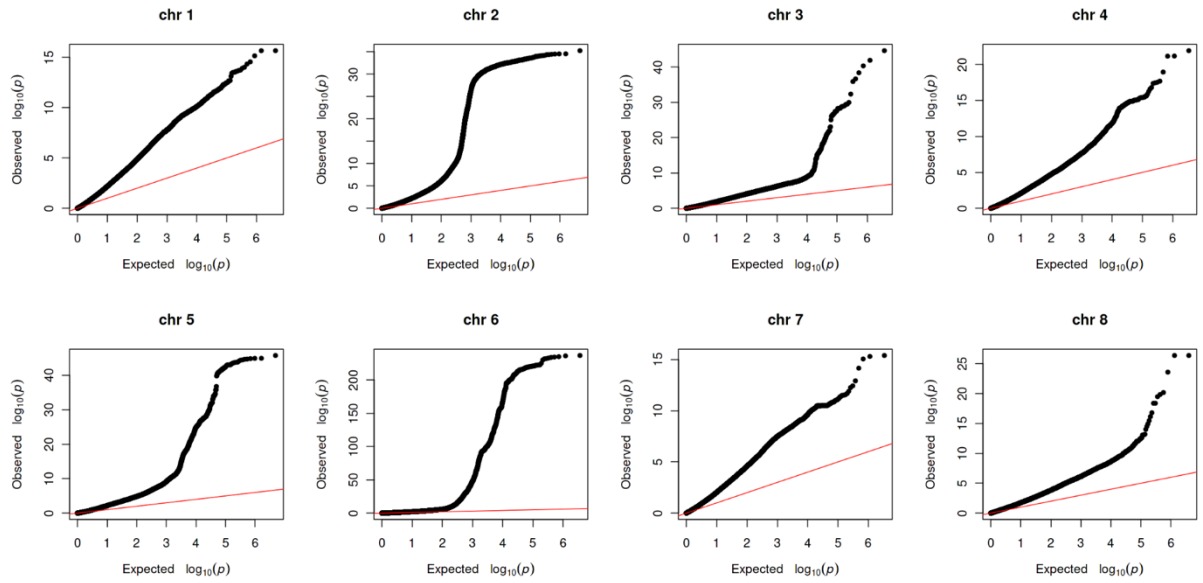

Figure S6. QQ plot for genome-wide association mapping using Linear Model (LM) on normalised magenta box colour scores for all Lower Road samples (*n* = 915).

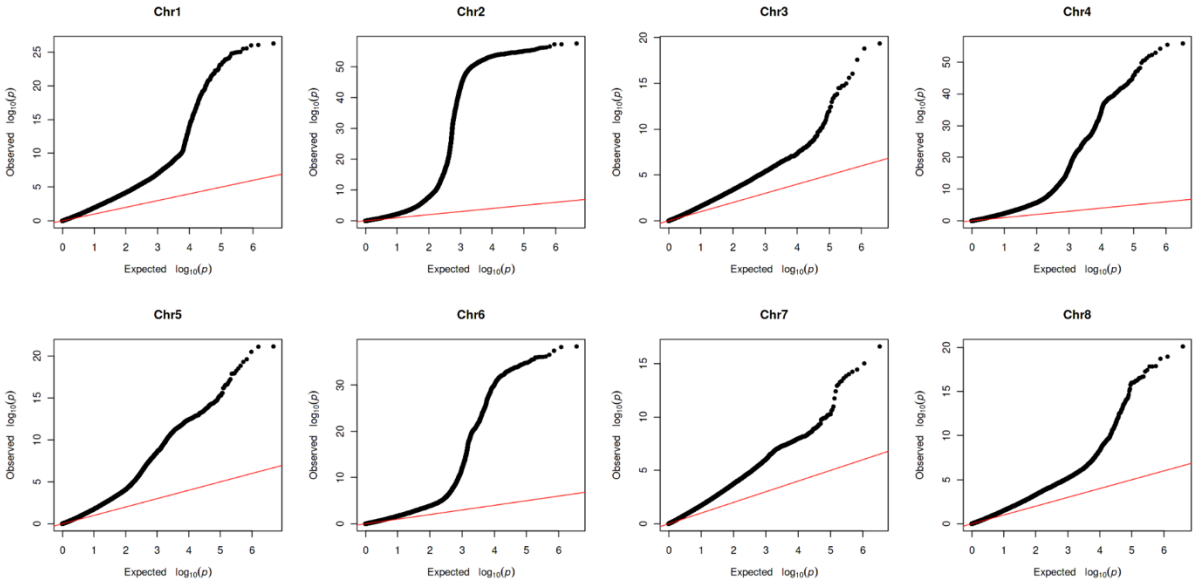

Figure S7. QQ plot for genome-wide association mapping using Linear Model (LM) on normalised yellow box colour scores for all Lower Road samples (*n* = 915).

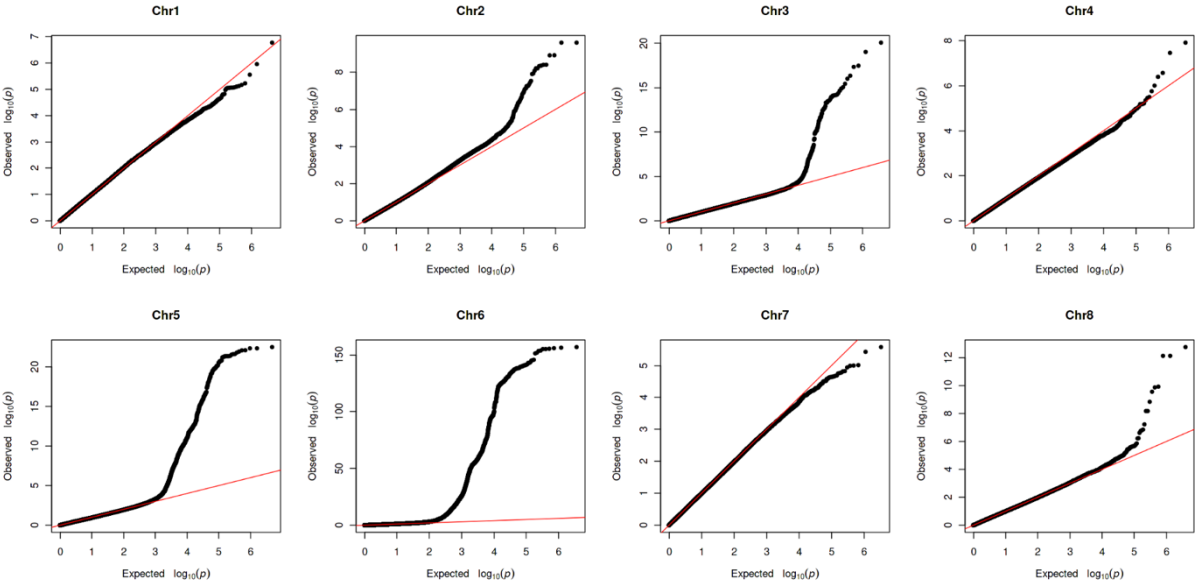

Figure S8. QQ plot for genome-wide association mapping using Linear Model (LM) on normalised magenta box colour scores with yellow score as fixed-effect covariate and kinship matrix as random effect for all Lower Road samples (*n* = 915).

