UNIVERSITY OF CALIFORNIA
SANTA CRUZ

GENOMIC CONFLICTS ACROSS EVOLUTIONARY SCALES

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in
BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Christopher Condon

August 2026

The Dissertation of Christopher Condon is
approved:

_____
Professor Russell Corbett-Detig, Chair

_____
Professor Shelbi Russell

_____
Professor David Haussler

_____
Donald Smith
Interim Vice Provost and Dean of
Graduate Studies

Table of Contents

List of Tables and Figures

Abstract

Genomic Conflicts across Evolutionary Scales

Genomic conflict arises when genetic elements increase their own proliferation despite imposing costs on their hosts. This dissertation examines how these conflicts emerge, persist, and reshape genomes across evolutionary timescales through three studies integrating gamete sequencing, comparative transcriptomics, long-read isoform sequencing, and population genomics. First, I investigated pollen-acting segregation distortion in crosses within and between *Arabidopsis lyrata* and *A. halleri* populations. Distortion was absent from crosses between closely related parents but widespread and variable in more divergent crosses, including crosses within species. Repeated, opposing distortions across chromosomes and genome-wide non-independence among loci were more consistent with negative epistatic incompatibilities than with independent single-locus drivers, demonstrating that reproductive barriers can arise early during population divergence. Second, I examined transcript processing in the non-recombining UV mating-type regions of four deeply diverged *Mamiellales* algae. Genes in these regions exhibited consistently elevated intron retention and other forms of alternative splicing. Long-read data from *Micromonas pusilla* revealed abundant aberrant isoforms with shortened coding sequences and disrupted protein domains, although many genes retained at least one likely productive transcript. These findings identify reduced splicing fidelity as a persistent, transcript-level form of degeneration in chromosomes constrained to retain

essential genes. Third, I characterized introner turnover across thirteen geographically diverse *M. pusilla* isolates. Two deeply diverged populations differed approximately fourfold in introner abundance yet showed convergent insertion patterns across genomic and functional contexts. Within the introner-rich population, allele-frequency patterns indicated ongoing gains and losses. Introner presence was associated with reduced gene expression, increased transcript isoform diversity, and greater nonsense-mediated decay, while depletion from essential genes revealed strong host filtering. Together, these results show that the evolutionary consequences of genomic conflict depend on interactions among transmission, recombination, genome architecture, molecular function, and purifying selection operating across populations, chromosomes, and molecular processes.

Acknowledgments

I could not have completed this PhD without the friendships I made in Santa Cruz, along with the endless support of my friends and family back home in San Diego. There were many times where I was unsure whether I would finish, but there was always someone there to give me the emotional space and support to help me persevere. I am grateful to my Advisor Russ Corbett-Detig, whose expertise, mentorship and support helped guide me through each of my thesis chapters. I am grateful to Manny Ares for his experimental expertise and guidance. I am deeply grateful to my labmates, present and past, including Alex, Lily, Cade and Gabe, for all their support and friendship. I also want to thank my friends in the broader PBSE community, including Chris, the first friend I made in grad school. I would not have made it this far without you.

Lastly, I am thankful for all my friends and family for their unconditional love and support. Brianna Wollard, who I met during my time in grad school, continues to be a source of love, support and inspiration to be a better person. I thank my parents, Martha and Vincent, for their boundless support, countless phone calls, patience and words of inspiration. I thank my brother Michael for his support, wisdom and friendship. I lastly want to thank my childhood friends, Shaker, Lawrence, Matthew, Zach, John and Tony. You all made me who I am today, and your friendship and support is priceless.

Introduction

Genomes are not evolutionarily unified entities. They contain alleles and mobile sequences whose success depends on increasing their own transmission, even when that success imposes costs on the organism that carries them. These selfish genetic elements exploit multiple stages of the life cycle: some elements bias which gametes survive or are transmitted, whereas others copy or mobilize themselves to create new genomic insertions. Their spread can reduce fertility, destabilize gene regulation, and restructure chromosomes, while hosts evolve suppressors and purifying selection filters harmful variants. Selfish genetic elements can therefore indirectly drive population-level evolution, where the same processes that benefit an element can consequently generate reproductive barriers, alter genome architecture, and leave persistent changes in gene expression and cellular function.

One route by which genetic conflict becomes visible is biased transmission. Under Mendelian inheritance, alternative parental alleles are expected to be represented equally among viable gametes, but segregation distortion arises when particular haplotypes are preferentially transmitted or when incompatible allele combinations reduce gamete survival. Such distortion may reflect a classic drive system, in which a selfish allele directly favors its own inheritance, or a genetic incompatibility between loci that evolved in different populations. These alternatives predict different genomic patterns, where single-locus drivers act largely independently, but Dobzhansky-Muller-type incompatibilities can produce paired, opposing distortions across chromosomes. Here, I use pooled pollen sequencing from

Arabidopsis hybrids spanning within-region, between-region, and between-species crosses to ask when these effects arise during divergence and whether their genomic organization is more consistent with drive or with negative epistasis.

A second route to selfish-element success is proliferation through transposition. Transposable elements can increase in copy number by moving or replicating within a genome, but the probability that a new insertion persists depends on local recombination, chromatin state, gene function, and the molecular consequences of insertion. These interactions are especially important in regions where recombination is suppressed, because reduced efficacy of selection and repeat accumulation can promote genomic erosion even when essential genes must be retained. They are also central to introner evolution: introners are mobile elements whose insertion into coding sequence creates spliceosomal introns, directly coupling transposition to transcript processing. Chapters two and three examine these complementary scales, from the long-term transcript-level consequences of recombination suppression in algal mating-type chromosomes to the population-level gain, loss, and functional filtering of an actively proliferating introner system.

Together, these projects address how selfish transmission and element proliferation are constrained by genomic context and converted into organismal and evolutionary consequences. By integrating gamete sequencing, comparative transcriptomics, long-read isoform sequencing, and population genomics, this dissertation connects transmission bias, genome architecture, and transposable-element turnover within a common framework of genomic conflict.

[The following is an adaptation of a publication which has been published in Evolution Letters: https://doi.org/10.1093/evlett/qraf013]

Chapter I: Diverging Arabidopsis populations quickly accumulate pollen-acting genetic incompatibilities

1.1 Abstract

The process by which species diverge from one another, gradually accumulate genetic incompatibilities and eventually reach full-fledged reproductive isolation is a key question in evolutionary biology. However, the nature of reproductive barriers, the pace at which they accumulate and their genomic distribution remain poorly documented. The disruption of co-adapted epistatic interactions in hybrids and the accumulation of selfish genetic elements are proposed contributors to this process, and can lead to the distortion of the mendelian segregation of the affected loci across the genome. In this study we detect and quantify segregation distortion across the genomes of crosses produced from a diverse sampling of *Arabidopsis lyrata* and *A. halleri* populations, two species at the early stages of speciation and that can still interbreed. We observe no distortion loci in crosses with geographically and genetically similar parents, but both their frequency of occurrence and their magnitude become highly variable in more distant crosses. We also observe that distorter loci evolve rapidly, as they occur not only in interspecific hybrids, but also in intraspecific hybrids produced by crossing individuals from two isolated regions. Finally, we identify both genome-wide non-independence and two specific genomic regions on different chromosomes where opposite distortion effects are repeatedly observed across multiple F1 individuals, suggesting negative epistasis is a major contributor to the evolution of hybrid segregation distortion. Our study demonstrates that pollen-acting segregation distortion is ubiquitous, and may contribute not only to

the ongoing reproductive isolation between *A. halleri* and *A. lyrata*, but also between recently diverged populations of the same species.

1.2 Introduction

An essential goal of evolutionary biology is to identify the genetic basis and underlying biological processes for speciation. Dobzhansky-Muller incompatibilities (DMIs) are thought to be a major contributor to speciation, as it is a primary explanation for the accumulation of reproductive incompatibilities between isolated or diverging populations (Coyne 1992; Coyne and Orr 2004; Dobzhansky 1982; Muller 1942). The DMI model hypothesizes that interactions between different loci explains hybrid breakdown. Hybrid genomes contain new combinations of alleles that have not co-evolved together. As a result, most of these new interactions consist of negative epistasis (Simon et al. 2018). The total amount of incompatibilities is expected to depend on the level of molecular divergence between parental genomes, driving the transition from "populations" to "isolated species" (Roux et al. 2016). DMIs may accumulate sufficiently rapidly to produce population-specific incompatibilities, representing the first steps of the speciation process (Bomblies et al. 2007; R. B. Corbett-Detig et al. 2013; Seymour et al. 2019). However, the pace with which incompatibilities accumulate between populations and between species and how they manifest is only poorly understood (Matute and Coyne 2010; Ravinet et al. 2017).

An early observation in the field (Haldane 1922; Cowell 2023) is the disproportionate expression of incompatibilities in hemizygous and heterogametic

hybrid individuals (Gadau et al. 1999; Turelli and Orr 1995), suggesting that DMIs contributing to hybrid sterility and inviability are generally partially recessive (J 1940; Muller 1942). Therefore, those expressed in haploid tissues (where their expression is not restricted by dominant alleles as would be the case during the diploid phase in a heterozygous individual) should be particularly influential in reproductive isolation. Specifically, factors acting during the haploid phase of the life cycle can directly affect fertility of F1 individuals by compromising proper gamete development, hence potentially representing efficient barriers to gene flow. Several studies have shown the presence of both segregation distortion and impacted viability in hybrid gametophytes (Lin et al. 1992; McDaniel et al. 2007; McDermott and Noor 2012). The haploid phase of the life cycle therefore offers particularly tractable and appealing insights to reveal processes contributing to the formation of species.

Genetic incompatibilities may contribute to gametic inviability in two ways (illustrated in Figure 1.1). Under the classical DMI model, as mutations accumulate in diverging populations, each has a chance of generating a negative interaction with alleles that evolved in the other lineage in a hybrid genetic background (Orr 1996). Genetic incompatibilities can also contribute to gametic inviability by the evolution of selfish genetic elements that either bias the meiotic process, or destroy or disable those gametes that do not inherit them (Burt and Trivers 2008). Such selfish elements are often coupled with suppressor alleles, masking their effects within populations. When these co-evolved driver and suppressor alleles are uncoupled in hybrids, reactivated drivers can reduce hybrid fertility (Phadnis and Orr 2009). The spread of

deleterious selfish elements reduces individual fitness, and can contribute to the build up of reproductive isolation in case their deleterious effects are stronger in hybrids than within each of their two parental populations. Either model can result in segregation distortion of incompatible alleles and distinguishing among these models is critical for understanding the contribution of haploid-acting factors to speciation.

Direct gamete sequencing is a promising approach to detecting segregation distortion in the gamete phase, and it does so by bulk sequencing the gametes of an individual and detecting a skew in the ratios of maternal to paternal alleles (Carioscia et al. 2022; R. Corbett-Detig et al. 2019; R. B. Corbett-Detig et al. 2015). The classical approach to reveal segregation distortion is to focus on interspecific crosses by generating F1 progeny, intercrossing F1s, then generating a large population of F2 individuals (Fishman and Willis 2005; Leppälä et al. 2013; Lin et al. 1992). This approach is limiting because it is challenging to raise a sufficiently large F2 population to detect loci that distort within F1 individuals (unless their effect is very strong), and because it is usually impossible to distinguish whether segregation distortion is occurring in the gametic phase or in the post-zygotic phase. Gamete sequencing overcomes these limitations, as it only requires generation of a single F1 hybrid, which eliminates the burden of generating an F2 population, and allows for investigation of a greater variety of F1 cross types. Another advantage is that each gamete (or gametophyte in the case of pollen) is effectively an individual in a population of millions or more covered by the analysis, and thus greatly increases the power to detect extremely small distortion effects that would otherwise be impossible

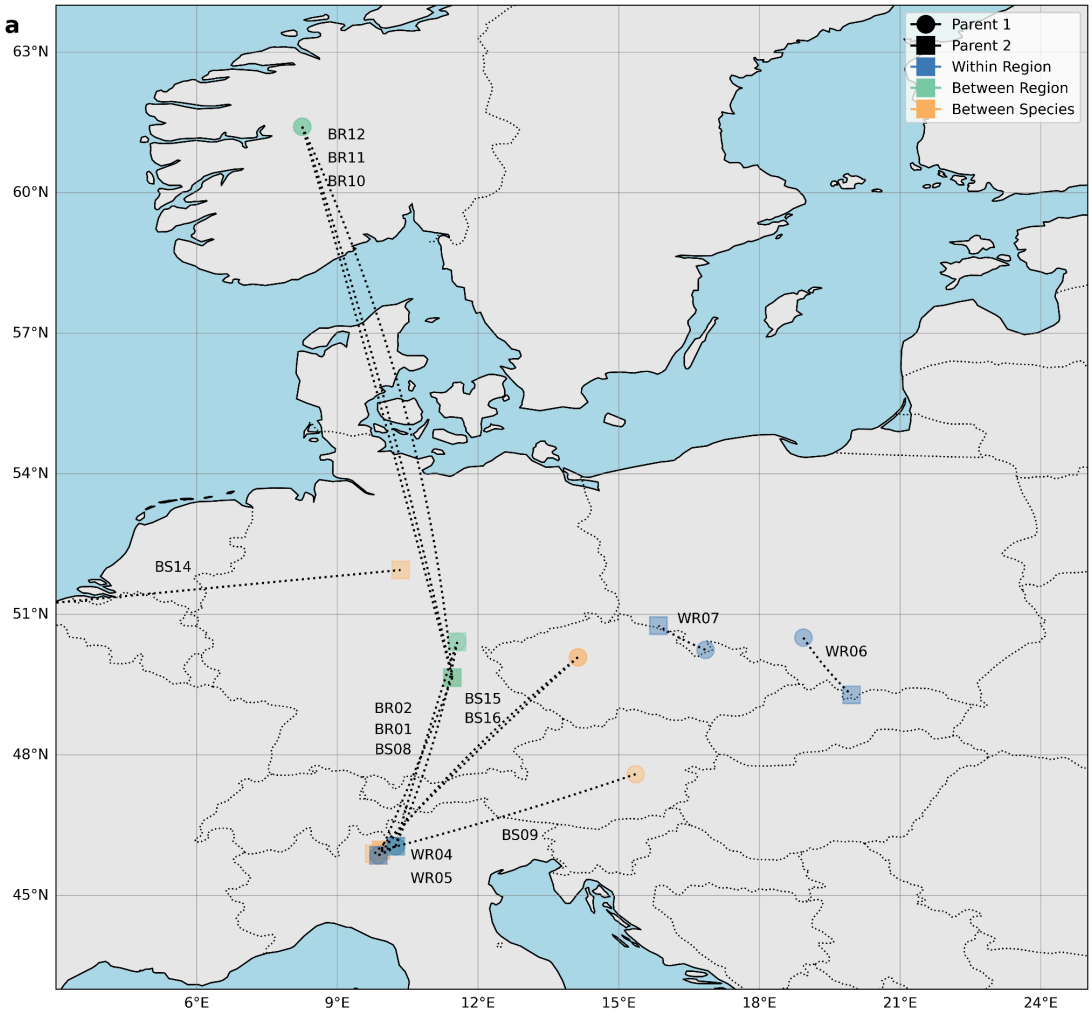
to measure. Finally, because gametes are sequenced prior to the zygotic phase, this ensures that any potential segregation distortion is occurring either during or right after meiosis, and that it involves processes that destroy rather than simply inactivate gametes.

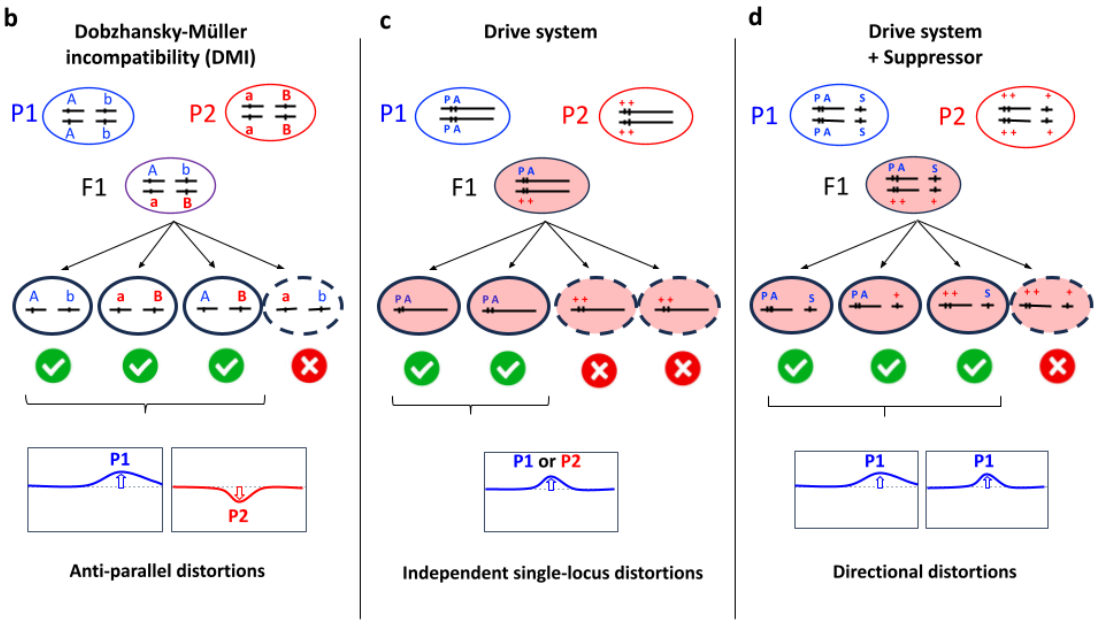
Bulk pollen sequencing has previously been used to identify segregation distortion in F1 hybrids derived from crosses between *Arabidopsis lyrata* and *A. halleri* (R. Corbett-Detig et al. 2019), but this earlier study left important questions unaddressed. *A. lyrata* and *A. halleri* are two recently diverged species in the Camelineae tribe of the Brassicaceae family that are currently undergoing reproductive isolation from one another (Roux et al. 2011; W.-K. Wang et al. 2010). Despite this, they are still capable of forming viable and fertile hybrids, and may still occasionally exchange genes between their natural populations (Castric et al. 2010; Novikova et al. 2016; Roux et al. 2011). Previous work using genetic markers (Leppälä et al. 2008) showed that segregation distortion is detectable in offspring both within and between *A. lyrata* populations, though distortion frequency of the latter group was much higher, suggesting an accumulation of DMIs. In our previous study (R. Corbett-Detig et al. 2019), we used bulk gamete sequencing on two F1 hybrids from *A. lyrata* x *A. halleri* crosses and identified segregation distortion on three chromosomes in both hybrids. However, this study did not answer questions related to the accumulation dynamics of segregation distortion, and whether it is a phenomenon that strictly occurs once lineages have diverged into different species.

Here, we investigate the pace of evolution and effect sizes of segregation distorters by identifying them at two different stages along the speciation process: between populations and between species of the *Arabidopsis* genus. We use gamete sequencing to measure segregation distortion in fourteen unique F1 hybrids, derived from a diverse combination of crosses within and between species in *A. lyrata* and *A. halleri,* encompassing a range of genetic distances between the parental individuals. By using gamete sequencing data from all F1 progeny, and somatic sequencing data from both F1 hybrids and their parents, we show widespread occurrences of segregation distortion loci across a majority of F1 hybrids. We estimate the genomic locations of these loci, and identify antiparallel distortion patterns that are consistent with the DMI rather than the selfish elements model. Distorter frequency and effect sizes are highly variable, but are generally higher in between-region and between-species crosses. Our results imply that segregation distortion in the haploid phase acts early during speciation, and is a likely contributor to hybrid infertility.