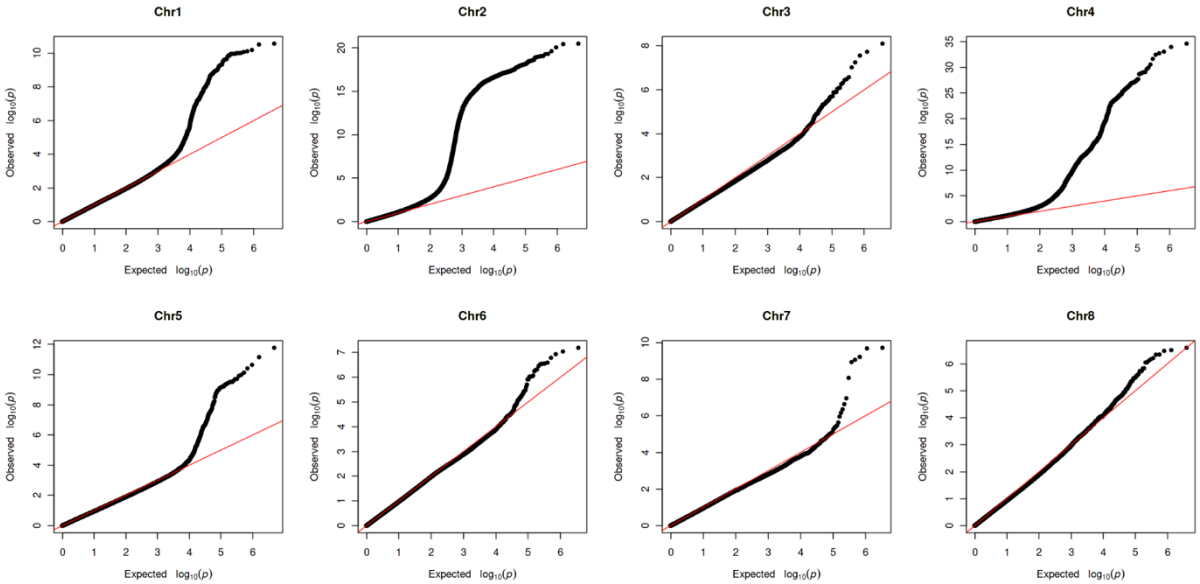

Figure S9. QQ plot for genome-wide association mapping using Linear Model (LM) on normalised yellow box colour scores with magenta score as fixed-effect covariate and kinship matrix as random effect for all Lower Road samples (*n* = 915).

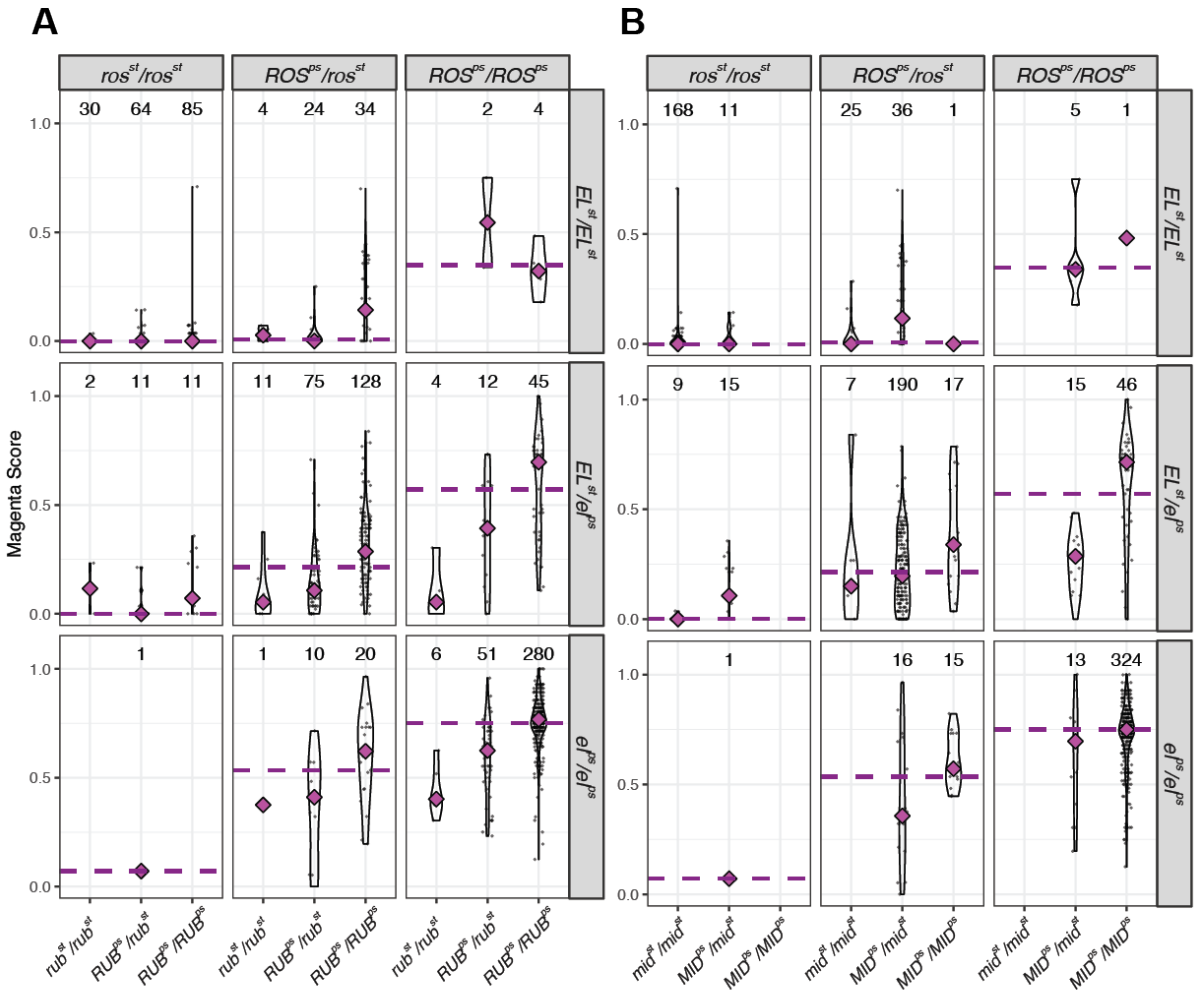

Figure S10. Magenta colour scores for samples with specific genotypic combinations at **(A)** *RUB*, *ROS3*, *EL* and **(B)** *ROS3*, *MID*, *EL*. For all plots, each dot represents the normalised magenta box score for each sample, violin plot describes the distribution of scores, and diamonds represent the mean colour score for that specific allelic combination at all 3 loci. Numbers over each violin plot denote the sample size for each allelic combination at all 3 loci. Dashed lines show mean colour scores for combinations of only *ROS3* and *EL*. *ps*: *pseudomajus*, *st*: *striatum*.

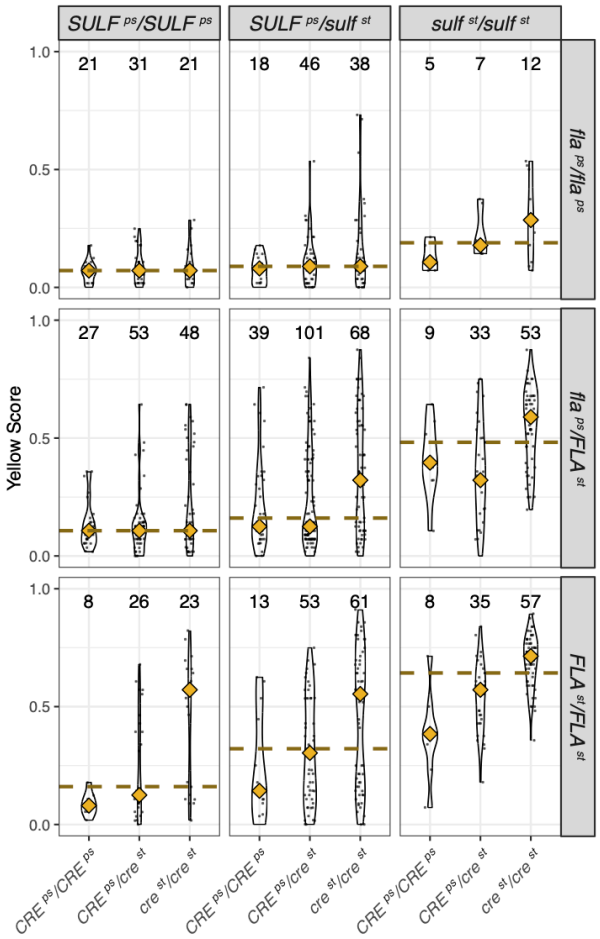

Figure S11. Yellow colour scores for samples with specific genotypic combinations at *CRE*, *FLA* and *SULF*. For all plots, each dot represents the normalised magenta box score for each sample, violin plot describes the distribution of scores, and diamonds represent the mean colour score for that specific allelic combination at all 3 loci. Numbers over each violin plot denote the sample size for each allelic combination at all 3 loci. Dashed lines show mean colour scores for combinations of only *SULF* and *FLA*. *ps*: *pseudomajus*, *st*: *striatum*.

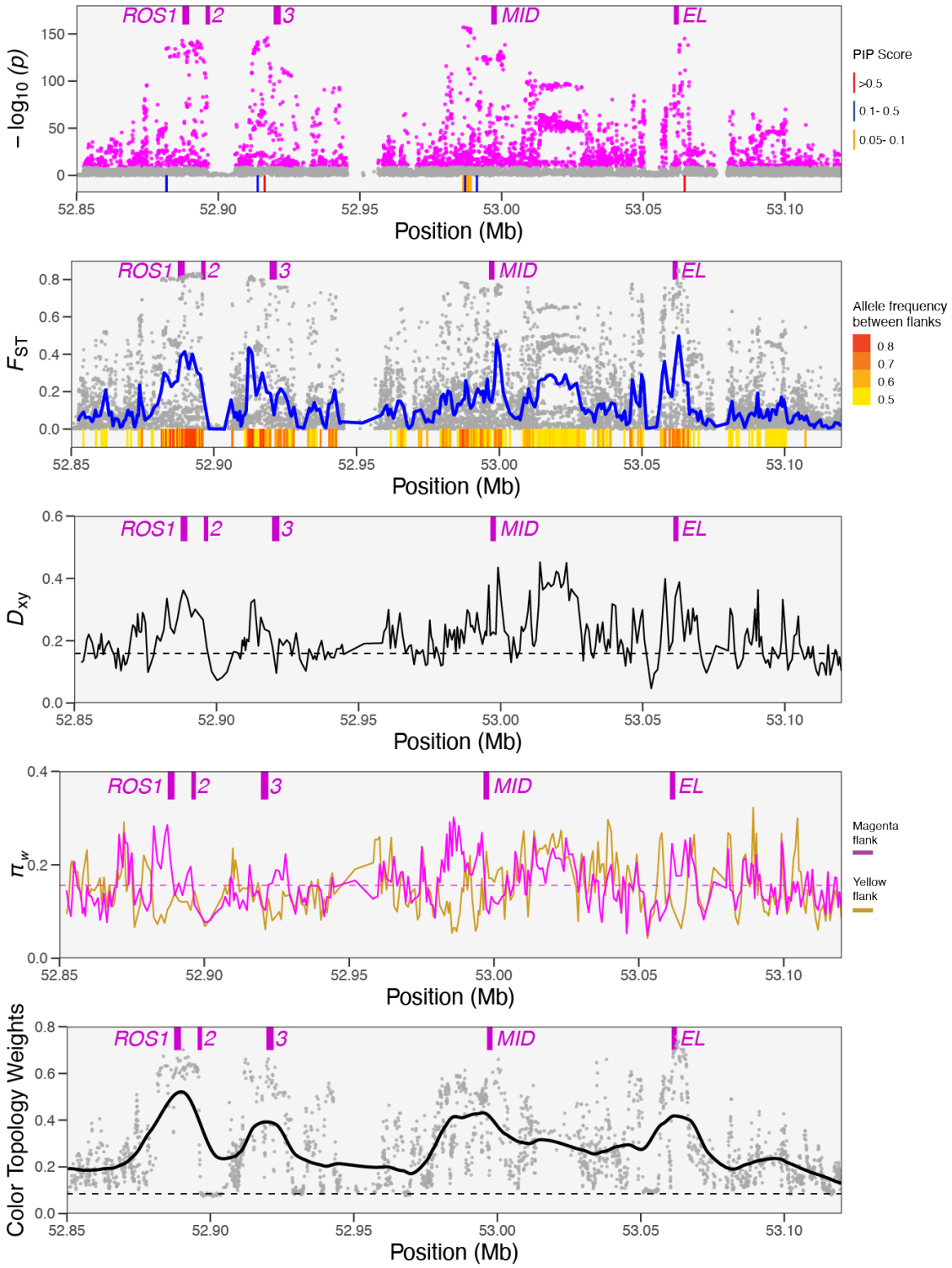

Figure S12. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in the *ROS/EL* region.