Figure 1.1 Parental accessions and models of segregation distortion

(a). Geographic map of parental accessions and crossing schemes used in this study. Illustration of three possible models of segregation distortion. (b). The DMI model illustrates gametophytic incompatibility between alleles a and b (derived alleles, independently fixed in either parental species), resulting in antiparallel distortions towards alleles from P1 and from P2 between the two loci. (c). In drive systems, a toxic element is expressed at the diploid stage, and only gametic products inheriting the antidote survive. All drive systems documented rely on genetic elements that are tightly linked (either a Poison-Antidote, PA system as represented here, or a Killer-Target system where a "killer" element is produced by one chromosome selectively kills chromosomes carrying a "target" element in trans). Such single-locus drive systems are expected to result in independent distortion patterns across chromosomes carrying them (either towards P1 or towards P2, but independently from one another). (d). Suppressors (S) of drive systems are commonly found in natural populations, and are often genetically unlinked to the original driver elements. In such a case, the suppressor would rescue gametes carrying it, resulting in distortion towards the parent contributing to the driver and the suppressor.


a
Parent 1
Parent 2
Within Region
Between Region
Between Species
63°N
60°N
57°N
54°N
51°N
48°N
45°N
6°E
9°E
12°E
15°E
18°E
21°E
24°E
BR12
BR11
BR10
BS14
WR07
WR06
BR02
BR01
BS08
BS15
BS16
BS09
WR04
WR05



1.3 Methods

1.3.1 Plant Sampling and Crosses

Parental plants were collected in natural populations and grown and propagated vegetatively under greenhouse environments either at the university of Cologne (crosses BS08, BS09, BR10, BR11 and BR12) or at the university of Bochum (crosses BR01, BR02, WR04, WR05, WR06, WR07, BS14, (Stein et al. 2017)). We crossed parental plants by hand pollination and one F1 individual was selected for each cross. The F1s were also grown under their respective greenhouse conditions and brought to flowering. To evaluate whether the distorters revealed in (R. Corbett-Detig et al. 2019) were fixed within species, we chose parents for inter-specific hybrids from geographic origins different from those in (R. Corbett-Detig et al. 2019). For F1s within *A. halleri*, we crossed parents either from

close geographic origins or from distant accessions. For F1s within *A. lyrata*, all parents came from accessions that were geographically relatively distant (Table S1).

Following (R. Corbett-Detig et al. 2019), we collected 20-50 open flowers in 95% EtOH in a 50mL Falcon tube for each F1. The tubes were gently vortexed to suspend pollen, and first centrifuged at low speed to collect flowers in the bottom. We then pipetted the supernatant with pollen in a clean tube, which we centrifuged for 10 minutes at 13,000 rpm. We discarded the supernatant and let pollen pellets dry overnight at room temperature. With an estimated 18,000 pollen grains per flower in *A. lyrata* (Willi 2013), we estimate that we sequenced up to 360,000 to 900,000 individual pollen grains per cross.

1.3.2 Library Preparation and Sequencing

For *A. halleri* parents, between 50 and 70 mg young leaf tissues were placed in a 1.5-mL polypropylene tube and shock-frozen in liquid nitrogen. A stainless steel metal bead (5 mm diameter) was added per tube, and leaf tissues were homogenized in adapters pre-cooled to -20°C using a Retsch mixer mill (Type MM 300, Retsch, Haan, Germany) for 1 min at 30 Hz. DNA was extracted from this homogenate using the NucleoMag® Plant kit (Macherey Nagel, Düren, Germany) on an epMotion 5075 robot according to the manufacturer's instructions (Eppendorf, Hamburg, Germany). The quality, quantity, and integrity of DNA samples were assessed on 0.8% (w/v) TAE-agarose gels, with a Nanodrop 2000 spectrophotometer (ThermoFisher Scientific, Darmstadt, Germany) and a Qubit 2.0 Fluorometer (Invitrogen, Darmstadt, Germany) using the dsDNA BR Assay Kit. The Illumina TruSeq DNA PCR-Free

Library Prep kit was used for producing libraries on an epMotion 5075 robot according to the manufacturer's instructions, with assessment of quantity and quality of PCR-free libraries on a Qubit 2.0 Fluorometer and an Agilent 2100 Bioanalyzer using dsDNA HS Assay kit. Libraries passing QC were diluted, pooled and dispatched to Novogene Europe (Cambridge, United Kingdom) for obtaining 150-bp paired-end sequencing reads at a targeted minimum depth of 20x genome coverage using an Illumina NovaSeq 6000 instrument (Illumina, San Diego, USA). Conditions for DNA extraction from leaves of *A. lyrata* parents, library construction and sequencing were described in (Takou et al. 2021). For F1 individuals, we extracted genomic DNA from ground leaf tissue using the Macherey Nagel Nucleospin Plant kit. For pollen samples, ten ceramic spheres were added to a tube containing a pollen pellet and crushed with Mp Biomedicals Lysing Matrix D and spun using the Fast prep 24-5G quick prep rotor with the following manual program: 2x(20 seconds at 4.0m/s). Following (R. Corbett-Detig et al. 2019), we isolated genomic DNA from pollen using the Macherey Nagel Nucleospin food kit with the slight modification that we used columns from the Tissu XS kit to accommodate the limited DNA concentrations. We then enzymatically sheared DNA and constructed indexed libraries (both pollen and leaf samples) using the Tecan – NGS Celero NuGen kit using 2 ng of starting material and 8 cycles of PCR. Illumina sequencing of 2x150bp paired-ends fragments was done at GENOSCREEN (Lille, France).

1.3.3 Short Read Alignment and Haplotype Phasing

To detect segregation distortion in the gametes of our F1 progeny, we first identified the group of single nucleotide polymorphisms (SNPs) that come from each ancestral haplotype. To do this, we first aligned all F1 and parental short read data to the MN47 *A. lyrata* reference genome (Hu et al. 2011) and jointly-genotyped each individual using the SNPArcher workflow (C. D. Mirchandani et al. 2024) using the Senteion-optimized versions of BWA and GATK (Freed et al. 2017). This workflow trims adapter sequences in reads resulting in a read length distribution that may contain reads with lengths less than the nominal read length. Below we describe how the read length distributions of the individual forward reads are made identical between pollen and somatic tissue pairs. We retained all aligned reads with a MAPQ of 30 or greater, and removed all SNPs below the 5% and above the 95% depth of coverage quantiles to mitigate against the effects of large structural variants and repeats which might affect allele frequency estimation.

To mitigate the effects of differential and biased mapping of somatic and germline libraries which might result in strongly skewed ancestry ratios, we mapped only the first read in each pair to reduce the potential sources of insert length on mapping properties which could affect the inferred ancestry ratios. Next, we trimmed the reads by order of read length in each leaf-pollen pair such that both samples have identical read length distributions (see (R. Corbett-Detig et al. 2019)) to further reduce the potential consequences of differential mapping on allele frequency biases

due to differences in somatic and germline libraries. We additionally removed any variant sites within 100 bp of each other.

Next, we haplotype-phased each F1 individual using the trio phasing function of WhatsHap (Martin et al. 2016). Trio phasing enables us to effectively phase haplotype in somatic and germline samples by using the maternal and paternal genotype information at variant sites to construct longer and more accurate haplotype blocks. To do this, we used the variant sites obtained from each parent, constructed a mother-father-child pedigree, and performed trio phasing to separate variant sites into maternal and paternal haplotype blocks. Once phased, we filtered any non-identically phased or genotyped variant sites in each leaf-pollen pair, as sites that have differential phasing are likely to be unreliable possibly stemming from a misalignment or structural difference between the reference and one or both haplotypes in a sequenced sample. We additionally performed the entirety of the above mentioned workflow using the *A. halleri* Auby-1 reference genome v1.0 (Pavan et al. 2024), and confirmed that our results are consistent between genomic references (Figure S1). The ability of this approach to phase entire chromosomes via trio-phasing is a critical element of this study because it makes each read within 50 cM of a given site potentially informative for detecting ancestry ratio distortion.

1.3.4 Ancestry Ratio Calculations and Estimation of SD Effect Sizes and Loci Positions

To calculate the ancestry ratios for each sample, we used the intersection of retained ancestry informative sites shared between pollen and leaf samples along each chromosome. At each site, we calculated the ancestry ratio as the ratio of

maternally-derived read counts to paternally-derived read counts. For both leaf and pollen samples, we plotted the average ancestry ratios in 1,000 SNP windows. We calculated and plotted the ancestry ratio difference as the difference between somatic and germline ancestry ratio for each window.

Using a maximum likelihood framework that models sequencing error, recombination rate and ancestry informative sites, we estimated both the position and effect size of each candidate distorter. Additionally, every variant site along the chromosome is informative to the effect size and location of a distorting locus due to linkage. At each ancestry informative site along the chromosome, we calculated the recombinational distance between our site, p, and a candidate distorting locus, i, in Morgans using the *A. lyrata* recombination map from (Hämälä et al. 2017). We then converted the distance to a probability that a recombination event has occurred between the two positions. Considering also the probability of sequencing error, there are four unique probabilities for mapping an allele A, at site p, given i. Conversely, there are four unique probabilities for mapping the alternate allele a, under the same conditions (See (R. Corbett-Detig et al. 2019) for details). Thus, we estimated the likelihood of a given distortion coefficient, k, at distorting locus i, by calculating the likelihood of the mapped allele frequencies across the entire chromosome.

To estimate the effect size at a given site p, we first estimated the likelihood of k at site p in both the somatic and germline data. Next, using the somatic estimate of k as a null model, we estimated the effect size at site p by calculating the likelihood ratio of germline k to somatic k. We then expanded this procedure to all ancestry

informative sites along the chromosome to find the maximum likelihood distorting position. All scripts and programs used for phasing and SD estimation are available from https://github.com/ccondon894/arabidopsis_SD_workflow.

1.3.5 Confidence Interval Estimation

We estimated the uncertainty around a presumed distorter's mapping position by bootstrapping with replacement variant sites along a chromosome and rerunning our maximum likelihood framework. We used the 2.5th and 97.5th percentiles as our confidence intervals after 100 bootstrap replicates. To further improve our confidence interval estimates, we selected a subset of samples for chromosomes 4 and 5 whose individual estimated distorter positions are very close to each other. Under the assumption that they are caused by the same underlying genetic element, we treated each sample as an independent event for the same distorter, and computed the joint likelihoods by taking the sum of log-likelihood among crosses for all bootstrap estimates across the selected samples. This approach significantly decreased the widths of estimated confidence intervals while keeping the joint confidence interval within the range of the individual sample intervals.

1.3.6 Analysis of Segregation Distortion Patterns in Relation to Parental Origin

We performed Fisher's Exact Tests on multiple cross groups classified by parental species or geographical location, including (i) within-region (WR), (ii) between-region (BR), within-species and (iii) between-species (BS) groups. We compared groups to one another and identified groups where there was strong evidence for significantly higher segregation distortion frequency. Using the same

groups as the Fisher's Exact Test, we also tested for significant difference between effect sizes by Mann Whitney-U test.

1.3.7 Genome-Wide Non-Independence Tests

To test for non-independence between distorting loci on separate chromosomes, we hypothesized that the mean directional effect size (k) for each individual across all chromosomes, and the mean across all individuals would be smaller than expected by chance. That is, we expect maternal and paternal distorting alleles to distort in an antiparallel fashion if these effects are consistent with a DMI model. Here, we define distortion in favor of paternal alleles as negative and distortion towards maternal alleles as positive deviations from 0.5. We then constructed a permutation test as follows: (1) Compute the grand mean segregation distortion effect size for all progeny, (2) Randomize the segregation distortion effects across the progeny by conditioning on the number of distorted chromosomes in each individual, (3) Recompute the mean distortion effect for the permuted progeny distribution and record the proportion with a mean effect smaller than or equal to the estimated mean from the true data, (4) Permute 100,000 times.

1.3.8 Pairwise Chromosomal Interaction Tests

We considered whether there were antiparallel interactions between specific pairs of chromosomes consistent across progeny, such that one chromosome distorts towards the maternal genome and the other chromosome distorts toward the paternal genome, or vice versa. We tested this hypothesis with a permutation test as follows: (1) Count the number of antiparallel distortion effects (i.e., one paternal and one maternal

effect) for each chromosome pair across all progeny, (2) Randomly permute the estimated distortion effects across all chromosomes, then recount all chromosome-pair interactions, (3) Repeat for 1,000,000 permutations and record the number of pairs that occur more frequently than the null distribution. To account for the genealogical non-independence caused by some of the crosses sharing particular parents, in these statistical tests we removed crosses with one or more shared parents.

1.3.9 Gene annotations in the chromosomal intervals.

Gene contents of the confidence intervals on chromosome 4 and 5 were retrieved from the annotation of the *A. lyrata* genome assembly (Hu et al. 2011). A list of pollen-related Plant Ontologies (PO) was manually compiled through searches at (https://ontobee.org/ontology/PO) (Dataset S1). PO annotations for the *A. thaliana* genome (po_anatomy_gene_arabidopsis_tair.assoc.gz, po_temporal_gene_arabidopsis_tair.assoc.gz, 2019-07-11) were downloaded from TAIR (https://www.arabidopsis.org/index.jsp). All Arabidopsis gene identifiers associated with pollen-related POs were extracted from both PO annotation lists, and subsequently combined in one list for annotating the gene contents of the chromosome 4 and chromosome 5 confidence intervals (Datasets S2, S3). A complete redundant gene list containing all PO numbers is also provided, with literature references (Dataset S4). Protein-protein interactions were downloaded from TAIR (TairProteinInteraction.20090527.txt, 2019-07-11).

1.4 Results and Discussion

1.4.1 Sequencing and Phasing

We obtained an average of 29 million reads for the parental individuals, corresponding to an expected sequencing depth of 41X and providing substantial power to identify SNPs distinguishing parental genomes. For the F1 samples, we obtained an average of 100 million pollen reads and 98 million leaf reads, corresponding to an expected sequencing depth of 144X, allowing for a precise estimation of the parental contribution at each SNP. After mapping all pollen and leaf sequence data to the *A. lyrata* genome, trio phasing, and filtering inconsistently genotyped or phased sites, we obtained an average of 848,054 retained sites shared between pollen and leaf, with hybrid WR06 retaining a low of 458,629 sites, and hybrid BS15 retaining a high of 1,487,203 sites.

1.4.2 Signatures of Segregation Distortion via Raw Ancestry Ratios

We computed the ancestry ratios across all chromosomes for each hybrid for both somatic and germline read data (Figure 1.2, Figure S1). The ancestry ratio is defined as the ratio of reads from the maternal haplotype versus reads from the paternal haplotype for a given site. The somatic genome sequence data provides a null model, and accordingly across each chromosome, the raw somatic ancestry ratios tend to stay near to 0.5, consistent with our expectations of 50:50 representation of haplotypes in somatic tissues. However, in the majority of F1s, and across numerous chromosomes, we observe signatures of segregation distortion in the pollen samples, with skews in both the maternal and paternal directions (Table 1, Figure 1.2, Figure

S1). From the somatic and germline ancestry ratios, we calculate the ancestry ratio difference between somatic and pollen samples (Figure 1.2). We observe a skew in germline ancestry ratios at least once on every chromosome, and ancestry ratio differences vary dramatically (0–19%).

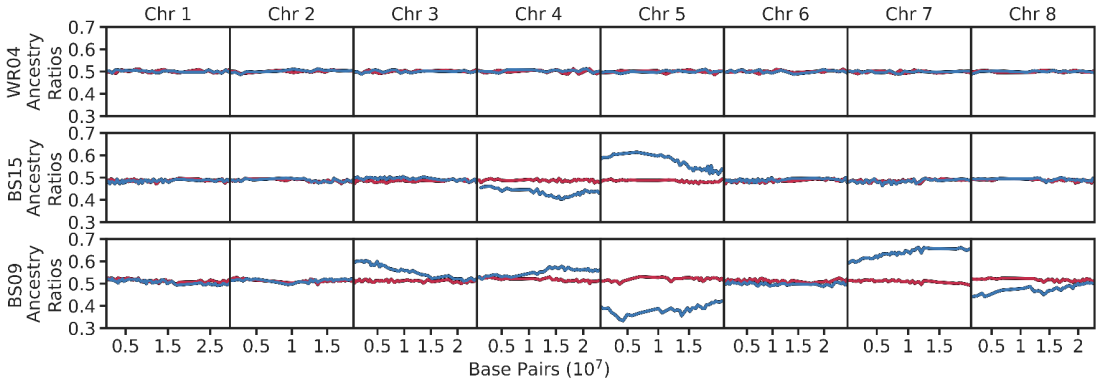


Figure 1.2 Raw ancestry ratios across selected hybrids
Included are WR04 (intraspecific *A. halleri*), BS15 (interspecific) and BS09 (interspecific). Ancestry ratios are plotted from somatic (red) and germline (blue) read data in 1000 SNP non-overlapping windows.

1.4.3 Segregation Distortion Patterns Vary Greatly Amongst Different Hybrid Groups

We detected many instances of pollen-acting segregation distortion in crosses both within and between species. We previously developed a maximum likelihood framework to estimate the effect size and position of a putative distorter locus along a chromosome (Figure 1.3) (R. Corbett-Detig et al. 2019). Using this framework, we estimated both segregation distorter effect size and position along each chromosome for all hybrids, and identified twenty five distorter loci with effect sizes greater than or equal to k = 0.02 (Table 1, Table S2). As expected, many distorter loci appear in interspecific hybrids derived from crosses between *A. lyrata* and *A. halleri* parents,

though the presence of distorter loci are not restricted to interspecific hybrids, and are also observed in several intraspecific *A. lyrata* or *A. halleri* hybrids.

Chromosome-specific distortion patterns vary and depend on the type of cross (Figure 1.2). In intraspecific *halleri,* we observe distortion in just two crosses, BR01 and BR02. This pair shares one parent, Pais09, which is the mother of BR01 and the father of BR02. Both crosses exhibit similar distorter loci on chromosomes 4 and 5, but each distorter skews in the opposite parental direction. Contrary to other documented cases e.g. in flour beetles (Beeman et al. 1992), mouse (Weichenhan et al. 1996), *Caenorhabditis elegans* (Ben-David et al. 2017) or *C. tropicalis* (Noble et al. 2021), the distortion patterns in this pair suggest that these distorters do not exhibit a maternal effect, because the ancestry ratio is associated with male or female contribution. The remaining four *halleri* hybrids do not show distortion (Table 1). Notably, all are derived from crosses whose parents are from nearby geographic locations. Genetic differences between parents in BR01 and BR02 is slightly higher than between those parents in non-distorting *halleri* hybrids ($\pi = 0.0052$ vs. $\pi =$ 0.0043-0.0049, see Table S2). Although they are the same species, our results suggest that the ancestral populations of BR01 and BR02 have accumulated sufficient genetic incompatibilities that they are readily measurable within the power of our experimental design.

Intraspecific *A. lyrata* crosses display distortion patterns distinct from intraspecific *halleri* crosses. All three have a mother from Norway and a father from Germany. Two crosses demonstrate distortion patterns on chromosomes 1 and 3 in

the same parental directions. However, BR12 shows stronger signatures of distortion genome-wide. Chromosomes 1 and 3 have much larger effect sizes, and BR12 also exhibits distortion on chromosomes 6 and 7, while BR10 contains a unique distorter locus on chromosome 5. These differences cannot be explained by parental genetic divergence, since all three have similar levels of genetic differentiation and are derived from the same set of populations (Table S2). Instead, these results demonstrate a level of variability in the presence or penetrance of distorting loci. This variability could be caused by the presence of modifiers across the genome modulating expression of the distorters, by allelic variability of the elements causing the distortion in the different individuals, or by other sources of individual variation in phenotypic expression, such as environmental effects.

The number of distorter loci and effect sizes in interspecific crosses exceeds those found in intraspecific crosses. Of the 29 total distorter loci we identified, 18 are in interspecific hybrids, and a distorter locus appears at least once on every chromosome within this group. Effect sizes vary substantially within this group (range $|k| = 0.02–0.19$). We replicated previously observed distortion signals in hybrids BS15 and BS16 (R. Corbett-Detig et al. 2019). Additionally, here we report two previously undetected distorter loci on chromosomes 1 and 2 in BS16. These additional loci could be explained by the increased statistical power we gained from trio phasing (which is new as compared to our previous analysis), which almost doubled the number of ancestry informative sites retained. BS14 and BS08 show modest signatures of segregation distortion, with only two loci identified in each

hybrid and an effect size of |k| = 0.02. BS09 however, exhibits distorter loci on six of its eight chromosomes, and two loci display effect sizes of |k| = 0.19, by far the most distorter loci and largest effect sizes of all the crosses. Notably, BS08 and BS09 share a parent, Hall2.2, yet show very contrasting distortion patterns. This reaffirms our observations that not only is segregation distortion evolving rapidly, but occurrences and effect sizes vary substantially.

Table 1: Table of all F1 hybrids included in this study.
Included within this table are distortion effect sizes within each chromosome. Effect sizes are colored dependent on their effect direction towards either the maternal (red) or paternal (blue) genome. Additional information about the parents and populations contributing to each cross are available in Table S2.