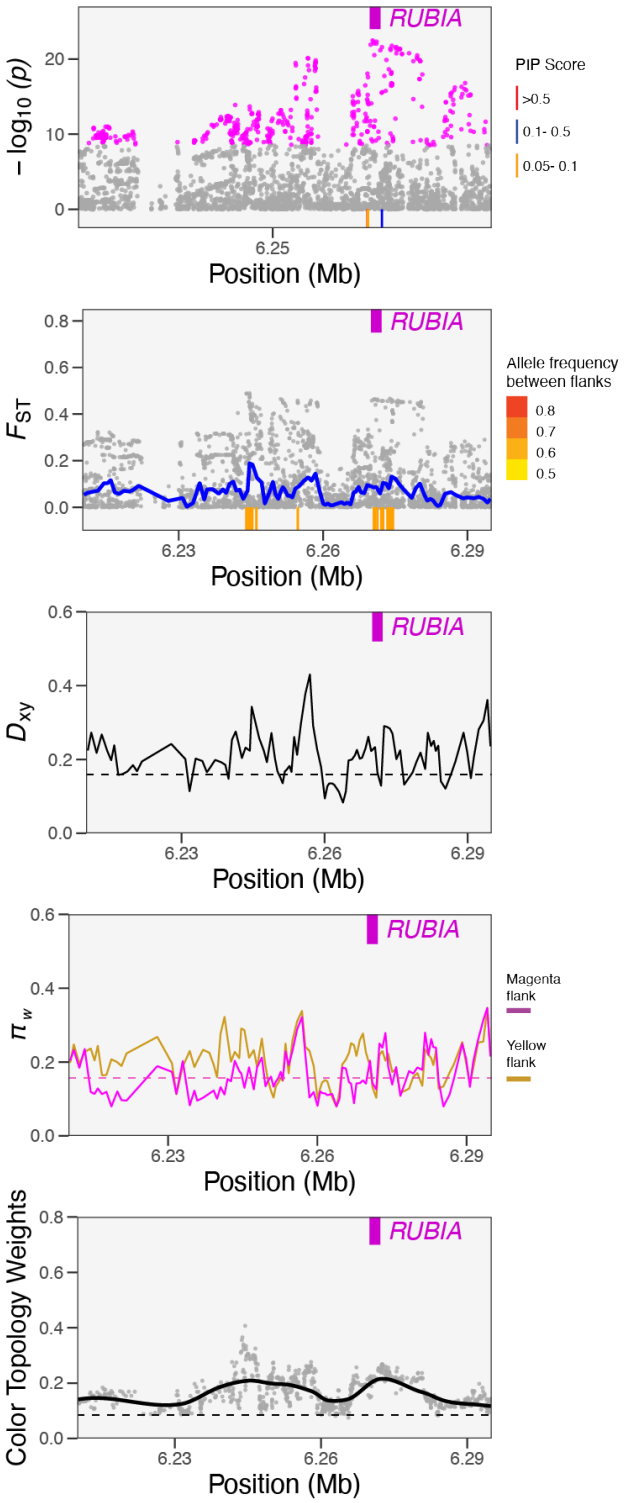

Figure S13. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in the *RUB* region.

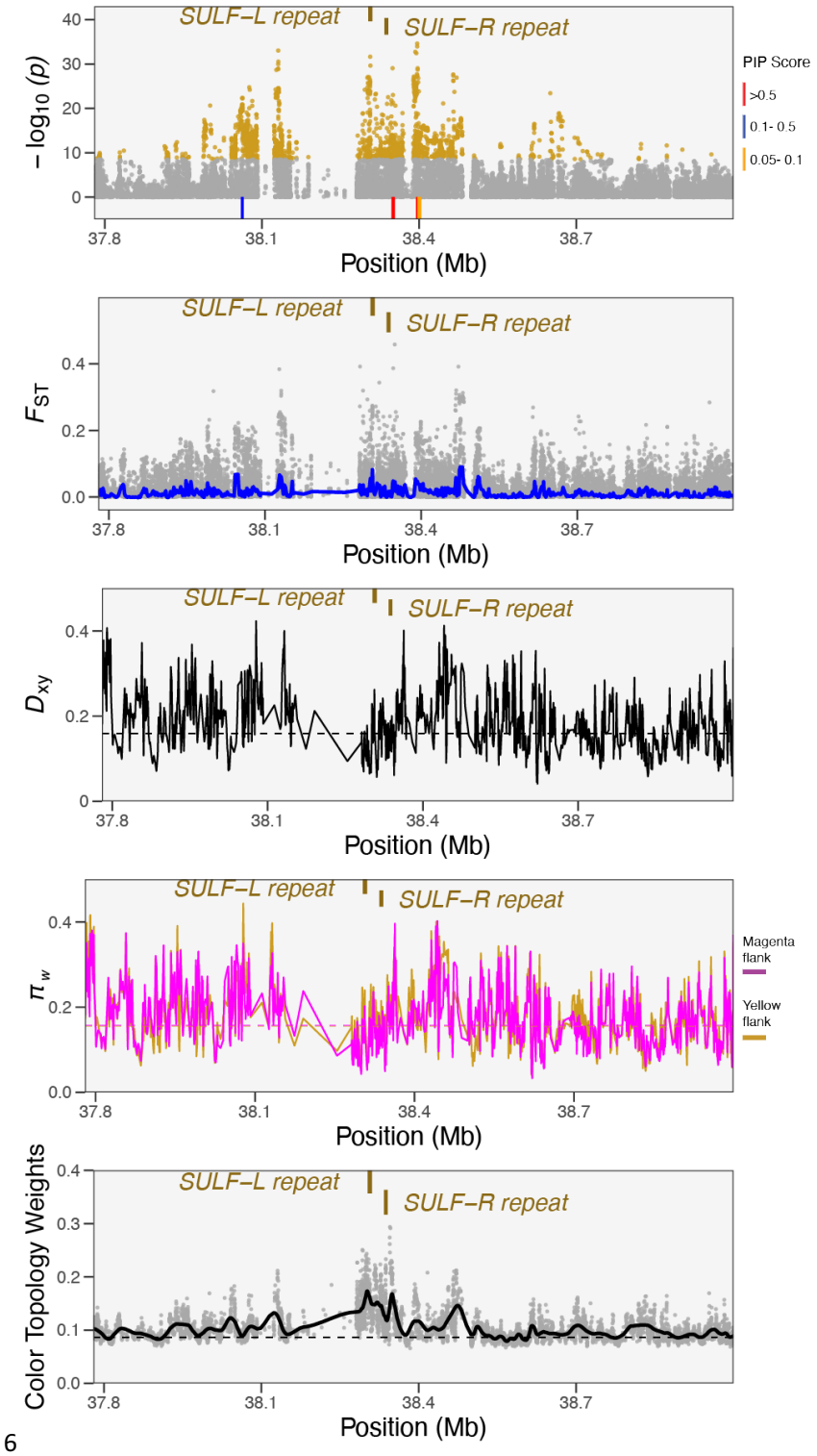

Figure S14. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in the *SULF* region.

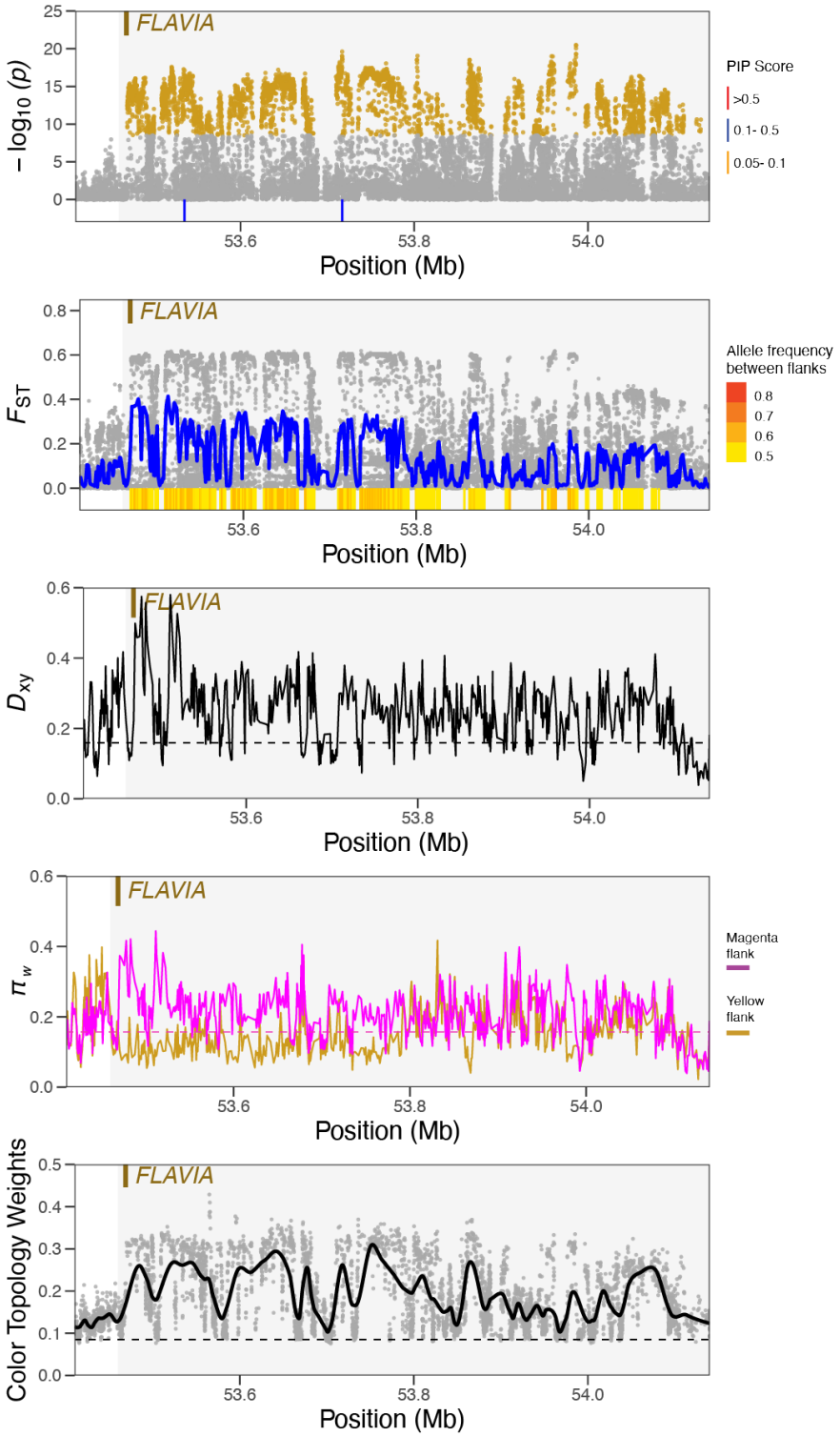

Figure S15. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in the *FLAVIA* region.

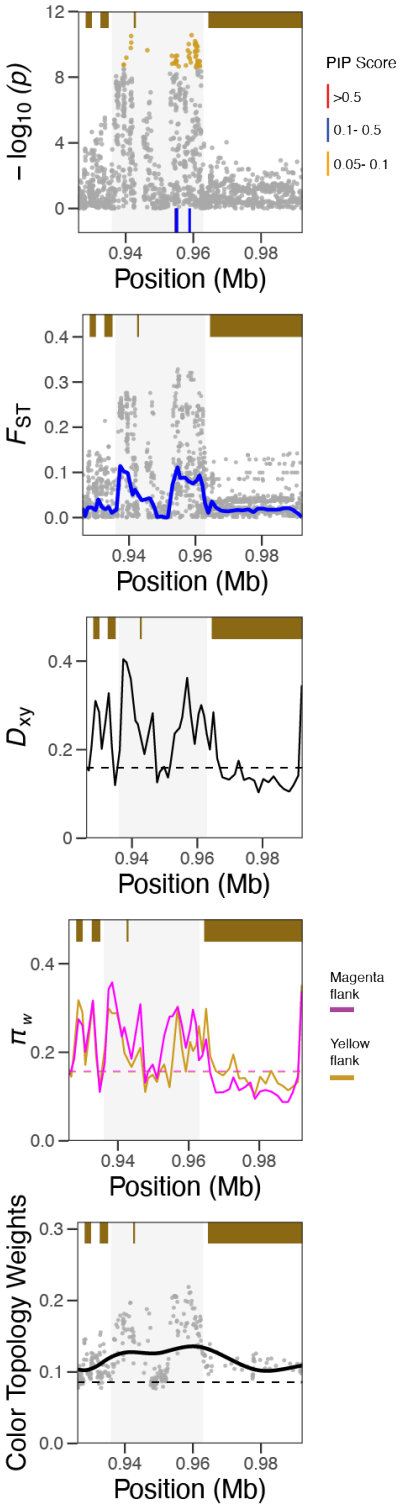

Figure S16. Details of **(A)** genome-wide association ($-\log_{10}P$ estimates from LMM on yellow colour scores and posterior inclusion probability, PIP from BSLMM), **(B)** genetic differentiation (F_{ST}) and allele frequency difference between magenta and yellow flank, **(C)** divergence between flanks (D_{XY}), **(D)** diversity within magenta and yellow flanks (π_w), and **(E)** colour topology (Tc) weights in the *CRE* region. Annotated genes with in this region are marked on top.

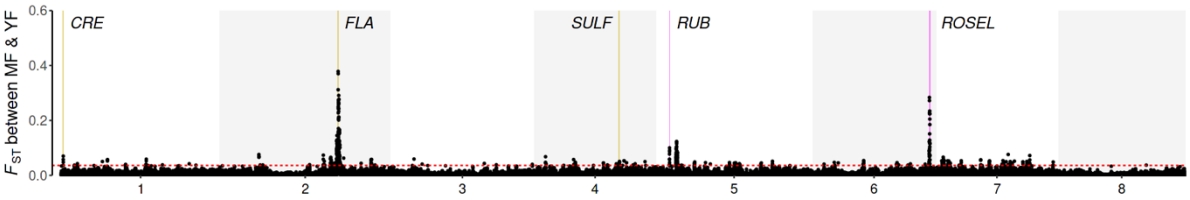

Figure S17. Genome-wide landscape of F_{ST} in 10 Kb non-overlapping windows, with the 99th and 95th percentile denoted in red and cyan dashed line. All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr 2, *SULF* in chr 4.