Cross Type	Cross ID	Estimated Distortion Effect Size (k)							
		Chr1	Chr2	Ch3	Chr4	Chr5	Chr6	Chr7	Chr8
Within Region	WR04	0	0	0	0	0	0	0	0
	WR05	0	0	0	0	0	0	0	0
	WR06	0	0	0	0	0	0	0	0
	WR07	0	0	0	0	0	0	0	0
Between Region	BR01	0	0	0	-0.07	0.06	0	0	0
	BR02	0	0	0	0.07	-0.05	0	0	0
	BR12	-0.05	0	0.1	0	0	0.04	-0.02	0
	BR10	-0.02	0	0.02	0	-0.02	0	0	0
	BR11	0	0	0	0	0	0	0	0
Between Species	BS09	0	0	0.1	0.06	-0.19	-0.02	0.19	-0.08
	BS08	0	0	-0.02	0	0	0.02	0	0
	BS15	0	0	0.02	-0.08	0.13	0	0	0
	BS16	0.02	0.02	0.03	-0.05	0.06	0	0	0
	BS14	-0.02	0	0	-0.02	0	0	0	0

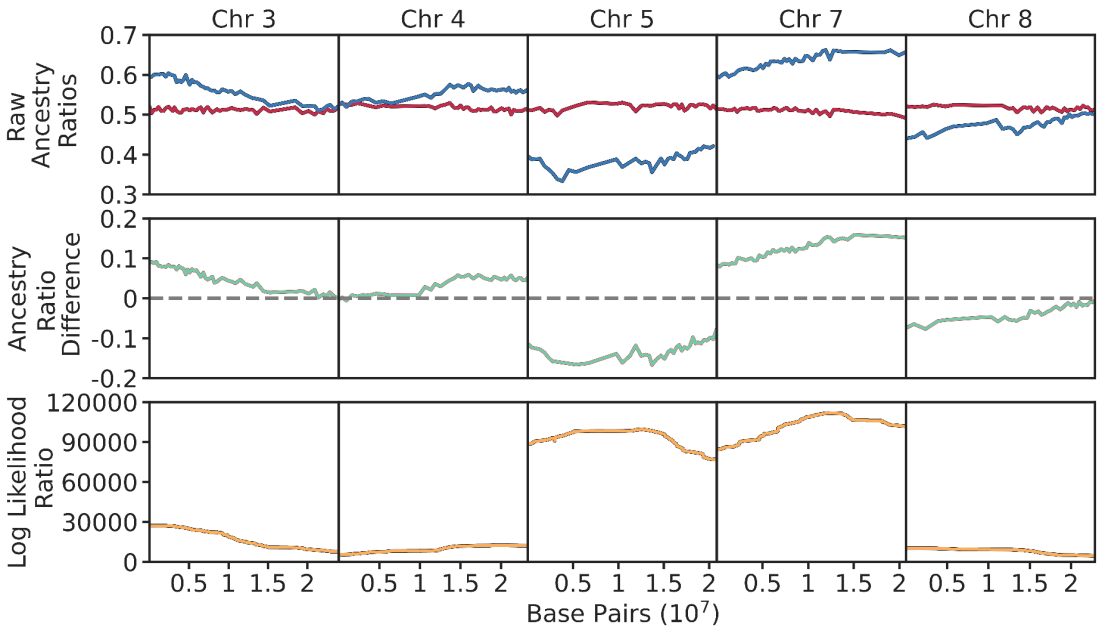


Figure 1.3 Ancestry-ratio differences and estimated distorter positions in hybrid BS09
Raw ratios are plotted from somatic (red) and germline (blue) read data in 1000 SNP non-overlapping windows. Differences between somatic and germline ancestry ratios are colored green. The dashed gray horizontal line indicates the expected difference, zero, when both ratios are equal. The estimated positions of distorting loci along each chromosome are colored orange.

1.4.4 Estimated Segregation Distortion Frequency and Effect Size Increase with Genetic Divergence Between Parents

Regardless of the proximal mechanism, a prediction is that segregation distortion should be more frequent in crosses with more genetically divergent parents. To formally test this, we grouped our hybrid progeny into three groups: within-region (n=4), between-regions (n=5), and between-species (n=5)(Figure 1.4a) . Collectively, within-region and between-regions also comprise the “within-species” group. Segregation distortion does not occur in any of the within-region crosses. Segregation distortion is apparent in both between-regions groups and between-species groups. When we compare occurrences between within-species and between-species groups,

we find that there are significantly more occurrences in the between-species group (Fisher's Exact Test, P=0.0013). While the difference between the between-species and between-region groups is not significant (FET, P=0.16), the odds ratio of 2.15 suggests that segregation distortion occurs more frequently in the between-species group. Overall, the proportion of distorting chromosomes increases with parental genetic divergence.

A second prediction is that effect sizes (*i.e.,* the proportion of affected pollen) should become larger as parental genetic divergence increases. While comparing between-species to within-species groups (Figure 1.4b), we find the effect sizes are significantly higher (Mann-Whitney U, P=0.0013). There is no significant difference in effect size when comparing between-species to between-region groups (Mann Whitney U, P=0.10), but the mean distortion effect size observed in between-species crosses is larger. Overall, these results demonstrate that distortion patterns are not limited to genomic differences between *A. lyrata* and *A. halleri*, because we observe them in many between-region crosses. Hence, parental genetic distance is a predictor of the distribution and abundance of distortion effects.

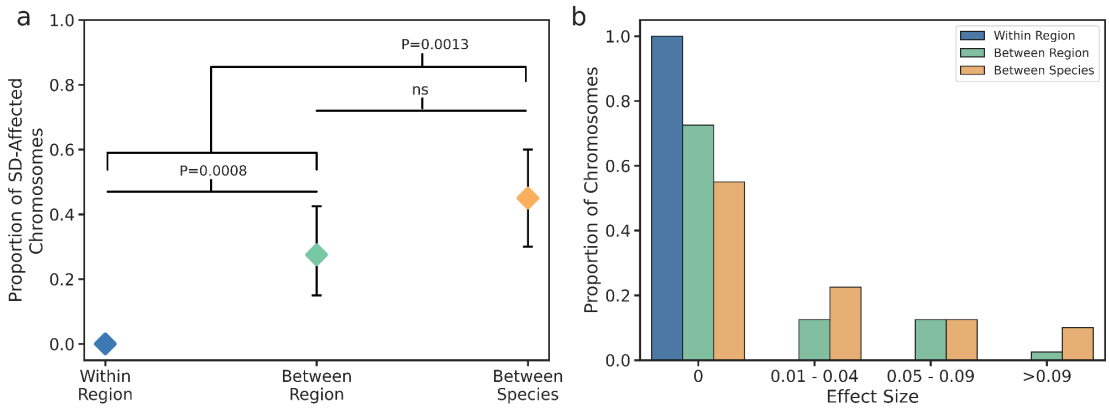


Figure 1.4 Segregation-distortion frequency and effect size across cross groups
(a) Proportion of segregation distortion presence/absences in chromosomes sorted by cross group. (b) Proportion of segregation distortion effect sizes, sorted by cross group and effect size cluster.

1.4.5 Evidence for Non-Independence Among Loci

If negative interactions between parental genomes (DMI) are the source of most segregation distortion, we expect that the effects of maternal and paternal alleles will often be paired by an apparent distortion effect from the opposite parental genome. In contrast, if selfish elements are causing the distortion, we expect independent or directional distortion effects (see Figure 1.1). To distinguish between these two possibilities, we first computed the sum of k for all detected sites within each progeny's genome, $\sum_{i=1}^{n}(k_{\{crossID\}})$. We then tested whether the mean of all such sums, was closer to 0 than we expect by chance (as expected under the DMI model, $\mu\left(\sum_{i=1}^{n}(k_{\{crossID\}})\right)$). In evaluating this hypothesis, we found that there is a significant correlation between maternal-paternal directional effect size such that the sum of all significant effects in a cross is closer to zero than expected by chance (P = 0.001,

Permutation Test, Figure 1.5a). This suggests that interactions among alleles contributed by the maternal and paternal genomes are a major contributor to pollen inviability.

Next, we considered whether there were non-independent, antiparallel interactions between specific pairs of chromosomes consistent across progeny. We found that the pair of chromosomes 4 and 5 show highly significant non-independence across the set of crosses (Bonferroni adj., $P=5x10^{-5}$, Permutation Test, Figure 1.5b), suggesting a genetic interaction between this specific pair of chromosomes. This further supports the idea that interactions between alleles contributed by each parental genomes is the major contributor to the pollen-acting segregation distortion. We note that this test is distinct from the more typical interchromosomal LD which may be revealed in F2 populations of crosses, and instead tests for correlated biased transmission effects across individuals rather than across progeny from a single experiment. This test therefore relies on the presence of identical or similar distorting loci in individuals. Further, to fully take into account phylogenetic non-independence in these tests (such as *e.g.* the fact that two *A. halleri* individuals are more likely to contain the same distorting allele rather than each being an independent evolutionary process generating the observation) would require experimental crosses to control in a more precise manner the genealogy of the distorting chromosomes being observed.

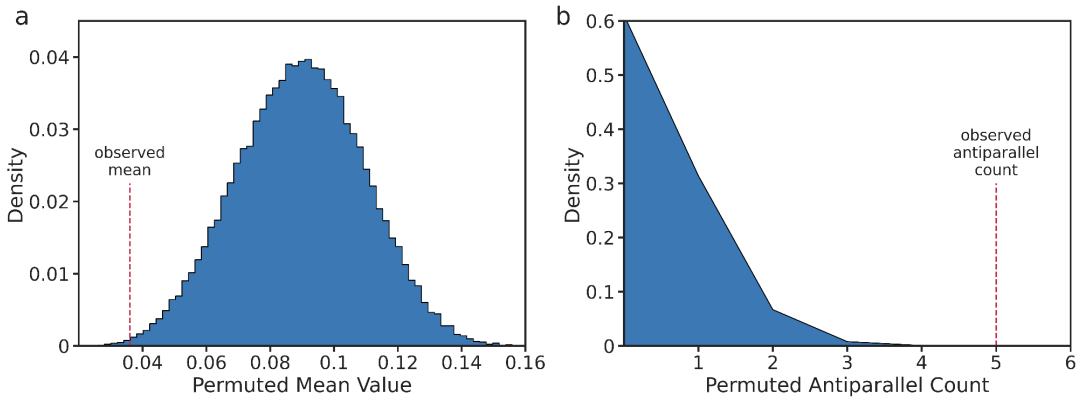


Figure 1.5 Tests of genome-wide and chromosome-specific non-independence
(a) The distribution of the mean of effect size sums for 100,000 permutations testing for non-independence across all F1 progeny. The original mean of effect size sums across F1 progeny is indicated by the dashed red line. (b) The distribution of antiparallel interaction occurrences between chromosomes 4 and 5 for 1,000,000 permutations. The original antiparallel count is indicated by the dashed red line.

1.4.6 Confidence Interval Estimation Around Distorting Loci on Chromosomes 4 and 5

Next we estimated the uncertainty around a putative distorter's mapping position by bootstrapping, as in (R. Corbett-Detig et al. 2019). To improve on confidence interval estimates from individual crosses, we selected a subset of hybrid samples for chromosomes 4 and 5 whose individual estimated distorter positions are close enough that they could presumably correspond to the same locus. Following this logic, we proposed that each sample could be treated as an independent event for the same distorter, and thus computed the joint likelihoods for all bootstraps across the selected samples (Figure 1.6). This approach greatly improved the confidence interval estimates compared to the analysis based on single crosses, decreasing the window size while also keeping the joint confidence interval within the range of the individual sample intervals (more so for chromosome 4 than chromosome 5). The final joint confidence interval sizes for chromosomes 4 and 5 were 2.14 Mbp

(17,065,274 – 19,205,774 bp) and 610 kbp (4,032,300 – 4,643,400 bp), respectively, on the *A. lyrata* reference genome. We acknowledge that this approach is only valid assuming that distorters in close proximity to each other are indeed the same distorting allele and that no secondary distorters are present. Thus, these joint confidence intervals should be interpreted cautiously, and any further examination of these distorter positions would require functional follow-up work.

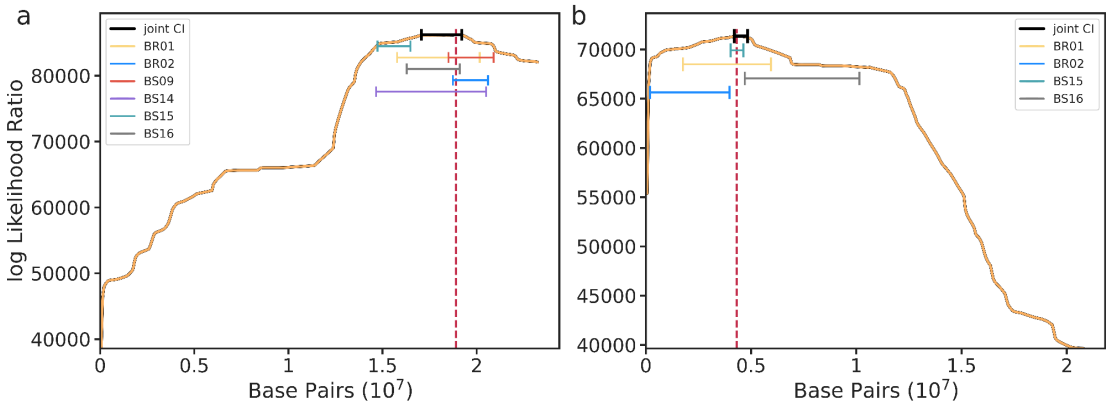


Figure 1.6 Confidence intervals for distorting loci on chromosomes 4 and 5
The solid orange line indicates the estimated likelihood of the distorting locus along the chromosome. The dashed red line indicates the joint maximum-likelihood position and solid black bars indicate the joint confidence interval. Colored bars indicate the individually estimated sample confidence intervals.

1.4.7 Candidate Segregation Distortion Genes

The joint chromosome 4 and chromosome 5 confidence intervals were predicted to contain 411 and 61 genes, respectively in the *A. lyrata* reference genome (Datasets S2, S3) (Hu et al. 2011). Among them, 216 and 26 had *A. thaliana* orthologs annotated as pollen-related in the Plant Ontology (PO) database (www.plantontology.org, see list of PO terms in Dataset S1). The predicted functions of a number of these genes could be relevant in the context of the observed

segregation distortion phenotype, such as genes encoding F-box proteins, proteases, proteins contributing to transcription and translation, cell division, nutrient and metabolite transport, endomembrane trafficking and signaling. For instance, the chromosome 5 interval contains a gene encoding a Concanavalin A-like lectin-type protein kinase family protein (AL5G17300) in syntenic position with At5G65600 (AtLecRK-IX.2). AtLEcRK-IX.2 was characterized as a regulator of cell death and a positive regulator of pathogen recognition receptor-triggered immunity in Arabidopsis acting to phosphorylate RbohD and thus triggering ROS production (Luo et al. 2017; Y. Wang et al. 2015). A gene within the chromosome 4 confidence interval (AL4G34480) is a non-syntenic homologue of AT3G56440 encoding AtATG18D, an uncharacterized protein related to yeast autophagy 18 (ATG18) D (Xiong et al. 2005). These are examples for which an alteration of the activity of gene products could trigger cell death during pollen development. None of the gene products in the chromosome 4 interval were previously identified to establish direct protein-protein interactions with the products of any of the genes in the chromosome 5 interval according to the protein-protein interactions available through TAIR, and none of the genes in the two intervals had previously been identified as being required for male gametophyte development (Muralla et al. 2011). We acknowledge that the sheer number of possible direct or indirect molecular interactions between any of these genes currently makes it difficult to formally pinpoint specific pairs of genes more precisely and that one or both causal genes might be absent in the confidence intervals of the reference genome used here. Hence, our method serves as an efficient

way to narrow down a segregation distortion locus to a limited chromosomal interval, but fine-mapping will now be necessary to further restrict the list of candidate genes.

1.5 Conclusion

Our results demonstrate that segregation distortion is widespread and rapidly evolving not only between different *Arabidopsis* species, but within species as well. A rapid spread of segregation distorters has been observed in several instances (Seymour et al. 2019), but is often only transient and followed by the rapid evolution of suppression (reviewed in (Price et al. 2019)). Here we find that segregation distortion is highly variable among individuals, where within-region progeny show no distortion, but all other progeny display frequent distortion and variable effect size. Intriguingly, our observation that neither the number nor effect size of distorters increases substantially with divergence time challenges conventional models of epistasis in hybrid fitness (Dagilis et al. 2019). This finding may also offer a more nuanced perspective on snowball theory, which predicts an increase in interacting loci with divergence, though we acknowledge that our study primarily detected pairwise interactions and may be underpowered to identify the higher-order interactions that have been rarely documented in other systems (Moran et al. 2024). This approach is straightforward to apply to many other species, and expanding the analysis to more species will now allow to determine the generality of this conclusion. An especially interesting comparison would be to analyze hybrids between selfing vs outcrossing lineages, as differences in the mating system are expected to lead to differences in the intensity of genetic conflicts (Burt and Trivers 1998; Mantel and Sweigart 2024). We

expect that the dynamics of driver-suppressors should be more apparent in such circumstances.

More generally, our results demonstrate that segregation distortion acts early during population divergence and may be a contributor towards the common pattern of hybrid infertility contributing to speciation. We cannot completely rule out selfish genetic elements as potential distorters, as our observations could also be attributed to select driver elements that have spread in their own populations. Even if distorters ultimately crossed population and species boundaries, they could temporarily drive reproductive isolation. Pollen viability data would also help link the observed distortion patterns to an expected phenotype under a DMI, although we cannot rule out that a DMI could act sufficiently early in gametogenesis that its effect would be unobservable in pollen grain viability measurements. The pattern of segregation distortion observed in our data, where skews in parental directions tend to compensate each other both overall and at specific loci, is more consistent with negative epistatic interactions between pairs of loci, where the viability of the gametes that contain a specific combination of alleles is impaired. We note again that this pattern is limited by the phylogenetic non-independence of our study, but this can be overcome in future studies with experimental crosses that more precisely handle the genealogy of distorting chromosome pairs. In particular, it will be exciting to determine how different groups of organisms accumulate such incompatibilities, especially since the "speciation clock" has been suggested to tick at different paces across the tree of life (e.g. in plants vs animals, (Monnet et al. 2023)). In addition,

here we focused on the male germline, but the comparison with the female germline would be necessary to provide a comprehensive picture.

[The following is an adaptation of a publication which has been published in PLOS Biology: https://doi.org/10.1371/journal.pbio.3003823]

Chapter II: Splicing deficiency is driven by genomic erosion in non-recombining algal mating type chromosomes

2.1 Abstract

Splicing deficiency may represent a critical yet underexplored form of genomic erosion in non-recombining regions. Across four phytoplankton species diverged ~333–639 million years ago, genes within U (female) and V (male) "UV" mating-type regions—non-recombining chromosomal regions that determine mating compatibility—show strikingly elevated intron retention relative to genes in other genomic regions. Long-read data reveal abundant aberrant, likely non-functional mRNA isoforms despite preserved coding potential. This preservation suggests that splicing defects arose early in UV evolution and have persisted over deep time. We propose that these defects arise from evolutionary changes in sequence composition and chromatin organization that accompany recombination suppression, such as reduced GC content, altered nucleosome occupancy, and disrupted methylation, that collectively compromise splicing fidelity. Unlike sex chromosomes, which often degenerate through gene loss, splicing-deficient UV regions in green algae retain hundreds of genes, indicating that transcript-level dysfunction provides an alternative route to functional decay. Our results identify chromatin-mediated splicing deficiency as a novel axis of genomic erosion and position algal UV systems as models for studying how recombination suppression reshapes RNA processing fidelity in essential, non-recombining genomes.

2.2 Introduction

In eukaryotes, suppressed recombination drives predictable genomic decay including disrupted gene order, repetitive element expansion, and biased nucleotide

composition (Lahn et al. 2001; Bachtrog 2013). Many studies focus on highly degenerate regions like mammalian Y chromosomes, where extensive gene loss reflects relaxed selection on non-essential genes (*e.g.,* (Lahn and Page 1999)). However, many non-recombining regions must retain essential functions—such as mating-type loci in fungi and algae, centromeric regions, and loci harboring large segregating structural rearrangements—creating evolutionary tension between genomic decay processes and strong functional constraints (Coelho et al. 2018; Bull 1978). Understanding how this tension resolves has broad implications for genome evolution and for identifying the factors that drive the emergence of genetic innovation.