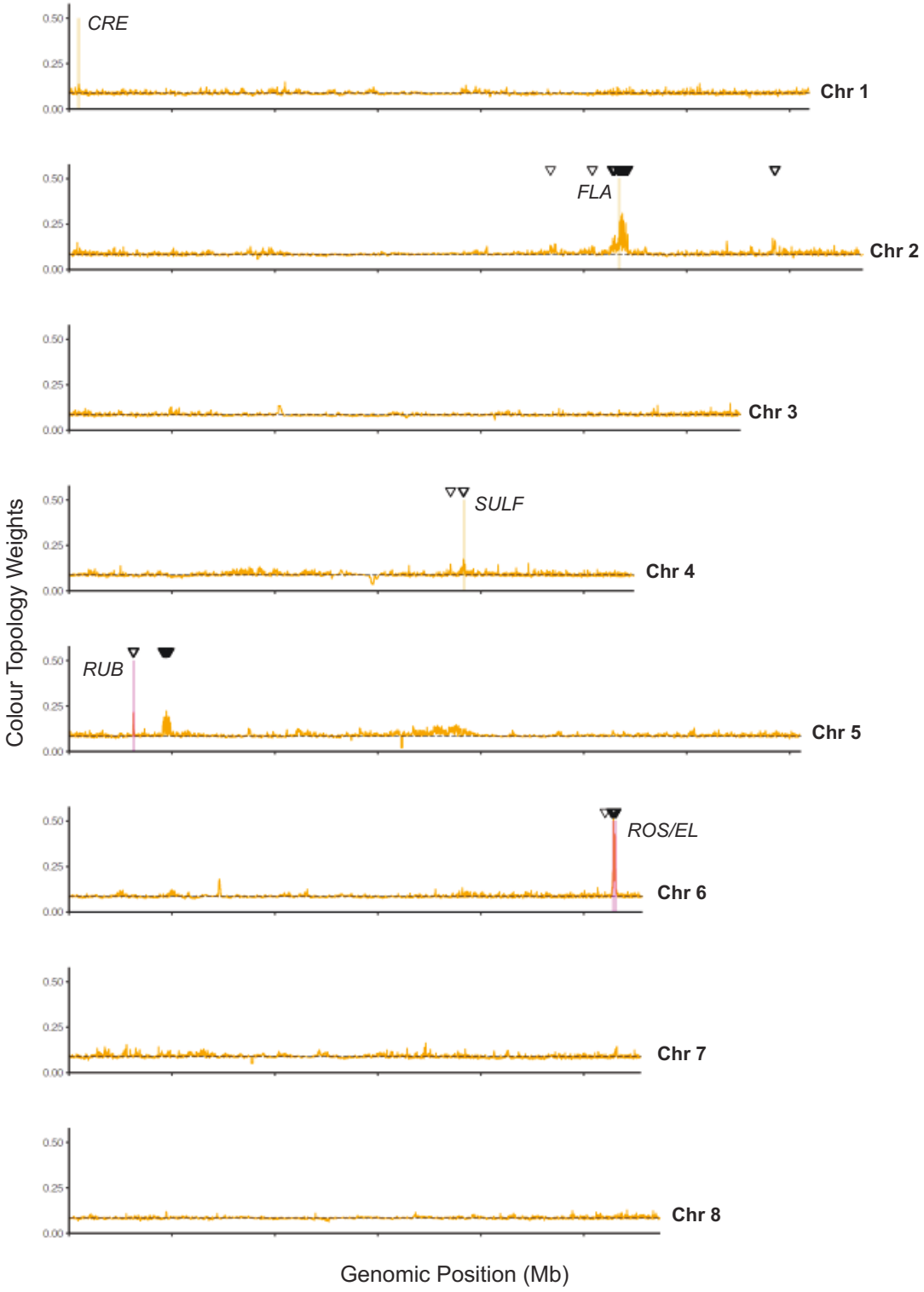

Figure S18. Genealogical landscape of colour topology (Tc) weights. Topology weights (loess smoothed, span = 50Kbp) for the 4,975,454 trees inferred by *Relate* plotted along each chromosome. Inverted triangles indicate trees with raw Tc weights ≥ 0.60. All previously described colour-controlling loci are highlighted and colour-coded magenta or yellow based on their function. Magenta-controlling loci: *RUB* in chr 5, *ROS/EL* in chr 6; and yellow-controlling loci: *CRE* in chr 1, *AUR* in chr 2, *FLA* in chr2, *SULF* in chr 4.

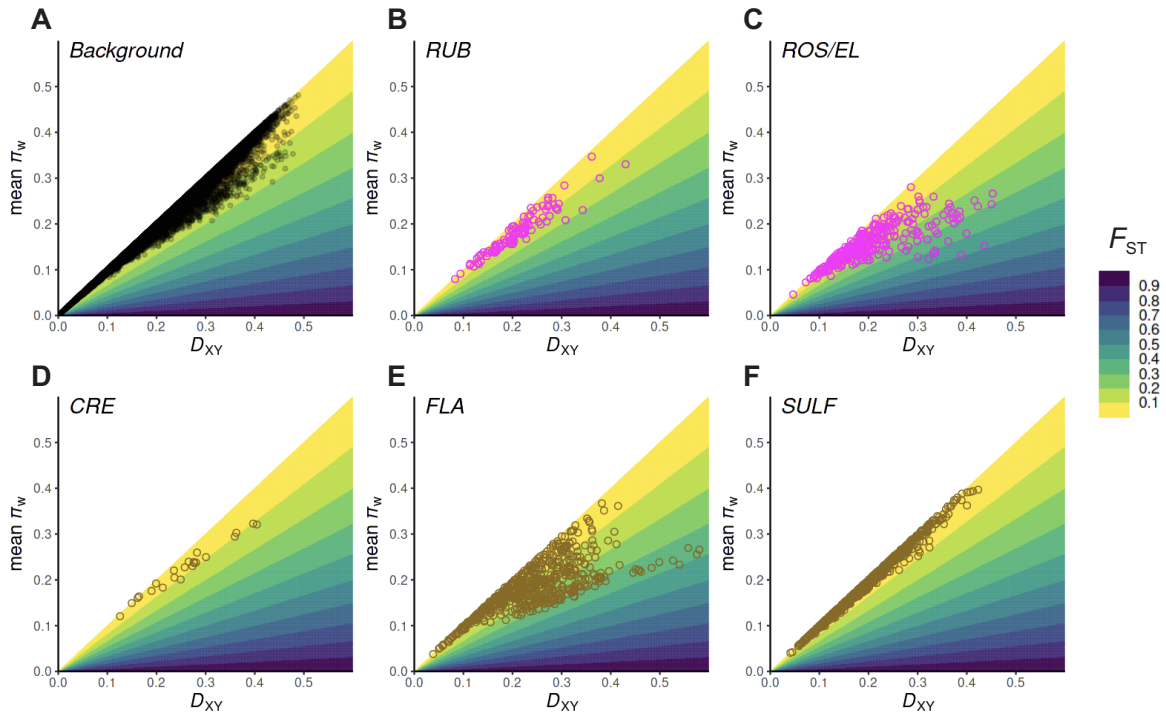

Figure S19. Relationship between divergence between (D_{XY}) and mean diversity within magenta and yellow flank (π_w). Each dot represents windows of 50 SNPs in the **(A)** genomic background, magenta associated loci—**(B)** *RUB* and **(C)** *ROS/EL*, and yellow associated loci—**(D)** *CRE*, **(E)** *FLA* and **(E)** *SULF*. Dots are coloured either black (not colour associated), magenta (anthoyanin associated) or yellow (aurone associated). The plots are shaded based on F_{ST} estimates in the same 50-SNP windows.

E.2 | Supplementary Tables

Table S1. Results of genome-wide association mapping using Bayesian sparse linear mixed models (BSLMM) on hybrids to characterized the genetic architecture of mate choice plumage traits. Posterior hyperparameter estimates (median ± standard deviation) of the proportion of phenotypic variance explained by all SNPs (PVE), proportion of variance explained by SNPs with measurable effects (PGE), and the number of SNPs in the models.

Phenotype	PVE	PGE	n SNPs
Magenta Colour	0.89 ± 0.025	0.88 ± 0.031	12 ± 6.5
Yellow Colour	0.86 ± 0.057	0.74 ± 0.095	38 ± 19.3

Appendix F

Supplementary Information for

Genealogical analysis of an island of divergence

F.1 | Supplementary Figures

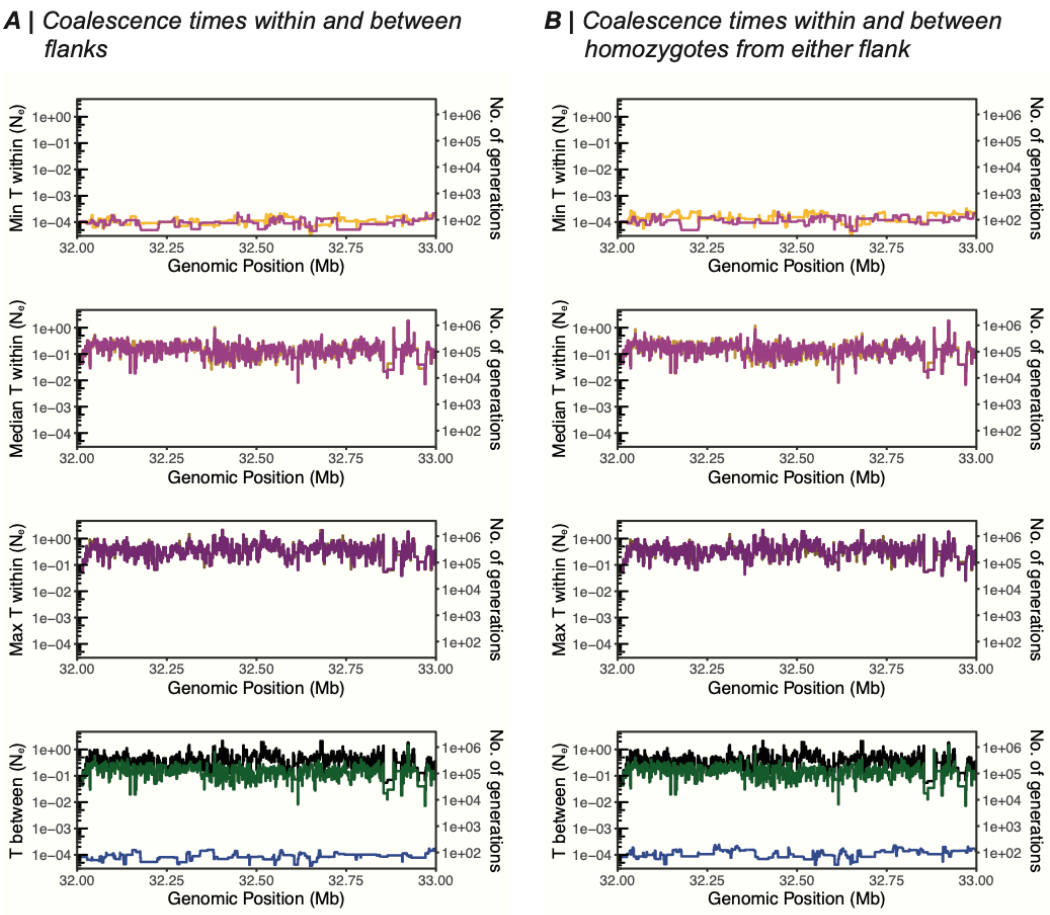

Figure S1. Coalescence history in a broader 1Mb region around *ROS/EL* within and between **(A)** all individuals from magenta and yellow flank, and (**B)** individuals homozygote for the *pseudomajus* background from magenta flank and individuals homozygote *striatum* background from yellow flank.