UV mating-type (MT) regions in the green algae offer a compelling model for studying genomic erosion with selective constraints to maintain genetic function. These largely non-recombining regions span 393-1,681 kb across the four Mamiellales species examined here, and contain 216-619 genes (Benites et al. 2021). The boundaries of these regions are defined by sharp transitions in GC content, which is a hallmark molecular signature of recombination suppression resulting in the loss of GC-biased gene conversion (Benites et al. 2021). Despite sharing 23 core gene families that define orthologous UV loci across Mamiellales, these regions show substantial species-specific gene content and structural divergence. Within O. tauri, the MT- and MT+ alleles differ in size (~650kb vs. ~450 kb) and contain both mating-type-specific genes and 75-79 shared genes beyond the core gene families.

Like non-recombining sex chromosomes, UV loci show disrupted synteny, lower gene density, expanded repeat content, decreased GC nucleotide content (Coelho et al. 2018; Benites et al. 2021; Worden et al. 2009), and more open chromatin (Valiadi et al. 2025) compared to autosomes. Notably, UV loci in *Mamiellales* also show higher gene expression than the autosomal regions (Worden et al. 2009; Valiadi et al. 2025). Comparative genomic analysis provides evidence for past translocations in O. tauri, where 46 gene families in MT- and 30 in MT+ have paralogs located on diverse autosomes (Benites et al. 2021). Unlike the usually highly degenerate Y or W chromosomes, UV loci encode hundreds of actively expressed genes with critical biological functions (Worden et al. 2009), suggesting that degeneration is tempered by selective pressures and may manifest in fundamentally different ways from more commonly studied sex chromosomes.

We hypothesize that transcript processing dysfunction represents a cryptic form of erosion in functionally constrained regions. RNA splicing fidelity depends on sequence context including both consensus splice sites and auxiliary splicing enhancers and silencers (Bénitière et al. 2024; Fu and Ares 2014; Sheth et al. 2006), chromatin structure (Jimeno-González et al. 2015), and DNA methylation patterns (Davey et al. 2004; Naftelberg et al. 2015). Recent comparative studies demonstrate that species with smaller effective population sizes accumulate more rare, likely non-functional alternatively spliced transcripts, suggesting that reduced efficacy of natural selection permits aberrant splicing (Bénitière et al. 2024). If suppressed recombination similarly degrades splicing while preserving coding sequences, this

could diminish gene function without complete gene loss—a form of "molecular erosion" largely invisible to traditional genomic analyses.

Using comparative transcriptomics across four Mamiellales species, we find widespread splicing dysfunction within recombination-suppressed mating-type regions. Long-read isoform sequencing further demonstrates that this dysfunction produces an excess of aberrant, likely non-functional transcripts, providing evidence for transcript-level genomic erosion. Our results reveal transcript-level defects as a novel axis of genomic decay and establish the mating-type regions of Mamiellales as a model for studying transcriptional dysregulation.

2.3 Methods

2.3.1 Software Development

We used LLM support to develop initial scripts and analyses. Each script was tested and ultimately validated by the authors prior to inclusion in this manuscript.

2.3.2 Comparative Intron Retention Analyses

We identified publicly available short-read RNAseq data for four species in the Mamiellales family through metadata search on the sequence read archive (Table S1). We further restricted the search to species whose genomes are chromosome level assemblies with a mating-type region defined in (Benites et al. 2021). We aligned RNAseq data to each genome using STAR (Dobin et al. 2013) version 2.7.10b and provided the appropriate gtf file during index construction, and the "--quantMode GeneCounts" option to obtain estimates of gene expression, but used otherwise default parameters. We used rMATs (Shen et al. 2014) version 4.3.0 to quantify intron

retention providing the parameters "--allow-clipping --novelSS" but otherwise default in running this analysis.

We fit a linear mixed-effects model of intron retention ratios, using gene ID as a random effect, to estimate the variance in intron retention explained by differences among genes and to calculate the intra-gene correlation.

2.3.3 Branchpoint-Sequence Detection

We extracted introns from annotated CDS features in GTF files and retrieved the corresponding 50-nt upstream branchpoint windows from the reference genome, outputting unique sequences in FASTA format for downstream analyses. These sequences were then analyzed with DREME (Bailey 2011) to identify enriched short motifs. In each case, the most significantly enriched motif is consistent with expected branchpoint sequences in these species.

2.3.4 Methylation and Chromatin Accessibility Data

For *M. pusilla*, we complimented gene expression estimates and intron retention rates with MNase coverage and CG methylation fractions obtained from (Huff and Zilberman 2014) under Gene Expression Omnibus accessions "GSM1134620" and "GSM1134621", respectively.

2.3.5 Random Forest Modelling

We trained a Random Forest classifier to predict intron retention events using a comprehensive set of genomic and transcriptomic features. Because intron retention ratios are highly skewed and difficult to model accurately using regression, we

framed the problem as a binary classification task. The classification target was defined as introns with retention ratios ≥0.025 (misspliced) versus <0.025 (properly spliced). Features included transcription levels (TPM), intron length, GC content, distance from the transcription start site, canonical splice site identity, frame disruption status, and one-hot encoded nucleotide sequences flanking the 5′ and 3′ splice sites (excluding the canonical dinucleotide positions). To prevent data leakage, we performed gene-level splitting, randomly partitioning genes into 80% training and 20% test sets such that all introns from a given gene were assigned to the same partition.

We optimized Random Forest hyperparameters (number of trees, maximum depth, minimum samples per split/leaf, and maximum features per split) using randomized search with 30 iterations and 3-fold grouped cross-validation on the training set, using ROC-AUC as the optimization metric. Class imbalance was addressed using balanced class weights. To identify a minimal feature set without sacrificing predictive performance, we implemented an automated feature selection procedure. We ranked features by their importance scores from the full model, then incrementally added in order of importance while tracking cumulative ROC-AUC on the held-out test set. We determined the optimal feature set using elbow detection via the Kneed algorithm. We produced partial dependence plots to visualize the marginal effect of each selected feature in the resulting minimal models on predicted intron retention probability. We also trained Random Forest models on autosomal introns

using the same procedure. The extreme rarity of retention events in autosomal regions prevented effective model training, yielding low F1 scores (Table S4).

2.3.6 M. pusilla Culture

The isolate of *M. pusilla* used here for RNA and chromatin analysis was obtained from the Roscoff Culture Collection. RCC834 is the same as CCMP1545, initially sequenced by Worden et al (Worden et al. 2009) and widely used as a reference for this species.

We grew cultures in low salinity (0.8X) artificial seawater with L1 supplements (except that Na_2SiO_3 was not added), (Enevoldsen and Unesco 2003) made using the L1 kit purchased from the National Center for Marine Algae/Bigelow (cat. No. MKL150L). We maintained small (10ml) cultures in vented T25 flasks at 18°C under a 14h light:10 h dark illumination cycle at 220 microEinsteins (uE).

To obtain cell pellets for RNA extraction, we transferred approximately 2 mL of the RCC834 culture into eight separate wells and maintained them in the above conditions for two days. On the third day, at 3 hours into the light cycle (1020 lux, 14 μE), we removed and pelleted each sample by centrifugation for 4 minutes at 5,000 g and 4°C, immediately flash froze the pellets in a dry ice-ethanol bath, and stored them at -80°C.

2.3.7 Full-length Isoform Sequencing with R2C2 and Analysis with Mandalorion

We extracted RNA from approximately 1E8 cells of *M. pusilla* strain RCC834 (also known as CCMP1545) using a mixture of Trizol and chloroform, followed by

purification with an RNeasy extraction kit. We prepared cDNA libraries through two sequential reactions. First, we performed reverse transcription on 200 ng of total RNA using Smartscribe reverse transcriptase, FS buffer, DTT, indexed oligo-dT primers, and a template switching oligo (TSO), incubating at 72°C for 3 minutes and 42°C for 60 minutes. Next, we amplified the resulting cDNA by PCR using KAPA HotStart master mix with primers targeting the oligo-dT and TSO sequences.

We circularized the cDNA libraries via Gibson assembly (NEBuilder HiFi) using a short, pre-annealed double-stranded DNA splint overlapping the cDNA ends. To remove uncircularized molecules, we digested the reaction with ExoI, ExoII, and Lambda Exonuclease (all NEB) for 16 hours at 37°C, then heat-inactivated the enzymes for 20 minutes at 80°C. We cleaned the reaction with SPRI beads at a 1:0.85 ratio.

We used the clean, circularized library as a template for rolling circle amplification (RCA) with Phi29 polymerase (NEB) and a random hexamer primer for 18 hours at 30°C, followed by heat inactivation for 10 minutes at 65°C. We then debranched the Phi29 reaction using T7 endonuclease for 2 hours at 37°C and cleaned and concentrated the product with a NEB Monarch DNA clean and concentrator column. We size-selected the resulting R2C2 library for DNA fragments larger than 3 kb using ProNex beads (Promega) and quantified DNA concentration with a Qubit fluorometer.

Table S6 lists the oligonucleotides used for library construction and circularization. We sequenced the multiplexed R2C2 library on an ONT PromethION

R10.4 flow cell and basecalled reads with the Dorado basecaller using model dna_r10.4.1_e8.2_400bps_sup@v5.0.0.

The resulting raw reads were converted into consensus reads and demultiplexed using C3POa (v3.2). Based on these demultiplexed consensus reads, Mandalorion (v4.6) was used to identify and quantify high-confidence isoforms.

We refined the initial Mandalorion isoform calls using "sqanti3_qc" (Tardaguila et al. 2018; Pardo-Palacios et al. 2024) using default parameters with the addition of "--isoAnnotLite" to attempt to liftover CDS features from the original gff3 file.

2.3.8 Junction-based Alternative Splicing Classification for R2C2 Isoforms

For each *M. pusilla* gene with at least two R2C2 isoforms, we classified alternative splicing events by comparing junction coordinates across isoforms: alternative 5' splice sites (A5SS) and 3' splice sites (A3SS) were identified as junctions sharing one boundary but differing at the other, exon skipping (SE) as junctions in one isoform spanning an exon present in another, and retained introns (RI) as junctions fully contained within an exon of another isoform. This analysis was restricted to the Mandalorion/SQANTI3-filtered isoform set. Per-region counts are provided in Table S8.

2.3.9 Isoform Usage Analyses

We quantified isoform usage for each gene using Evenness, defined as the Shannon diversity index of transcript usage normalized by the maximum possible

Shannon diversity for the observed number of isoforms. We fit linear models with log-transformed gene expression (log(TPM+1)) and intron number as covariates, and we included mating-type region status (inside vs. outside) as the main predictor. We tested the significance of each predictor, and the coefficient for mating-type region was highly significant, indicating that isoform usage is more even for genes inside the mating-type region after correcting for expression and intron number.

2.3.10 Disruptions to protein function

We quantified possible effects on protein function in two ways. First, we compared the length of CDS annotated for each isoform with the CDS length in the canonical gene annotation. The expectation is that shorter CDS are less likely to be functional. Second, we first used interproscan (version 5.59-91.0) (Blum et al. 2021; Jones et al. 2014) with the flag "-goterms" to annotate functional domains and ontologies in all *M. pusilla* proteins. Then, to assess the effect of isoforms on protein function, we compared the proportion of isoforms that contained *X* percent of the total functional domains present for each gene across multiple values of *X*. We performed the same analysis for gene ontology terms. We then compared these results between genes in the MT region and the rest of the genome.

2.3.11 Logistic regression for internal priming

To test whether oligo-dT mispriming on AT-rich sequences could explain elevated intron retention in the mating-type region, we fit a logistic regression of the form P(has_IR) ~ perc_A + gene_id, where perc_A is the percent adenine in the 20

bp genomic window downstream of each isoform's transcript termination site and gene_id is a categorical fixed effect. The model was restricted to the 164 mating-type genes containing both IR and non-IR isoforms (n = 1,393 isoforms).

2.4 Results

2.4.1 Splicing In Green Algae Mating-Type Region is Highly Aberrant

Transcriptome analysis among four Mamiellales species revealed widespread and apparently conserved splicing misregulation within the mating-type region (Figure 2.1). We compared splicing patterns between previously identified mating-type regions to other regions in four Mamiellales species (Benites et al. 2021). In *Micromonas pusilla* the mean intron retention frequency outside of the mating-type region was 0.011, while within the mating-type region the mean frequency was 0.33 ($p < 0.0001$, Mann-Whitney U). We estimate that nearly half (45%) of all transcribed introns derived from genes within the mating-type region are retained in these data. This difference (33% vs 45%) indicates that the frequency of intron retention is more common in transcripts of highly expressed genes (see also below). Despite approximately 339-630 million years since the last common ancestor of these taxa (Parfrey et al. 2011), this pattern is qualitatively consistent across each species (Figure 2.1, Table S1), implying that the processes that contribute to this apparent aberrant splicing have been operating since shortly after the emergence of the UV mating-type region and have been retained in these diverging lineages.

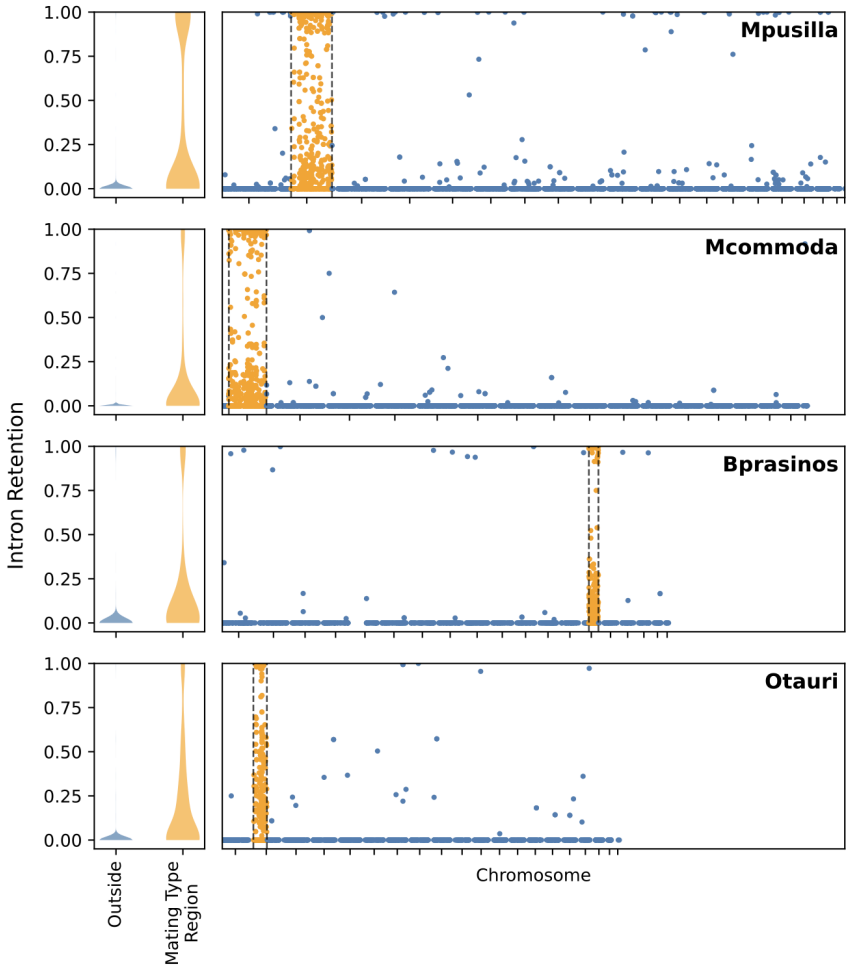


Figure 2.1 Genome-wide intron-retention patterns across Mamiellales (Left), per intron retention frequency along autosomal regions (blue) and within the mating-type region (orange). (Right) Retention frequency for each intron plotted along each chromosome with colors as on the left panel. We obtained coordinates for mating type regions from (Benites et al. 2021). Tick-marks indicate the midpoints of chromosomes in each genome assembly and we excluded any contig below 50Kbp. Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/{species}.features.branchpoints.tsv).

2.4.2 Possible Mechanistic Basis of Aberrant Splicing

Introns in mating-type genes are highly divergent from those elsewhere in the genome. They are significantly shorter with lower GC content (Figure S1), both features that affect splicing efficiency (Maayan Amit, Maya Donyo, Dror Hollander,

Amir Goren, Eddo Kim, Sahar Gelfman, Galit Lev-Maor, David Burstein, Schraga Schwartz, Benny Postolsky, Tal Pupko, Gil Ast 2012; Sterner et al. 1996). Most strikingly, branchpoint sequences which are critical for spliceosome assembly, are significantly depleted in all species within the mating-type region ($p < 0.0001$ for all species; Table S3). Splice-site flanking sequences also reflect the broader GC depletion in the mating-type region, though specific positions within the extended splice site consensus motifs retain their nucleotide compositions (Figure 2.2, Figure S2). Splice-site flanking sequences reflect the broader GC depletion in the mating-type region, though specific positions within the extended splice site consensus motifs retain their initial nucleotide compositions (Figure 2.2, Figure S2). For example, guanine at the fifth intron position of the 5' splice site—which contributes to splicing efficiency by base pairing to U6 snRNA (Carmel et al. 2004; Wassarman and Steitz 1993; Sawa and Abelson 1992)—is significantly less frequent in mating-type introns (8%-23% reduction; $p < 0.0001$ for all species, Fisher's exact test; Table S2), yet these sites remain enriched for guanine relative to other intronic positions, suggesting that natural selection is relaxed but detectably present. In contrast, canonical splice sites (5' GT/GC and 3' AG) occur at similar frequencies inside versus outside the mating-type region (Figure S2), with only two subtle exceptions: *M. commoda* shows slightly fewer canonical 3' splice sites in mating-type genes (96.1% vs. 98.6%, $p < 0.0001$), while *M. pusilla* shows slightly more canonical 5' splice sites (96.6% vs. 98.2%, p=0.41E=3). These results indicate that while mating-type introns differ markedly in composition and length, natural selection

maintains core splicing signals but acts less efficiently on secondary elements that affect splicing efficiency.

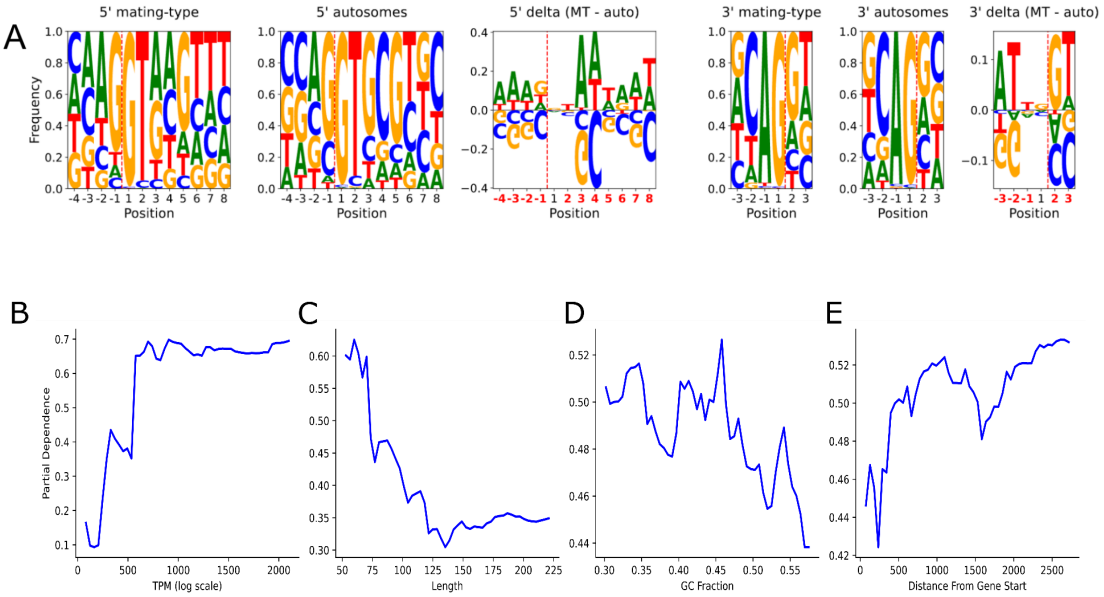


Figure 2.2 Splicing-related features in M. pusilla
(A) 5’ and 3’ splice sites in the mating-type region, 5’ and 3’ splice-sites in autosomal regions, and the difference in sequence composition between mating-type and autosomal regions’. Intron boundaries are denoted with a dashed vertical line. Partial dependence plots for (B) TPM, (C) intron length, (D) GC fraction, and (E) intron distance from the gene start. The Y-axis shows the marginal effect of each feature on predicted intron retention probability, holding all other features at their average values. Higher values indicate increased likelihood of retention. Feature importances are provided in Table S4. Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (files: data/Mpusilla.features.branchpoints.tsv, rf/{species}_MT_importances.tsv).

Random forest modeling considering only introns within the mating-type loci confirmed that shared transcript and genomic features predict the occurrence of splicing dysfunction across all four species (AUC-ROC: 0.735–0.867). Transcripts per million, a measure of gene expression, was the strongest predictor, wherein highly expressed genes showed the highest intron retention rates across all species. Shorter introns, lower GC content, and greater distance between the intron and the

transcription start sites also predicted retention (Figure 2.2, Table S4), while potential functional consequences (frameshifts, stop codons) and the presence of specific splicing features (branchpoints and splice-site adjacent nucleotides) showed little predictive power and were rarely retained in final models after applying minimal feature selection techniques. In contrast, Random Forest models trained on autosomal introns failed to achieve meaningful predictive performance (F1=0.00-0.17) due to extreme class imbalance where retained introns comprise only 0.6-2.6% of autosomal introns (Table S4). The absence of predictive power for frameshifts or stop codons may reflect active nonsense-mediated decay (NMD). Core NMD machinery (UPF1, UPF2)(Kurosaki et al. 2019) is expressed across all species (Table S7) and transcripts with retention induced frameshifts or stop codons would be primary degradation targets. Our measurements therefore likely capture the subset of retained introns that saturate or escape NMD. Notably, 16–35% of variance in intron retention occurred at the gene level across all species, indicating coordinated splicing regulation within genes rather than independent intron-by-intron effects. Consistent with this coordination, intron number per gene was strongly correlated with retention rate in mating-type genes (Spearman's $\rho = 0.56–0.65$, $p < 0.0001$) but not in autosomal genes ($\rho = 0.06–0.22$), suggesting that splicing defects compound across multiple introns specifically in the mating-type region.

The mating-type region of *M. pusilla* is substantially different from the other genomic regions with respect to the local chromatin environment (Valiadi et al. 2025). MNase-based measurement of chromatin accessibility (Huff and Zilberman 2014)

suggests that genes residing in the mating-type region have substantially altered nucleosome occupancy. Relatedly, there are fewer CpG dinucleotide sites in exonic regions flanking each intron within the mating-type region consistent with lower overall GC nucleotide content (Figure S3). Nonetheless, CG dinucleotides that are present within the mating-type region are significantly less likely to be methylated on a per-site level than are those in the autosomal regions of the genome (Figure S3). This implies that overall GC content and methylation efficiency are critical determinants of differences in methylation patterns.

2.4.3 Diverse Isoforms are Created by Widespread Missplicing

Additional categories of alternative splicing are also enriched in genes in the mating-type region beyond intron retention, consistent with a broad disruption of splicing fidelity. We circumvented the vagaries of interpreting short-read RNA sequencing data by applying a long-read circularized consensus method (Volden et al. 2018) to generate 14,350 high quality full length isoforms for the *M. pusilla* reference strain. We quantified additional classes of alternative splicing events using junction coordinates extracted directly from R2C2 isoform consensus sequences. For each gene with at least two isoforms, we identified 5' splice sites (A5SS), alternative 3' splice sites (A3SS), exon skipping events (SE), and retained introns (RI) based on comparisons of intron–exon boundaries across isoforms of the same gene (see Methods). Because these event types are defined exclusively by splice junctions, their detection cannot be inflated by intron retention. Classical alternative splicing events (A5SS, A3SS, and SE) are rare in *M. pusilla*: combined, they account for

approximately 100 events among mating-type genes and 190 events across 2,458 multi-isoform autosomal genes, one to two orders of magnitude fewer than retained introns (629 in the mating-type region; 1,595 autosomal). Critically, every alternative splicing event category is enriched on a per-gene basis in the mating-type region relative to autosomes (A5SS 8.2x, A3SS 6.9x, SE 2.3x, RI 2.8x; $p<0.0001$ for each; Table S8). These results indicate that the mechanism driving elevated splicing variability in the mating-type region operates broadly across event types and is not limited to intron retention.

Analysis of full length isoforms in *M. pusilla* produced detailed insights into mRNA processing at the transcript level. Using a linear regression model, isoform number increases with both expression and intron count (0.27 isoforms per log-expression unit, 0.76 per intron). Genes in the mating-type region produce ~1.9 more isoforms than autosomal genes, with predicted means of 1.83 (non-mating-type region genes) vs. 3.72 isoforms for mating type region genes at average expression and intron number ($p < 0.0001$, OLS; Figure 2.3). Additionally, genes within the mating-type region show more evenly distributed isoform expression, with less dominance by a single major isoform, compared to autosomal genes ($p < 0.0001$, GLM; see Methods). This pattern suggests that purifying selection is less efficient at suppressing non-canonical isoforms in the mating-type region.

We also compared other features of transcript diversity including transcript start sites and polyadenylation sites for genes within and outside of the mating-type regions. This analysis revealed selective disruption of 3′ end regulation whereby

genes in the mating-type region exhibit significantly more distinct transcript termination sites compared to genes outside the region (39.4% increase; $p < 0.0001$, Poisson GLM). This indicates that like splicing, termination and polyadenylation may be similarly disrupted. However, genes within and outside the mating-type region display comparable diversity in transcription start sites across isoforms ($p = 0.563$, GLM). We speculate that while the accumulation during evolution of AT-rich sequences may contribute to increased alternative polyadenylation, mispriming of oligo-dT primers during library construction could artifactually contribute to this pattern (although see below). Together, these findings suggest that co-transcriptional and post-transcriptional RNA processing—but not transcription initiation—are perturbed in the recombination-suppressed mating-type region.

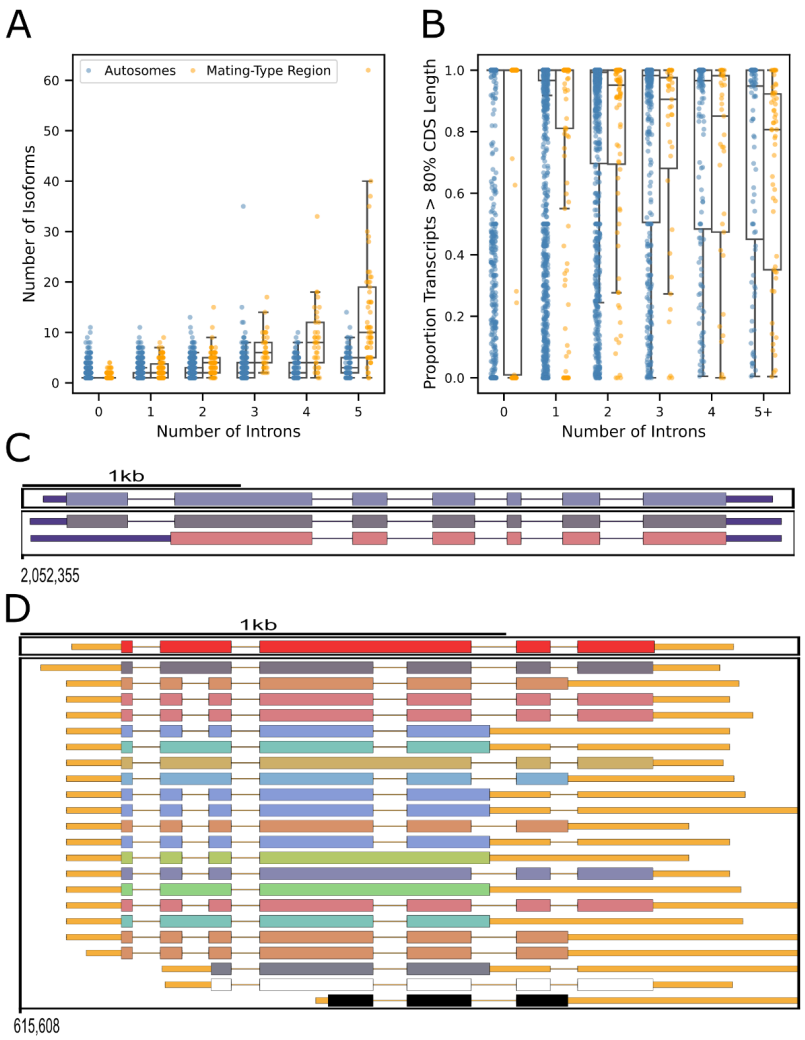


Figure 2.3 Isoform diversity inside and outside the M. pusilla mating-type region
(A) The number of isoforms per gene in mating-type (orange) and autosomal genes (blue). (B) The proportion of transcripts that contain a minimum of 80% of the CDS of the canonical gene's CDS. In A and B all genes with more than 5 introns are collapsed and displayed as 5+ introns. (C) The complete isoform set from a representative gene, MicpuC2.PR_0200000013, selected from the autosomal regions. (D) The complete isoform set from a representative gene, MicpuC2.EuGene.0000020604, selected from the mating-type regions. Each horizontal track represents a distinct isoform (colors assigned for visual distinction). Thick colored bars indicate exons, thin black lines indicate spliced introns, and orange medium-width lines indicate untranslated regions (UTRs). Representative genes were selected based on the median number of isoforms for each region and genes with 6 introns. Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/834.isoform.diversity.function.tsv).

Transcripts derived from genes within the mating-type region are less likely to be functional than those in other regions of the genome. In the absence of direct experimental validation, it is challenging to determine if a given transcript produces a functional protein. We therefore indirectly estimated functionality by asking what fraction of full-length transcripts are derived from isoforms that contain at least 80% of the annotated CDS in the correct reading frame without premature termination. Strikingly, transcripts from genes within the mating-type region showed a significantly lower fraction meeting this threshold ($p < 0.0001$, GLM, Figure 2.3B). To further assess the effect of alternative isoforms on protein function, we annotated functional protein domains and ontologies for each isoform. Isoforms derived from genes in the mating-type region also more frequently exhibit functional domain disruptions and changes in functional ontology compared to genes in the rest of the genome (Table S5). For example, for genes in the mating-type region, ~87% of isoforms contained at least 90% of annotated functional domains on average, compared to >92% for genes in autosomal regions ($p < 0.001$, Mann-Whitney U; Table S5). These patterns suggest that the elevated alternative splicing observed in the mating-type region is primarily deleterious, though we cannot exclude the possibility that some alternatively-spliced isoforms are neutral or positively selected.

2.4.4 Mispriming during cDNA preparation is unlikely to explain increasing missplicing in the mating-type region

A potential confounder for intron retention estimates in AT-rich regions is the capture of immature transcripts through oligo-dT mispriming, but several lines of evidence argue against this explanation. First, R2C2 isoforms were processed with

Mandalorion v4.6.0 using an A-content filter (Acutoff = 0.5) that discards any read whose 20bp downstream genomic window exceeds 50% adenine. Second, the GC content of the *M. pusilla* mating-type region (~42%) is comparable to that of the human genome, where oligo-dT mispriming is not typically considered a dominant artifact. Third, as described above, the alternative-splicing events beyond intron retention are enriched in the mating-type region, consistent with widespread misprocessing. Finally, to directly test whether downstream A-content predicts intron retention, we fit a logistic regression model within the 164 mating-type genes that contain both IR and non-IR isoforms (n = 1,393 isoforms). The per-isoform percentage of adenine in the 20 bp window downstream of the transcript termination site showed no significant association with intron retention (p = 0.31; Figure S5). Together, these results demonstrate that internal priming artifacts of immature transcripts are unlikely to account for the widespread splicing deficiency we observe.

2.5 Discussion

We propose that recombination suppression triggered a cascade of interconnected molecular changes and distinct selective pressures that collectively explain these widespread transcript-level defects (Figure 2.4). The suppression of recombination in the mating-type region dismantles the primary evolutionary mechanism that maintains genomic integrity against specific, underlying mutational pressures that have been empirically demonstrated in Mamiellales (Krasovec et al. 2017) and common to most eukaryotes (Katju and Bergthorsson 2019). This mutational pressure driving GC to AT transitions is counteracted in the recombining

portions of the genome by GC-biased gene conversion (gBGC). The reduction of recombination in the mating-type region reduces gBGC, diminishing the main impediment. The strong intrinsic bias for GC→AT mutations drives the dramatic shift toward an AT-rich sequence composition. Simultaneously, a deletion bias causes a net loss of base pairs per generation. This effect systematically shortens non-coding regions, and due to decreased efficacy of selection in non-recombining genomics segments, introns gradually evolve to be shorter. The collective action of mutational factors and natural selection thereby result in a distinct nucleotide composition within the mating-type region.

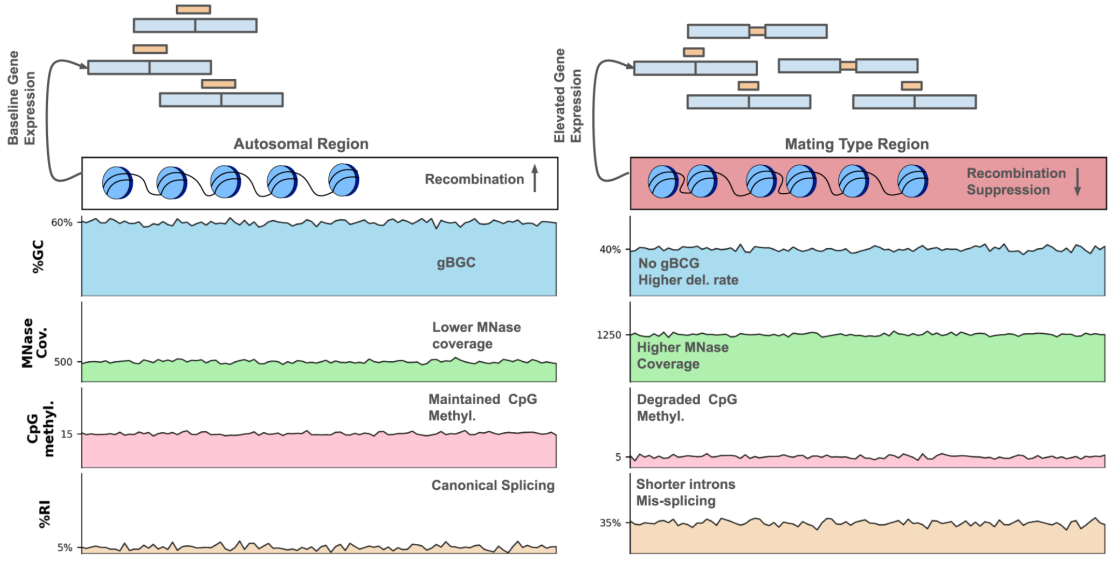


Figure 2.4 Proposed model of genomic erosion in Mamiellales mating-type regions
Schematic comparing autosomal regions (left) with mating-type regions (right), illustrating how recombination suppression initiates a cascade of molecular changes. Suppression of recombination eliminates GC-biased gene conversion (gBGC), allowing GC→AT mutational bias to reduce GC content from ~60% to ~40%. This compositional shift reduces CpG dinucleotide availability and methylation efficiency, disrupts nucleosome positioning (elevated MNase coverage indicates more open chromatin), and shortens introns through deletion bias. These chromatin and sequence changes collectively impair splicing fidelity: ~35% of mating-type introns are retained (thick orange bars in gene models) compared to ~3% in autosomal regions, producing abundant truncated, dysfunctional transcripts despite preservation of coding sequences. Expression levels (number of gene boxes) are elevated in the mating-type region due to altered chromatin accessibility, further exacerbating splicing defects through kinetic mismatches between transcription and co-transcriptional splicing.

We propose that this collective action of mutational biases and natural selection then causes decreased efficacy of splicing in the evolving mating-type region through a series of molecular consequences. First, this dramatic shift in GC content has compounding effects on chromatin structure: fewer CpG dinucleotides provide fewer sites for DNA methylation, while AT-rich sequences alter nucleosome positioning and consistency in the mating-type region (Davey et al. 2004; Huff and

Zilberman 2014; Tillo and Hughes 2009; Valiadi et al. 2025). Second, changes in GC content and nucleosome positioning can disrupt transcription kinetics essential for splicing efficiency and fidelity (Gelfman et al. 2013; Schwartz et al. 2009). Third, altered chromatin structure appears to elevate overall gene expression levels in humans (Jimeno-González et al. 2015), consistent with our finding that highly expressed genes in the mating-type region show the highest intron retention rates across species. This positive correlation between expression and retention is paradoxical given that some species show the opposite pattern (Bénitière et al. 2024; Saudemont et al. 2017), where selection minimizes splicing errors in highly expressed genes. However, in the *Mammiellales* species considered here, gene expression and intron retention are weakly positively correlated even in autosomal regions, where selection can operate efficiently to minimize intron retention (Spearman's ρ = 0.06-0.12 across species, $p < 0.0001$). This suggests fundamentally different mechanisms may govern transcription and splicing fidelity across eukaryotes. Furthermore, in the mating-type region, altered chromatin structure enables a loss of kinetic coupling between transcription and splicing. More rapid transcription by RNA polymerase II through more altered chromatin might outpace or disrupt co-transcriptional spliceosome assembly, leading to inappropriate exon inclusion or skipping (e. g. (Nogués et al. 2003; Fong et al. 2014)). Fourth, functional coupling between splicing and 3' end formation can lead to transcript 3' end heterogeneity when splicing is inefficient (Reimer et al. 2021), an effect we also observe in transcripts from the mating type regions. This cascade provides a mechanistic basis

for how non-recombining but functionally essential genomic regions accumulate progressive transcript-level defects while retaining the capacity to produce a fraction of functional transcripts sufficient to maintain viability.

A critical question remains why genes whose functions are unrelated to mating compatibility remain inside the UV region despite apparent persistent dysfunction. While comparative genomic analysis demonstrates that gene trafficking between UV and autosomal regions has occurred during Mamiellales evolution (Benites et al. 2021), we hypothesize that relocation alone does not immediately restore splicing function. In the event of a translocation, a gene would carry the sequence-level changes accumulated during its residence in the UV region, including degraded GC content that has been shown to persist in genes relocated from brown algal SDRs to autosomes (Lipinska et al. 2017). Studies of X chromosome gene trafficking demonstrate that relocated genes require substantial evolutionary time to re-establish proper regulatory contexts (Betrán et al. 2002; Emerson et al. 2004). Furthermore, genes would need to evolve compensatory mutations that restore lost splice site strength and intron length. This evolutionary lag may exceed the timescale over which selection maintains even partial gene function in situ, explaining why the system tolerates ongoing transcript-level dysfunction rather than rapidly relocating non-mating genes.

This mode of genomic erosion represents a fundamentally distinct evolutionary trajectory than the gene-loss paradigm exemplified by highly degenerate mammalian Y chromosomes. Rather than wholesale elimination of dispensable genes

under relaxed selection, mating-type regions experience a form of molecular compromise: genes are retained but function at reduced efficiency due to pervasive splicing deficiency. The persistence of this pattern across more than 300 million years of Mamiellales divergence suggests that producing even a modest fraction of correctly spliced transcripts may satisfy selective constraints when complete gene loss would be lethal. This framework has broad implications beyond green algal UV systems. Other non-recombining but essential genomic regions, including the young sex chromosomes of many plant and animal species, large structural polymorphisms segregating in natural populations, and even highly repetitive centromeric regions, may similarly accumulate cryptic transcript-level dysfunction that precedes or accompanies more visible forms of degeneration. Our model generates testable predictions: recently formed sex chromosomes should show intermediate splicing deficiency that scales with the time since recombination suppression, and experimental restoration of recombination should improve splicing fidelity. These findings therefore establish Mamiellales mating-type chromosomes as a powerful comparative system for dissecting how genome architecture, chromatin state, and RNA processing fidelity interact in shaping the evolution of essential non-recombining regions across eukaryotes.