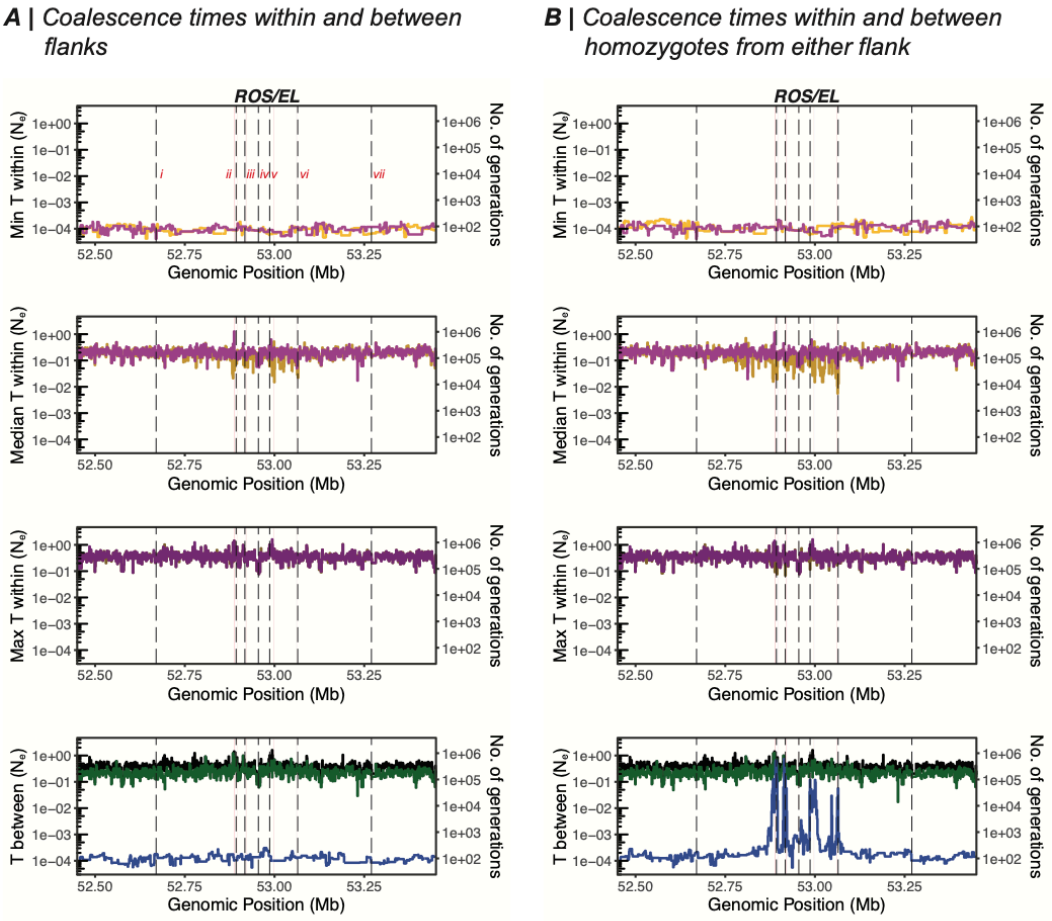

Figure S2. Coalescence history in a neutral 1Mb genomic region within and between **(A)** all individuals from magenta and yellow flank, and (**B)** individuals homozygote for the *pseudomajus* background from magenta flank and individuals homozygote *striatum* background from yellow flank. Vertical dotted lines (*i* – *vii*) in top left plot represent positions at which genealogical trees are drawn in Fig S3.

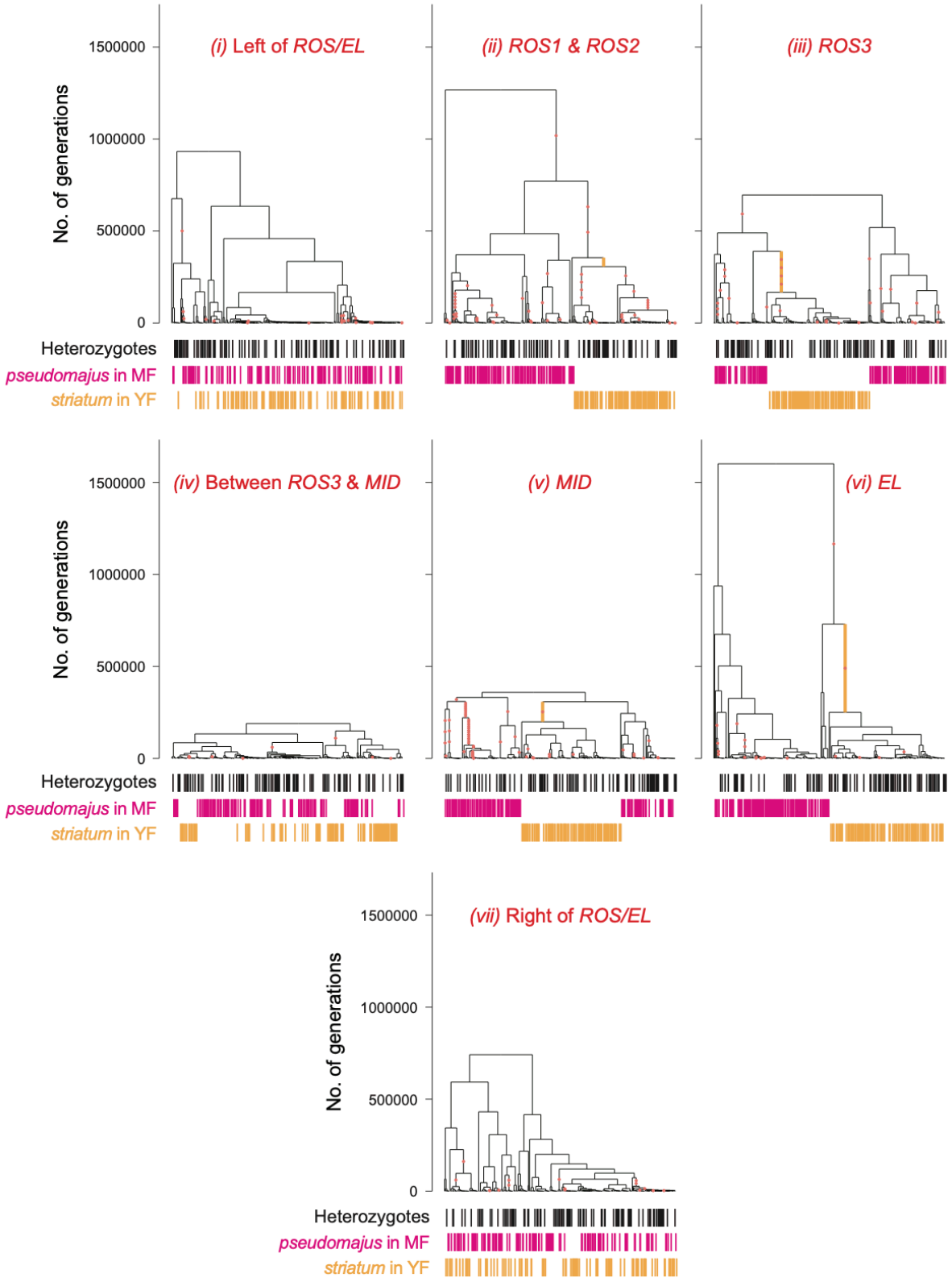

Figure S3. Genealogical trees drawn at positions denoted by vertical dotted lines (*i* – *vi*) in Fig S2. Highlighted yellow branches denote lineages ancestral to all individuals with striatum background in yellow flank. Red dots on branches denote mutations. Vertical tracks below each tree mark the group of each tree tip.

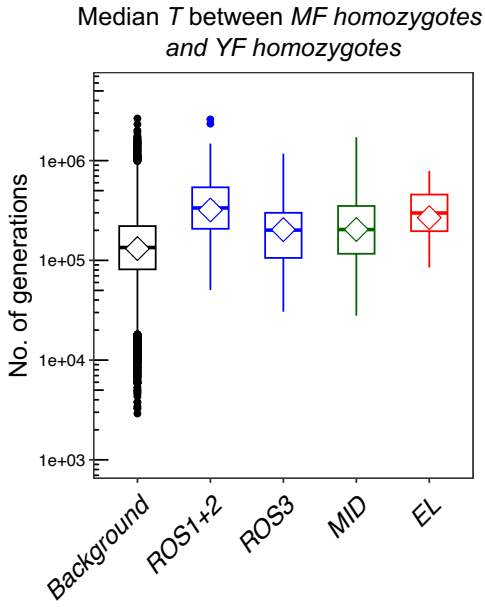

Figure S4. Comparison of median pairwise coalescence time between individuals homozygote for *pseudomajus* background in magenta flank and individuals homozygote for *striatum* background in yellow flank ($T_{b,ps/st}$) between neutral genomic background, *ROS1/ROS2*, *ROS3*, *MID*, *EL*. *ps*: *pseudomajus* background: $ROS^{ps}MID^{ps}el^{ps}$ / $ROS^{ps}MID^{ps}el^{ps}$, *st*: *striatum* background: $ros^{st}mid^{st}EL^{st}$ / $ros^{st}mid^{st}EL^{s}$