2.6 Supporting Information

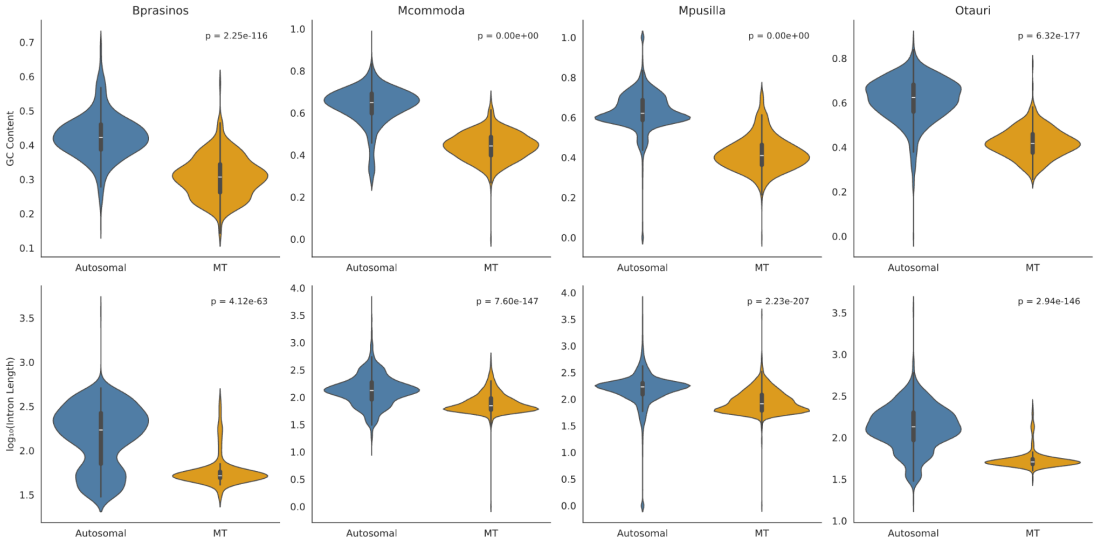


Figure S1 Intron length and GC content across genomic regions
Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/{species}.features.branchpoints.tsv).

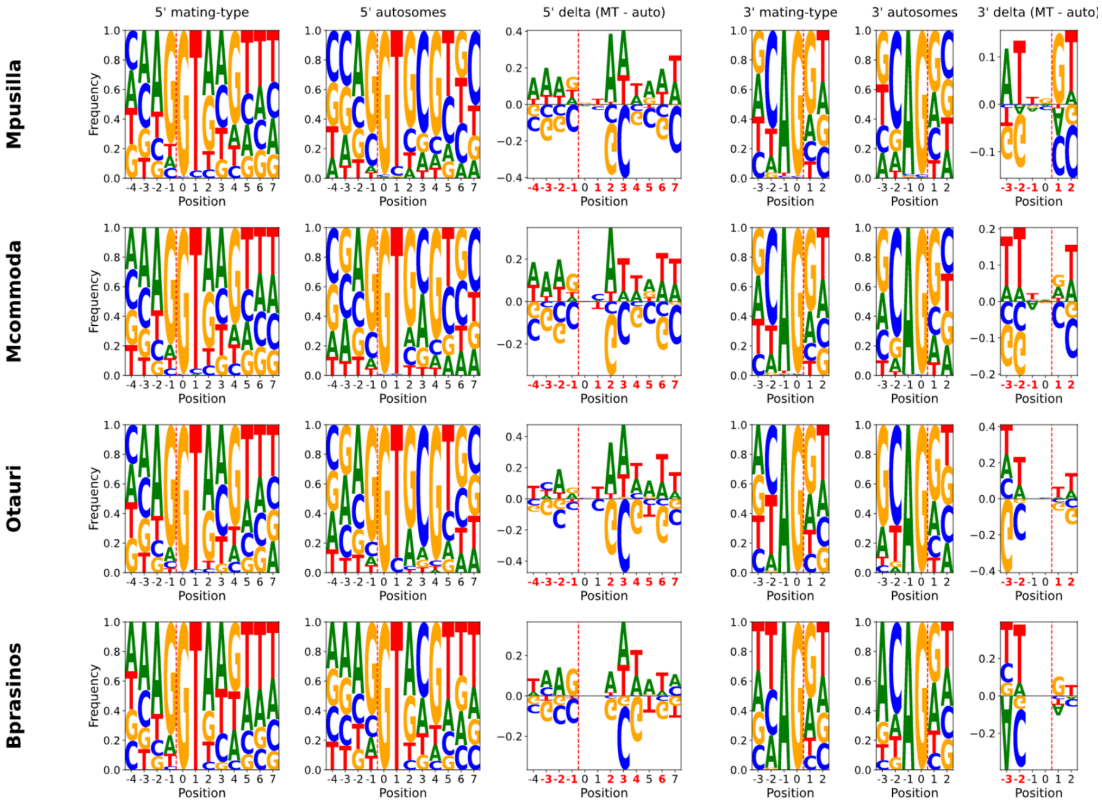


Figure S2 Splice-site sequence composition across genomic regions
Data from *M. pusilla* is shown in Figure 2.2 and replicated here to facilitate comparison across species. Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/{species}.features.branchpoints.tsv).

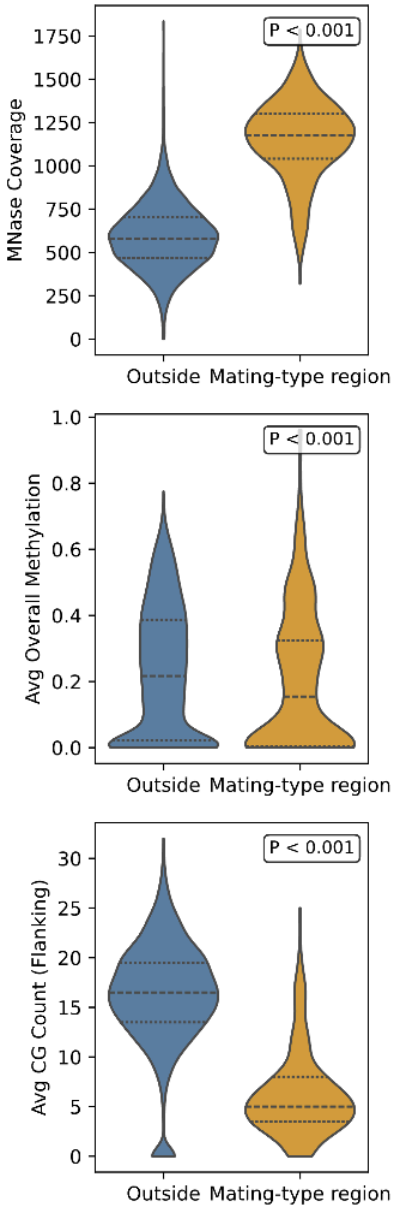


Figure S3 Nucleosome occupancy and methylation around *M. pusilla* introns
(Top) MNase-based library preparation coverage within introns. (Middle) Average per CG dinucleotide modification rates in bisulfite sequencing data. (Bottom) Number of CG dinucleotide sites in the 50bp exons flanking each intron. We used a Mann-Whitney-U test to determine p-values. MNase and methylation data from (Huff and Zilberman 2014). Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/Mpusilla.features.methylation.tsv).

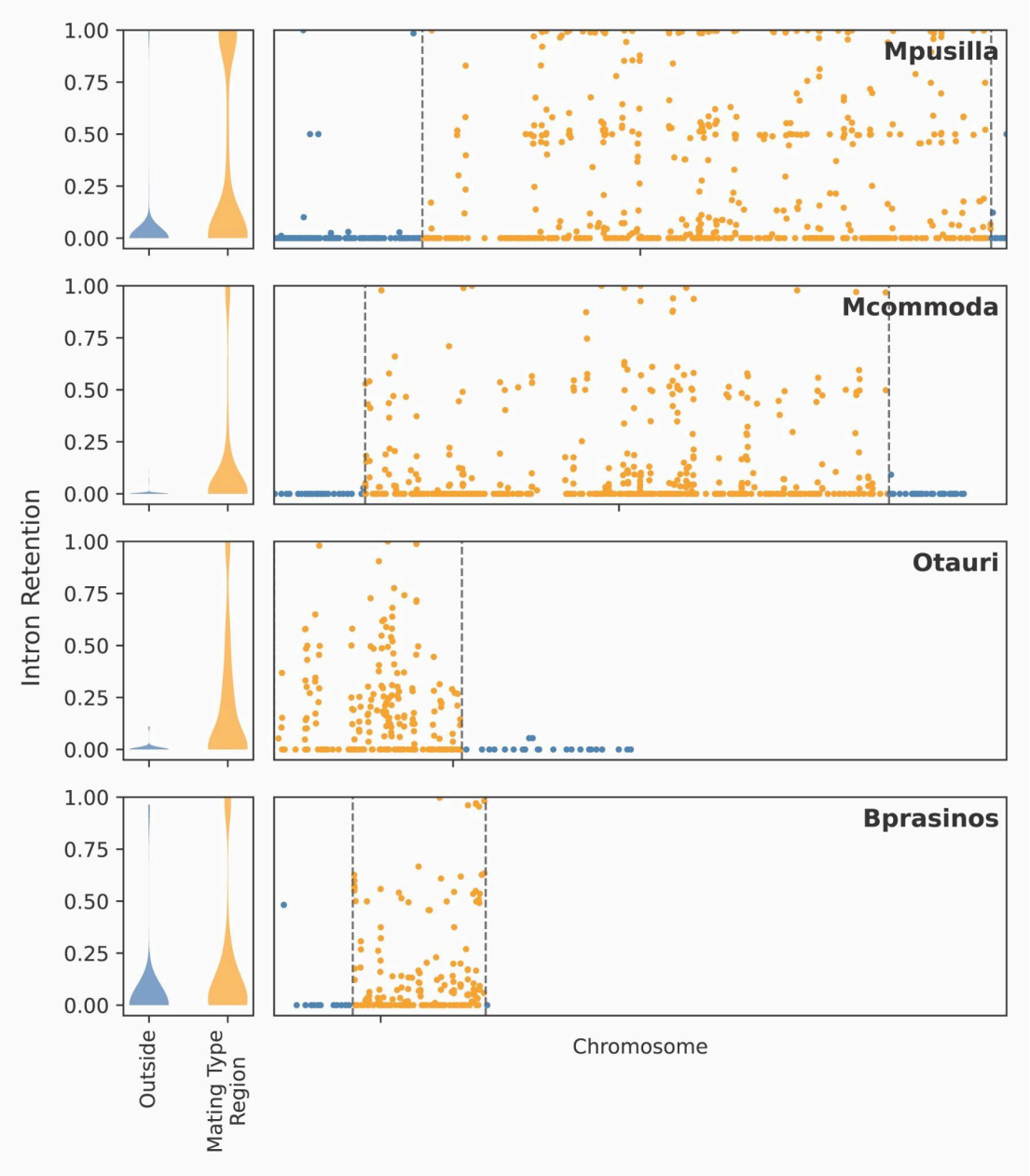


Figure S4 Chromosome-scale intron-retention patterns across Mamiellales (Left), per intron retention frequency along autosomal regions (blue) and within the mating-type region (orange). (Right) Retention frequency for each intron plotted along each chromosome with colors as on the left panel. We obtained coordinates for mating type regions from (Benites et al. 2021). Tick-marks indicate the midpoint of chromosome in each genome assembly. Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/{species}.features.branchpoints.tsv).

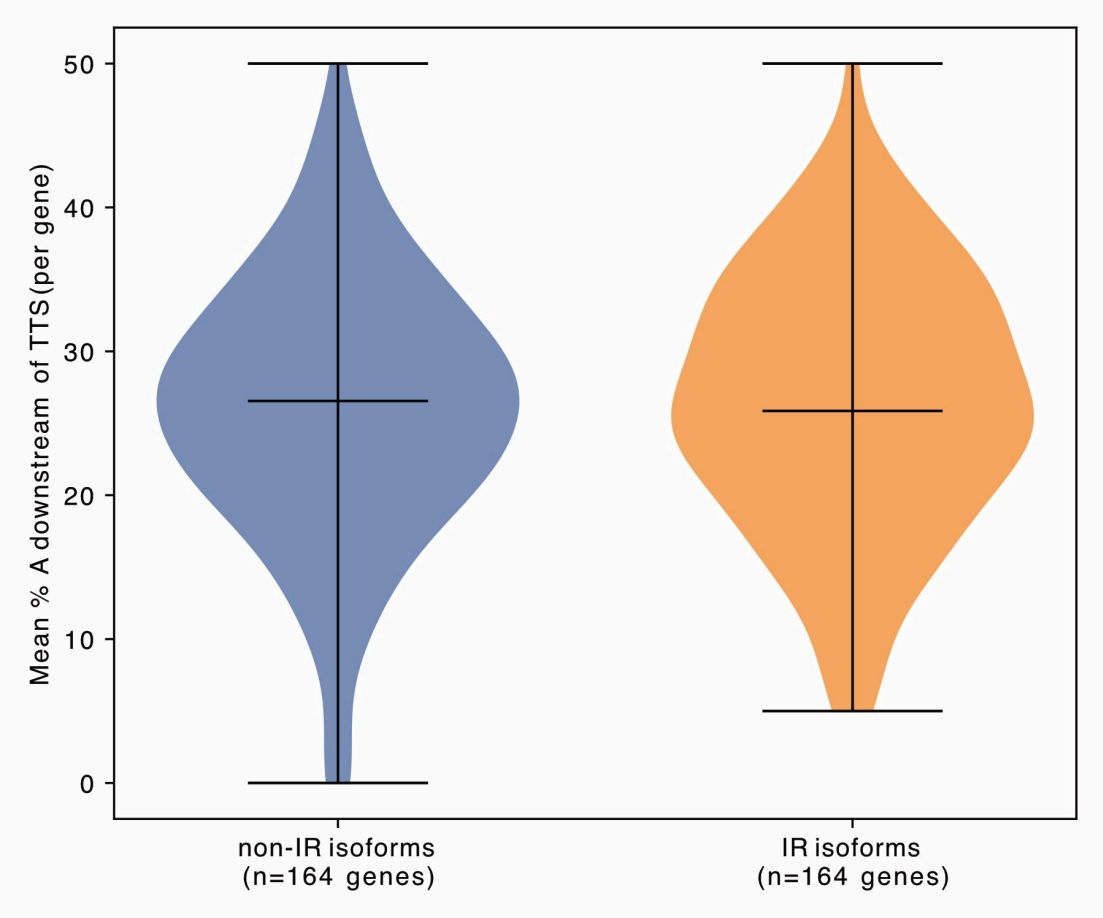


Figure S5 Adenine content downstream of transcript termination sites
Each gene contributes one mean to the non-intron-retained violin (blue) and one to the intron-retained violin (orange). Underlying data for this figure are available in the GitHub repository at https://github.com/russcd/mating-type-missplicing (file: data/internal_priming_vs_IR_per_isoform.tsv).

Table S1: Data sources and intron retention rates across Mamiellales species.

Species	Genome Accession	RNA-seq Accession	Mean Retention (Autosomes)	Mean Retention (Mating-Type)	Retention Mating-Type per Intron, per Transcript
M. pusilla	GCF_000151265.4	SRR7814466	0.011	0.33	0.45
M. commoda	GCF_000090985.2	SRR7814480	0.0017	0.15	0.17
B. prasinos	GCF_002220235.1	SRR14396982	0.015	0.18	0.17
O. tauri	GCF_000214015.3	SRR5986291	0.010	0.2	0.23

Table S2: Numbers of introns with a G at the 5th position in mating-type and non-mating type introns.

	Mating-Type		Non-mating type		
Nucleotide in 5th intron position	G	Other	G	Other	Fisher's Exact Test P-value
M. pusilla	651	371	5668	2299	3.9E-05
M. commoda	617	324	3395	1090	1.6E-12
O. tauri	299	136	1475	136	1.4E-30
B. prasinos	155	175	812	352	3.7E-10

Table S3: Consensus branchpoint sequences were identified genome-wide using DREME motif discovery. 'Present' = exact 6/6 nucleotide match within the 50-nt branchpoint window upstream of the 3' splice site. Ambiguity codes: V = A/C/G, R = A/G, S = C/G. Branchpoint is underlined.

		Mating-Type		Non-mating type		
	Branchpoint Sequence	Present	Absent	Present	Absent	P-value
M. pusilla	CTG<u>A</u>CS	109	914	3180	4787	3.64E-215
M. commoda	CTV<u>A</u>CC	81	860	1741	2744	8.11E-85
O. tauri	ACTR<u>A</u>C	50	385	1219	392	3.58E-136
B. prasinos	RCTS<u>A</u>C	22	308	527	637	6.22E-45

Table S4: Random forest classifier model accuracy, F1 scores and feature importances.

	Species	M. pusilla	M. commoda	B. prasinos	O. tauri
Evaluation	ROC-AUC (Full model)	0.876	0.785	0.837	0.833
	ROC-AUC (Minimal model)	0.867	0.735	0.812	0.830
Feature Importances (Minimal model)	TPM	0.434	0.339	0.397	0.338
	Distance from TSS	0.168	0.214	0.289	0.287
	Intron Length	0.249	0.252	0.171	0.148
	Intron GC Content	0.150	0.174	0.143	0.191
	Frame Disruption	Not Retained	0.022	Not Retained	Not Retained
	C at position 1 of 3’ Exon	Not Retained	Not Retained	Not Retained	0.036

Table S5: Protein structure preservation in annotated isoforms. Columns 2-4: proportion of isoforms retaining ≥X% of GO (Gene Ontology) terms assigned to the canonical protein. Columns 5-7: proportion of isoforms retaining ≥X% of annotated protein domains. These represent complementary measures of functional preservation.

threshold (X)	mean proportion of transcripts containing X total functional ontology terms for MT genes	mean proportion of transcripts containing X total functional ontology terms for other genes	MWU P-value	mean proportion of transcripts containing X total functional domains for MT genes	mean proportion of transcripts containing X total functional domains for other genes	MWU P-value
50	0.9621444077	0.9747142755	0.004539050829	0.9502255549	0.9684631655	0.000104108238 2
60	0.9571463833	0.9716433957	0.000538468706	0.9388186285	0.9600907444	1.20E-05
70	0.9437828405	0.9671404104	9.22E-06	0.9193801852	0.9490108626	4.59E-08
80	0.9302962741	0.961164801	8.06E-07	0.882107797	0.9300767338	7.39E-10
90	0.9282327821	0.9584154652	6.78E-07	0.8678387912	0.9220152699	2.50E-10

Table S6: Oligonucleotides used in smartseq and R2C2 library construction. Index corresponds to a random 8 nucleotide sequence unique to each replicate library.

Oligo ID	**Sequence**
TSO-Smart-seq2	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG
Oligo_dT_Index1	AAGCAGTGGTATCAACGCAGAGT [Index] ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
ISPCR primer	AAGCAGTGGTATCAACGCAGAGT
ISPCR_Anneal_Splint_1_F	ACTCTGCGTTGATACCACTGCTT GCACGCACCGACAAACTCTGGCCGATTGTACTTTCTTATAAG GCGTAACACTAGACCATATCGTTTCTATAGATTAATTATGGCG AGAATTGCTAGCGAACTAGACAATTTTCGAAATAATCCTTTTT ATAT AAGCAGTGGTATCAACGCAGAGT
ISPCR_Anneal_Splint_1_R	ACTCTGCGTTGATACCACTGCTT ATATAAAAAGGATTATTTCGAAAATTGTCTAGTTCGCTAGCAA TTCTCGCCATAATTAATCTATAGAAACGATATGGTCTAGTGTT ACGCCTTATAAGAAAGTACAATCGGCCAGAGTTTGTCGGTGC GTGC AAGCAGTGGTATCAACGCAGAGT

Table S7: NMD protein orthologs and their measured expression levels.

NMD Protein	**M. pusilla ortholog**	**M. commoda ortholog**	**B. prasinos ortholog**	**O. tauri ortholog**	**M. pusilla TPM**	**M. commoda TPM**	**B. prasinos TPM**	**O. tauri TPM**
UPF1	MICPUCDRAFT_56002	MICPUN_65270	Bathy10g01700	OT_ostta09g02800	28.0	55.2	48.5	39.7
UPF2	MICPUCDRAFT_46957	MICPUN_58110	Bathy18g00500	OT_ostta04g01650	31.8	51.0	25.9	101.8
UPF3	NA	NA	NA	OT_ostta15g01265	NA	NA	NA	17.7

Table S8: Alternative splicing event quantification from R2C2 long reads within the MT and autosomal regions in M. pusilla. Here we examined only genes with at least one intron, therefore requiring splicing for mature transcripts.

	MT Region		Autosomal Regions		
Event	Total Events	Mean per gene	Total Events	Mean per gene	P-value Fisher's Exact Test
A5SS	41	0.12	36	0.01	<0.0001
A3SS	57	0.17	59	0.02	<0.0001
ES	31	0.09	97	0.04	<0.0001
RI	629	1.83	1595	0.65	<0.0001

Chapter III: Population-specific introner turnover in *M. pusilla* reveals convergent genomic patterns and functional costs

3.1 Abstract

Spliceosomal introns vary dramatically in abundance across eukaryotes, yet the population-level processes governing intron gain, loss, and persistence remain poorly understood. We investigated introner turnover in thirteen geographically diverse isolates of the marine picoeukaryote *Micromonas pusilla* using short-read genome sequencing, pangenome analysis, long-read isoform sequencing, transcriptomics, and population-genetic modeling. Two deeply diverged populations differed approximately fourfold in introner abundance and shared few orthologous insertion sites, indicating extensive lineage-specific turnover. Nevertheless, introners accumulated in similar broad genomic regions and functional categories in both populations, suggesting conserved genomic constraints on insertion or retention. Within the introner-rich population, a U-shaped presence-absence frequency spectrum and reduced sequence diversity among polymorphic introners supported ongoing gains and losses. Introner polymorphisms were associated with elevated recombination rates but not with recent selective sweeps. Within orthologous genes, each additional introner was associated with 24% more alternative isoforms and 36% lower expression, while introner-containing genes produced more transcripts predicted to undergo nonsense-mediated decay. Introners were depleted from genes involved in core ribosomal and transcriptional functions, consistent with purifying selection against insertions in dosage-sensitive genes. Together, these results show that introner landscapes reflect recurrent mobility filtered by genome architecture,

molecular consequences, and host selection across short evolutionary timescales in nature.

3.2 Introduction

Spliceosomal introns are ubiquitous features of eukaryotic genomes that shape gene structure, enable alternative splicing, and modulate gene expression. Despite their ancient origins and fundamental role in eukaryotic complexity, intron numbers vary dramatically across lineages, from nearly intron-free genomes in some microsporidia to highly intron-dense genomes in vertebrates and plants (Irimia and Roy 2014; Csuros et al. 2011). Even among closely related species, intron content can differ by orders of magnitude, suggesting rapid evolutionary turnover in some lineages (Farlow et al. 2010). Understanding the population-level dynamics that generate this variation remains a central challenge in genome evolution.

Introners, transposable elements (TEs) that create functional spliceosomal introns upon insertion into exons, represent one mechanism capable of generating rapid, genome-wide intron gains. Introners have been proposed as a major driver of episodic intron proliferation events. Recent comparative genomic surveys have revealed that introner-derived introns are present in ~5% of surveyed eukaryotic species, spanning over 1.7 billion years of evolutionary history (Gozashti et al. 2022). However, these large-scale surveys necessarily focus on ancient, fixed introner insertions detectable through comparative analysis. The real-time population

dynamics of introner gain and loss that determine whether insertions fix, persist as balanced polymorphisms, or are rapidly purged, remain poorly characterized.

The marine picoeukaryote *Micromonas pusilla* harbors one of the most introner-dense genomes yet characterized, with thousands of introner insertions distributed across its compact ~22 Mb genome (Worden et al. 2009; Huff et al. 2016). Although most *M. pusilla* introners are supported by spliced transcripts, introner insertion can alter splice-site architecture and, in some cases, change the resulting mature transcript, suggesting that imperfectly processed insertions may impose transcriptional or splicing-associated costs on their host (Gozashti et al. 2022; Huff et al. 2016). However, the population-level consequences of these molecular effects—whether introner insertions experience purifying selection, drift to fixation, or exhibit more complex dynamics—have not been systematically examined. Furthermore, the relationship between introner load and organismal fitness, and whether different introner abundances reflect stable equilibria or transient states, remains unknown.

Here, we characterize introner population dynamics across 13 geographically diverse *M. pusilla* isolates, combining whole-genome sequencing, long-read isoform sequencing, and population genetic analysis. We reveal two deeply diverged population groups differing four-fold in introner abundance, yet exhibiting nearly identical introner insertion preferences across genomic regions and functional gene categories. Within the introner-rich population, we observe active introner turnover characterized by a distinctive U-shaped allele frequency spectrum likely reflecting

ongoing gains and losses, and we demonstrate that polymorphic introners are preferentially located in regions with higher recombination rates. Using generalized linear models integrating RNA-seq and long-read isoform data across three strains, we quantify the functional consequences of introner presence where each introner reduces gene expression by 36% while increasing transcript isoform diversity by 24%, with introner-containing genes showing elevated rates of nonsense-mediated decay. Together, these findings illuminate the multi-scale evolutionary forces shaping introner dynamics, from the immediate molecular costs driving purifying selection to the recombination-mediated processes facilitating both insertion and deletion over ecological timescales.

3.3 Methods

3.3.1 Acquisition and Culture of 13 M. pusilla strains

We obtained sixteen isolates of *M. pusilla* from the Roscoff Culture Collection. We identified three of these isolates to be identical matches to another isolate in the dataset (CCMP490, RCC647, and RCC835), and thus these three were excluded from this study. We note that isolate RCC834 used in RNA-seq analysis is the same isolate as CCMP1545, initially sequenced by Worden et al (Worden et al. 2009) and widely used as a reference for this species. The full list of strains and their metadata can be found in Table S1.

We grew all *M.pusilla* cultures in low salinity (0.8X) artificial seawater with L1 supplements (without Na_2SiO_3), (Enevoldsen and Unesco 2003) using the L1 kit purchased from the National Center for Marine Algae/Bigelow (cat. No. MKL150L).

We maintained small (10ml) cultures in vented T25 flasks at 18°C under a 14h light:10 h dark illumination cycle at 220 microEinsteins (uE)

3.3.2 RCC1749 long read assembly and alignment to RCC1749

We constructed a long-read assembly of RCC1749 using Oxford Nanopore sequencing. Raw reads were quality-filtered to retain only sequences ≥5,000 bp in length, resulting in ~244× estimated genome coverage. We performed de novo assembly using Flye v2.9.1-b1780 with the --nano-hq parameter, an expected genome size of 22 Mbp, and target assembly coverage of 50×. The resulting draft assembly was polished through a three-step process: first with Racon (Vaser et al. 2017) using nanopore read alignments generated by BWA-MEM, followed by consensus calling with Medaka (model R941_min_high_g303) , and final polishing with Homopolish (Huang et al. 2021) using the R9.4.pkl model. Assembly quality was assessed using BUSCO and QUAST (Tegenfeldt et al. 2025; Gurevich et al. 2013).

We used MUMmer v4.0 (Marçais et al. 2018) to perform a whole-genome nucmer alignment between the 21.95-Mb CCMP1545 and 19.44-Mb RCC1749 assemblies. Alignments were filtered with delta-filter -1 to retain one-to-one matches. The filtered alignment contained 11,027 alignment blocks spanning 15.79 Mb of CCMP1545 and 15.85 Mb of RCC1749, corresponding to 71.9% and 81.5% of the respective assemblies. Aligned regions had a length-weighted mean nucleotide identity of 90.9%. Alignment blocks had a median length of 667 bp and a maximum length of 54.2 kb.

3.3.3 RNA extraction and cDNA library preparation for isolates RCC834, RCC1614 and RCC1749

We extracted RNA in four replicates from approximately 1E8 cells of *M. pusilla* strains RCC834, RCC1614 and RCC1749 using a mixture of 750uL Trizol and 250uL chloroform, followed by purification with a Qiagen RNeasy extraction kit. We prepared cDNA libraries through two sequential reactions. First, we performed reverse transcription on 200 ng of total RNA using Smartscribe reverse transcriptase, FS buffer, DTT, indexed oligo-dT primers, and a template switching oligo (TSO), incubating at 72°C for 3 minutes and 42°C for 60 minutes. Next, we amplified the resulting cDNA by PCR using KAPA HotStart master mix with primers targeting the oligo-dT and TSO sequences. All oligo and primer sequences can be found in supplementary table S4.

3.3.4 Illumina sequencing

We prepared Illumina paired-end libraries from *M. pusilla* genomic DNA or RNA-derived cDNA libraries using an in-house Tn5 transposase-based protocol (C. Mirchandani et al. 2024) followed by selection for fragments >= 300 bp with a Zymo Select-a-Size DNA Clean & Concentrator kit for each sample. We sequenced both genomic DNA and cDNA libraries using Illumina NextSeq paired-end sequencing with a 150bp read length. Libraries covered an average read depth of 204x per isolate for genomic DNA, and an average 35.8 million read pairs per replicate for cDNA libraries.

3.3.5 RNA-seq analysis

RNA-seq reads were aligned to strain-specific reference genomes using HISAT2 v2 (Kim et al. 2019). Resulting alignments were sorted and indexed using samtools (Li et al. 2009). Gene expression quantification was performed using featureCounts from the Subread package (Liao et al. 2013), counting paired-end reads at the exon level and grouping by gene_id.

3.3.6 M. pusilla variant calling

We performed variant calling against both the Population 1 (CCMP1545) and Population 2 (RCC1749) reference genomes to generate accurate haplotypes for the 12 non-reference *M. pusilla* strains. First, alignment files for each strain were preprocessed, where reads with TP:Z tags were removed using awk, and files were processed with Picard AddOrReplaceGroups to add standardized read group information. Duplicated reads were subsequently identified and marked using GATK MarkDuplicates (Van der Auwera et al. 2013).

Variant discovery was performed using a two-pronged approach. SNPs and small indels were called for each strain using GATK HaplotypeCaller (Van der Auwera et al. 2013). In parallel, large, non-reference insertions (>40bp) were identified and assembled *de novo* from each alignment file using INSurVeyor (Rajaby et al. 2023). The variant calls from GATK and large insertions from INSurVeyor were then merged into a comprehensive, non-redundant variant set using bcftools isec and bcftools concat (Danecek et al. 2021). The final VCF file was filtered to retain only high-quality variants (QUAL > 30) using GATK VariantFiltration. Finally, the filtered

variants were applied to the corresponding reference genome using GATK FastaAlternateReferenceMaker to produce a final haplotype FASTA file for each strain against each of the two references.

3.3.7 Cactus pangenome construction

We constructed a pangenome from 13 reference and strain-specific local de novo assemblies using minigraph-Cactus (Hickey et al. 2024), designating CCMP1545 and RCC1749 as reference genomes. The workflow partitioned the initial minigraph alignment into scaffold-level subproblems, performed progressive Cactus alignments for each partition, and combined the alignments into a whole-genome HAL representation. The resulting graph contained 25.36 Mb of sequence distributed across 3,363,118 nodes and 4,578,075 edges. Both a VCF containing the graph variants and the deconstructed graph assemblies were used for downstream analysis.

3.3.8 Strain-specific gene annotation liftover

We extracted protein sequences from the CCMP1545 genome and GTF using agat_sp_extract_sequences.pl (https://github.com/NBISweden/AGAT) and aligned them to each sample-specific assembly using miniprot (Li 2023) with protein-to-genome alignment and GTF output enabled. When a reference protein aligned to multiple candidate loci, the nucleotide sequence of each candidate interval was extracted and compared with the corresponding concatenated CCMP1545 CDS using splice-aware minimap2 alignments (Li 2018). We then resolved protein mapping globally to retain one locus per reference protein and prevent the same locus from being assigned to multiple proteins. Resolved mappings were required to have at

least 70% sequence identity based on the miniprot PAF alignment, and when retained gene models overlapped on the same target contig, the model with lower sequence identity was removed.

3.3.9 Seed introner discovery

We identified and catalogued introner sequences from multiple sources of genomic evidence. We first curated a high-confidence "seed" database of introners from known introns previously identified via RNA-Seq analysis for both reference genomes. To create a non-redundant set, any overlapping intron annotations were resolved by retaining only the longest sequence. This curated set was then subjected to an all-vs-all BLASTn (Camacho et al. 2009) search to identify families of related sequences. Only sequences that shared homology with another sequence in the set (>= 80% identity over >=80% of both query and subject length) were retained, forming our reference introner database. We used this seed introner set to identify additional homologous sequences from insertion variants derived from our Cactus graph. Each of these sequence sets was searched against our reference introner database using BLASTn. Sequences from these datasets that exhibited high homology to the reference introners (>=80% identity over >=80% length) were classified as introners.

3.3.10 M. pusilla Introner Genotyping

Using our high-confidence introner reference set, we identified candidate loci in each assembly and constructed a genotype matrix spanning all orthologous loci. We searched each assembly with BLASTn using reference introners as queries,

retaining hits with ≥80% nucleotide identity and alignment lengths 80–120% of query length. Hits at identical or overlapping target coordinates were collapsed by favoring longer queries and then higher identity, and low-complexity candidates were removed. Candidate bodies were then compared by BLASTn with individual *M. pusilla* introners from the families described by (Gozashti et al. 2022). The highest-bitscore hit determined family assignment, and candidates required ≥70% identity and ≥80% query coverage. Genic candidates additionally required a GT or GC 5′ splice donor within the first 20 bp; candidates spanning gene boundaries were excluded.

We next compared loci among assemblies using the 100-bp sequences immediately flanking each validated candidate. Left and right flanks from every query strain were aligned separately to every other assembly with BWA-MEM (Li and Durbin 2009). In the primary pass, retained flank alignments required MAPQ ≥30 and at least 70 aligned bases, with both flanks mapping to the same target contig. A target locus was called present when the mapped interval overlapped a validated introner by at least 60% of the shorter interval; concordantly mapped flanks without a validated introner were called absent, including intervals with a gap <40 bp or total span <200 bp. Loci with an unmapped flank, flanks mapping to different contigs, or an implausibly long forward-strand span (>500 bp) were called missing. Unresolved comparisons between populations were re-evaluated using relaxed BWA-MEM thresholds and searches guided by conserved host-gene and neighboring-gene context.

Pairwise relationships were represented as a network in which nodes corresponded to sample-specific genomic loci and edges represented supported present–present or present–absent relationships. Connected components defined initial locus groups; components containing multiple loci from one strain were resolved using edge support and were partitioned when their present members belonged to different introner families. Locus groups were then iteratively refined using gene-relative insertion fingerprints, codon position, flanking amino-acid context, and tandem-duplication evidence, followed by introner-body sequence comparisons to assess within- and between-population orthology. Unresolved sample–group combinations were encoded as missing data. Finally, short reads from each nonreference strain were aligned to its own assembly, and depth-based calls from reads with MAPQ ≥30 were incorporated into the genotype matrix.

3.3.10 Population-wide SNP Analysis

We used the VCF derived from the Cactus graph and filtered out all indels and identified 4-fold degenerate (4D) sites using degenotate (C. D. Mirchandani et al. 2024). We performed PCA on the 4D sites using PLINK (Purcell et al. 2007) with window size 50, step 5, and r^2 threshold 0.2, with analysis restricted to LD-pruned SNPs to reduce correlation effects. We then identified synonymous, missense and nonsense SNPs in the full VCF using SNPEff (Cingolani et al. 2012).

3.3.11 Demographic Modeling

We inferred the demographic history of *M. pusilla* using the diffusion approximation method implemented in moments (Jouganous et al. 2017). The analysis

was based on the two-dimensional joint site frequency spectrum (2D-SFS) constructed from 4D sites. We defined the two populations: "Population 1" (n=11) and "Population 2" (n=2). We fitted a demographic model of population splitting with subsequent symmetric migration to simulate an ancestral population that splits at time (T) into two daughter populations, which then exchange migrants. The model was optimized to estimate the effective population sizes of Population 1 (N1) and Population 2 (N2), the split time (T), and the migration rate (M). The population-scaled mutation rate, Theta ($4N_e\mu$), was also estimated from the data.

3.3.12 Phylogenetic Tree Construction

We performed a maximum likelihood (ML) phylogenetic analysis on all *M. pusilla* strains using a PHYLIP-formatted alignment derived from the graph-derived 4D site VCF. Of the 180,518 sites, 147,244 were phylogenetically informative. We used IQ-TREE v2.2.0 (Nguyen et al. 2015) to conduct the tree search and inference with the General Time Reversible (GTR) model, incorporating empirical base frequencies (+F) and an ascertainment bias correction for SNP data (+ASC).

3.3.13 LD Decay Analysis

We excluded out strains RCC1749 and RCC3052 from the following analysis due to their divergence from the rest of the strains. Using the 11 individuals of the *M. pusilla* Population 1 4D VCF, we performed LD decay analysis, using VCFtools (Danecek et al. 2011), we calculated the squared correlation coefficient (r^2) for all SNP pairs located on the same chromosome and within a 100 kb window of each

other. The resulting pairwise r^2 values were then averaged across 25 bp distance bins to visualize the decay of LD with physical distance.

3.3.14 Recombination Rate Analysis of Introner-containing Genomic Features

We inferred fine-scale recombination rates for Population 1 with pyrho v0.2.0 (Spence and Song 2019) from phased, biallelic SNPs divided into separate VCFs for the 19 CCMP1545 scaffolds. A two-locus likelihood table was generated using a mutation rate of 9.8×10^{-10} substitutions per site per generation and a two-epoch demographic history simulation. Final maps were inferred independently for each scaffold with pyrho optimize, using phased haplotypes (--ploidy 1), a likelihood window of 125 SNPs and a block penalty of 150. We used the CCMP1545 Pyrho recombination map to project onto RCC1749 using the MUMmer whole-genome alignment between the two assemblies. Pyrho intervals overlapping aligned regions were transformed to RCC1749 coordinates while accounting for forward, reverse, and split alignments, with the original recombination-rate estimates retained. Pyrho output was interpreted as a piecewise-constant estimate of the per-base, per-generation recombination rate between consecutive SNPs.

To compare recombination environments, exons from the CCMP1545 RCC1749 reference annotations were merged within genes and partitioned into introner-containing and exonic regions. Introners separated by no more than 10 kb within the same gene were grouped into clusters, and a 5 kb window was centered on each cluster midpoint. Exonic regions remaining after subtraction of introner clusters were tiled with non-overlapping windows of up to 5 kb; terminal windows shorter

than 2.5 kb were excluded. For each analysis window, the recombination rate was calculated as the overlap-length-weighted mean of all intersecting pyrho intervals. Windows within the mating-type region on chromosome 2 (49,808–1,730,591 bp) were excluded.

3.3.15 Population 1 vs. Population 2 introner genomic correlation

We assessed the correlation of introner density distributions across aligned genomic regions to determine whether introner placement patterns were shared between homologous sequences in CCMP1545 and RCC1749. We used MUMmer whole-genome alignments to define corresponding, non-overlapping windows (1, 5, 10, 20, and 50 kb) between the CCMP1545 reference and RCC1749 query genomes. For each window pair, we calculated the introner density as the ratio of introner count to window size, normalized to introners per kilobase. We excluded windows containing exclusively fixed introners (introners present in both strains) to focus on variable introner distributions. We computed Spearman rank correlation coefficients and associated p-values for each window size independently.

3.3.16 Introner Genetic Diversity Estimation

We estimated the genetic diversity across orthologous introner loci by calculating the genetic divergence of fixed loci (Dxy) between Population 1 and Population 2, π within fixed loci in Population 1 and Population 2, and π across polymorphic loci in Population 1. We excluded any groups of loci where the introner was not callable due to insufficient read coverage. Consensus introner flanking regions were constructed and aligned in an identical approach as 4-fold degenerate

regions. Diversity was computed between and within Population 1 vs. Population 2 multiple sequence alignments in three groupings: π(Population 1), π(Population 2) and Dxy(Population 1 vs. Population 2). Diversity near polymorphic Population 1 loci were calculated in three groupings: π(present), π(absent) and π(present vs. absent. We additionally computed Dxy and π for orthologous introner loci bodies, following the same procedure as described above.

3.3.17 Introner and canonical intron gain & loss rate inference

To infer introner gain and loss dynamics, we modeled the Population 1 introner presence/absence spectrum as a demographic-SFS-scaled mixture of recent gains and losses. The workflow fits a one-population moments demographic model to the folded 4D SFS, and tabulates a two-state SFS for introner loci, with bins indexed by the number of Population 1 samples carrying the introner-present state. Our MLE model, based on (Vogl et al. 2020), treats the observed segregating counts y_i for present-count bins i = 1, … , n-1, as independent Poisson random-field counts. Let a_i be the unfolded neutral SFS shape at bin i under the fitted demographic model. The expected contribution to bin i for a gain event is $\theta_\lambda a_i$. For a loss, the absent allele is derived and the present-count bin i corresponds to the absent-derived count n - i, giving an expected contribution $\theta_\mu a_{n-i}$. Thus, the fitted mean is $m(\theta_\lambda, \theta_\mu) = \theta_\lambda a_i + \theta_\mu a_{n-i}$, and the model estimates θ_λ and θ_μ by minimizing the negative Poisson log-likelihood $-\ l = \sum_{i=1}^{n-1} [m_i - y_i log(m_i)]$ and optimizing on

log-transformed θ values for positivity. The same procedure was applied to matched canonical introns, and the resulting θ_{λ} and θ_{μ} estimates were normalized by the number of callable loci in each class so that introner and canonical intron gain/loss rates could be compared on a per-locus basis.

3.3.18 Selective Sweep Analysis

We constructed a Population 1 allele-frequency dataset from fourfold-degenerate SNPs, retaining biallelic segregating sites with sufficient genotype calls across Population 1 samples and excluding the CCMP1545 mating-type region. A folded empirical site-frequency spectrum was estimated genome-wide with SweepFinder2 (DeGiorgio et al. 2016) and used as the neutral frequency spectrum for composite-likelihood-ratio (CLR) scans. CLR values were calculated at 1 kb grid intervals across eligible CCMP1545 contigs, and high-CLR sweep-like regions were defined from grid points exceeding the empirical 99th percentile of the genome-wide CLR distribution, merging adjacent high-CLR grid points within 2 kb. We then compared the genomic positions of fixed and polymorphic Population 1 introners, and regular spliceosomal intron controls to the called SweepFinder2 regions. For each locus class, we calculated the fraction of loci overlapping sweep-like regions and assigned each locus the nearest SweepFinder2 CLR value. Significance of sweep-region overlap was evaluated by permutation, sampling the same number of positions from the eligible non-mating-type genomic space and comparing the observed overlap fraction to the resulting null distribution.

3.3.19 R2C2 Full Length Isoform Sequencing and Analysis with Mandalorion

We circularized the RNA-seq cDNA libraries via Gibson assembly (NEBuilder HiFi) using a short, pre-annealed double-stranded DNA splint overlapping the cDNA ends. To remove uncircularized molecules, we digested the reaction with ExoI, ExoII, and Lambda Exonuclease (all NEB) for 16 hours at 37°C, then heat-inactivated the enzymes for 20 minutes at 80°C. We cleaned the reaction with SPRI beads at a 1:0.85 ratio.

We used the clean, circularized library as a template for rolling circle amplification (RCA) with Phi29 polymerase (NEB) and a random hexamer primer for 18 hours at 30°C, followed by heat inactivation for 10 minutes at 65°C. We then debranched the Phi29 reaction using T7 endonuclease for 2 hours at 37°C and cleaned and concentrated the product with a NEB Monarch DNA clean and concentrator column. We size-selected the resulting R2C2 library (Volden et al. 2018) for DNA fragments larger than 3 kb using ProNex beads (Promega) and quantified DNA concentration with a Qubit fluorometer.

We sequenced the multiplexed R2C2 library on an ONT PromethION R10.4 flow cell and basecalled reads with the Dorado basecaller using model dna_r10.4.1_e8.2_400bps_sup@v5.0.0. The resulting raw reads were converted into consensus reads and demultiplexed using C3POa (v3.2). Based on these demultiplexed consensus reads, Mandalorion (v4.6) was used to identify and quantify high-confidence isoforms.

We refined the initial Mandalorion isoform calls using SQANTI3_QC (Pardo-Palacios et al. 2024) using default parameters with the addition of "--isoAnnotLite" to attempt to liftover CDS features from the original gff3 file.

3.3.20 Isoform Usage Analysis

We quantified isoform expression evenness using Shannon diversity metrics calculated from SQANTI3-classified isoforms across CCMP1545, RCC1614 and RCC1749. For each gene with at least two isoforms, we computed the Shannon index to measure overall isoform diversity and Shannon evenness to assess the proportionality of isoform expression independent of isoform number. We filtered isoforms to high-confidence categories (full-splice-match, novel-in-catalog, novel-not-in-catalog) and compared Shannon metrics between genes dominated by annotated (FSM) versus novel isoforms using Mann-Whitney U tests.

We tested whether changes in introner number were associated with transcript isoform richness in strain-specific introner genotypes and SQANTI3 isoforms across CCMP1545, RCC1614, and RCC1749 using common gene identifiers. Isoform richness was defined as the number of distinct high-confidence isoforms beyond the baseline transcript. We used a gene-conditioned Poisson model comparing the same orthologous gene across strains, with current introner number, strain, and within-gene long-read sequencing support as predictors. This formulation controlled for gene-specific characteristics and did not require assigning individual differences as ancestral gains or losses. Analyses were restricted to genes with at least 10 supporting long reads and complete introner callability.

Finally, we tested whether introner-containing genes produced a greater proportion of isoforms predicted to undergo nonsense-mediated decay. SQANTI3-classified, multi-exon isoforms with an NMD prediction were aggregated into NMD-positive and NMD-negative counts for each gene and strain. These counts were analyzed using a binomial GLM clustered by common gene, with introner presence as the primary predictor and strain, total long-read support, detected isoform number, gene span, and background exon count as covariates. Genes within the mating-type region were excluded from these functional analyses.

3.3.21 Expression and Splicing Analysis

We modeled gene expression using a negative-binomial GLM fitted to raw short-read RNA-seq counts from each gene and replicate, with log library size included as an offset. Predictors included strain, current introner number, CCMP1545-relative introner gains and losses, log gene span, number of alternative isoforms, and gene GC content. Model coefficients were exponentiated to obtain multiplicative effects on expected expression.

We compared splicing efficiency between fixed and polymorphic introners byquantifying intron retention from reads crossing the 5′ or 3′ splice boundaries and reads supporting the exact splice junction. Retention was calculated as ((U_5+U_3) / (U_5+U_3+2S)), where (U_5) and (U_3) are unspliced boundary-crossing counts and (S) is exact junction support. Retained and spliced read counts were analyzed with a beta-binomial regression including intron class, gene expression, their interaction, and RNA-seq replicate.

3.3.22 GO Enrichment Analysis

We compiled a CCMP1545 gene-to-GO annotation set by merging direct GO annotations from the reference annotation file with GO terms inferred from Pfam/InterPro2GO, KEGG Orthology, TAIR ortholog, PANTHER, and eggNOG-mapper annotations (Kanehisa et al. 2025; Reiser et al. 2024; Mi et al. 2021; Cantalapiedra et al. 2021). GO terms were standardized and propagated to ontology ancestors using the GO ontology. Introner-associated gene sets were defined from the introner genotype matrix, including genes containing any present introner, ancestral shared introners, likely independent insertions, Population 1-polymorphic introners, Population 1-fixed introners, Population 2-fixed introners, and introners belonging to each sequence family.

For each gene set, we tested GO-term enrichment and depletion relative to the background of all GO-annotated CCMP1545 genes using two-sided Fisher's exact tests, requiring at least five background genes per GO term. P-values were corrected across all group-term tests using the Benjamini-Hochberg false discovery rate procedure, with significance defined as global FDR < 0.05. To control for the greater opportunity for longer coding sequences to contain introners, significant GO enrichments were further evaluated with CDS-length-weighted empirical permutations: for each introner group, we randomly sampled the observed number of GO-annotated study genes from the GO/CDS-eligible background with probability proportional to CDS length, repeated this procedure 1,000,000 times, and calculated directional empirical P-values for enrichment or depletion with a +1 correction.

3.4 Results

3.4.1 Population Structure

Genome-wide variation revealed two deeply diverged populations of *M. pusilla*. We obtained 12 clonal *M. pusilla* isolates from the Roscoff Culture Collection and sequenced them using Illumina short reads. We generated a high-quality long-read assembly for the more divergent isolate RCC1749 using Oxford Nanopore sequencing and constructed local *de novo* assemblies for all remaining isolates by integrating SNP and structural variant calls against the CCMP1545 and RCC1749 references (see Methods). We constructed a pangenome graph from these assemblies with minigraph-cactus, extracted variants in VCF format using CCMP1545 as reference, and classified SNPs into synonymous, nonsynonymous, and four-fold degenerate sites for population genetic analysis. Across the 12 isolates, Illumina sequencing generated 529.6 million paired-end read pairs (median, 26.0 million per isolate; range, 19.6–217.3 million). The joint call set contained 876,278 biallelic SNPs genome-wide and 791,466 after excluding chromosome 2. The latter represented 4.0% of analyzed bases and comprised 239,929 nonsynonymous, 315,214 synonymous, and 236,323 noncoding SNPs, including 180,869 SNPs at fourfold-degenerate sites. Across the fourfold-degenerate analysis set, mean per-isolate genotype depth ranged from 15.9× to 203.3× (median, 140.8×), and genotypes were called at 93.0 to 99.3% of sites per isolate (median, 99.0%).

Multiple lines of evidence indicate that these populations have experienced long-term genetic isolation. The folded site-frequency spectrum contained 130,541 sites fixed for alternate alleles between populations but very

few shared polymorphisms, indicating minimal recent gene flow (Figure 3.1a). Consistent with this pattern, demographic modeling with moments supported an ancient population split (~14.3 million generations), subsequent reductions in effective population size in both lineages, and an extremely low migration rate (6.0×10^{-11} migrants per generation) (Figure 3.1c).

Within Population 1, patterns of linkage disequilibrium indicate an actively recombining population suitable for population genetic analyses. Linkage disequilibrium decayed rapidly from an initial r^2 of 0.47 to a background of 0.14, while Tajima's D (-0.45) indicated a modest excess of rare variants consistent with purifying selection or recent population expansion (Figure S1). Together, these patterns indicate that Population 1 behaves as a recombining population with sufficient haplotypic variability to faciliate subsequent population genomic analyses.

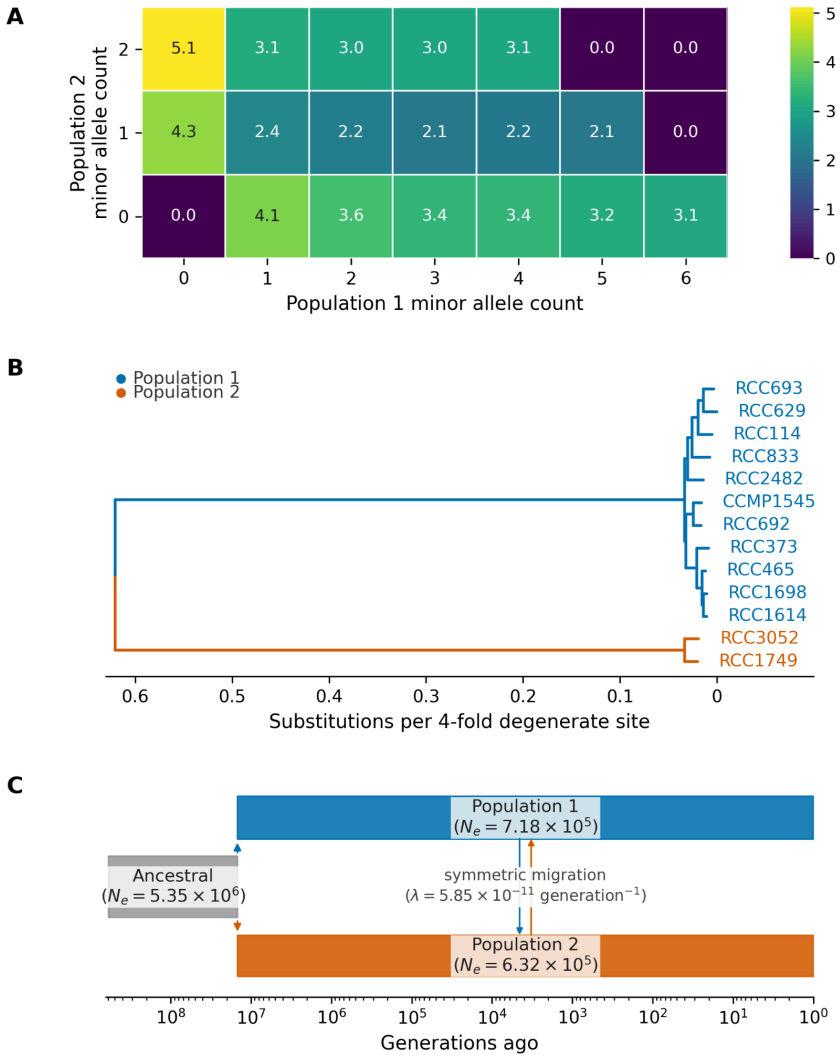


Figure 3.1 Population structure and demographic history of M. pusilla
(A) Folded joint two-dimensional allele-frequency spectrum for Populations 1 and 2 based on fourfold-degenerate SNPs; numbers indicate log-scaled allele counts and color intensity indicates the number of variants in each allele-count bin.(B) Maximum-likelihood phylogeny inferred from fourfold-degenerate SNPs, rooted using Population 2, showing the genetic separation between Populations 1 and 2. (C) Best-fitting two-population split-with-migration demographic model, illustrating the inferred population divergence and subsequent gene flow.

3.4.2 Comparison of two divergent M. pusilla assemblies

Despite their deep sequence divergence, the two populations retain extensive chromosome-scale synteny. Whole-genome alignment between the CCMP1545

(Population 1) and RCC1749 (Population 2) assemblies showed 90.7% average nucleotide identity, with 72% and 82% of the two genomes aligning, respectively. Nineteen of 21 chromosomes exhibited clear one-to-one syntenic relationships (Figure 3.2), indicating that most divergence has accumulated without large-scale chromosomal restructuring. The primary exception was chromosome 2, which contains the mating-type region, a non-recombining region spanning 1.68 Mb, or 77.5% of the chromosome. This region was extensively rearranged and incompletely aligned: 1.11 Mb (66.2%) was covered by one-to-one MUMmer alignments to the orthologous RCC1749 contig, distributed across 347 blocks, of which 166 (47.8%) were in reverse orientation. Because this chromosome is known to experience suppressed recombination and unusual evolutionary dynamics in Mamiellales (Benites et al. 2021), we excluded the mating-type region from subsequent genome-wide population genetic analyses.

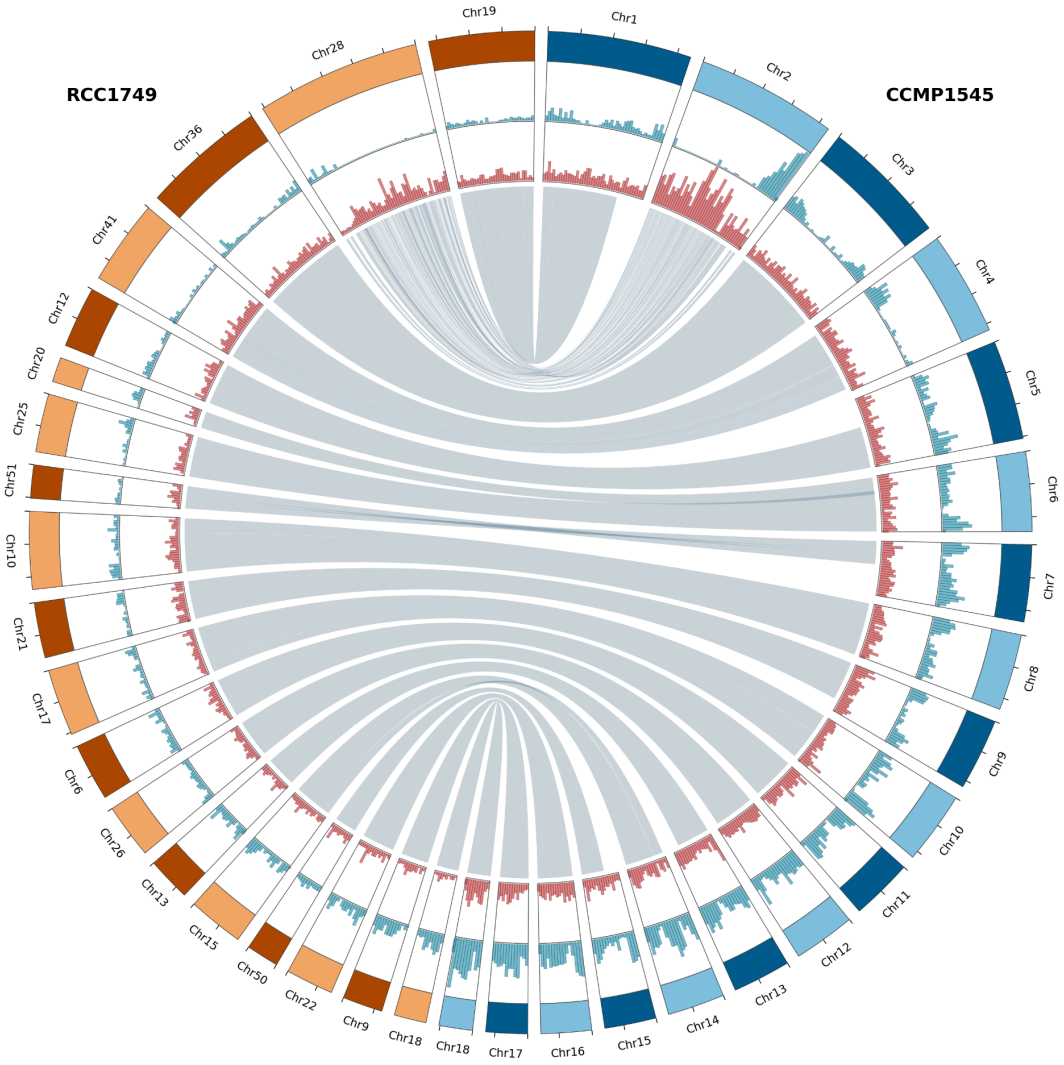


Figure 3.2 Whole-genome synteny and intron distributions between reference assemblies

Chromosomes from CCMP1545 (blue; right) and RCC1749 (orange; left) are arranged by their principal syntenic relationships. Gray ribbons connect merged MUMmer alignment blocks between assemblies. Histograms show introner counts (cyan, outer track) and canonical intron counts (red, inner track) in 50-kb windows.

3.4.3 Variation in Introner presence mirrors SNP-level phylogenetic patterns

Introner insertion polymorphism revealed extensive population differentiation between the two *M. pusilla* lineages. We developed a custom pipeline to identify orthologous introner insertions using flanking-sequence homology and refined

presence/absence calls with read-depth information to construct an introner genotype matrix (Methods). Across all isolates, we identified 6,066 orthologous introner loci, of which 4,706 occurred in Population 1 and 1,453 in Population 2. Remarkably, only 90 loci were shared between populations, while 4,613 and 1,360 were unique to Populations 1 and 2, respectively, indicating that the vast majority of introner insertion sites accumulated after the populations diverged. Population 1 also contained substantial within-population insertion polymorphism, including 619 segregating loci in addition to 3,421 fixed insertions, whereas Population 2 exhibited little detectable polymorphism likely due to the limited sample size. Finally, among the 90 shared loci, 64 contained conserved amino acid flanking sequences inferred to represent ancestral insertions retained across the population split, while 13 loci were identified as intergenic in Population 2, and 13 were identified as intergenic in both populations. Among the introner loci that were called as present or absent in all samples, we found 44 ancestral loci fixed in both populations (Figure 3.3a). We did not find any evidence of independent insertion events between the two populations.

Introners in *M. pusilla* cluster into six distinct families based on sequence composition and length (Figure 3.3b, see methods). Family 2 is by far the most common, accounting for 80.3% of all introners in Population 1. Families 1 and 3 are much less common in Population 1, representing 11.8% and 7.0% of loci, respectively, while families 4, 14, and 15 are rare, accounting for less than 1% of Population 1 introners. Population 2 shows a distinct family composition: family 2 is still the most abundant class, but represents a smaller fraction of loci (57.1%), while

family 1 is substantially enriched relative to Population 1 (31.1%). Family 3 is also modestly more common (11.6%). Thus, the major difference in introner composition between populations is an enrichment of family 2 in Population 1, and a corresponding shift toward family 1 in Population 2, consistent with lineage-specific changes in introner dynamics after divergence of the two populations.

Broad-scale patterns of introner accumulation are highly conserved between Population 1 and Population 2 despite extensive turnover of individual insertion sites, suggesting that largely conserved genomic features shape introner distributions in both lineages. Excluding shared ancestral introners, correlations in introner density were weak at fine scales (Spearman's $\rho = 0.16$, $p = 1.49 \times 10^{-102}$ in 1 kb windows) but increased steadily with window size, reaching $\rho = 0.69$ ($p = 5.36 \times 10^{-54}$) in 50 kb windows (Figure S4). Thus, although individual insertion events are largely population-specific, both populations have independently accumulated introners in the same broad genomic regions. Family 2, the dominant introner family in both genome assemblies, showed the strongest large-scale correlation ($\rho = 0.55$, $p = 6.44 \times 10^{-31}$ at 50 kb), whereas families 1 and 3 exhibited weaker but still positive correlations ($\rho = 0.26$, $p = 5.00 \times 10^{-7}$ and $\rho = 0.28$, $p = 3.59 \times 10^{-8}$, respectively).Together, these results suggest that conserved genomic features, such as chromatin organization or recombination landscape, shape where introners accumulate over evolutionary time.

3.4.4 Shared introner loci preserve deep population divergence

Introner loci reside within highly divergent genomic regions between Population 1 and Population 2, reinforcing the observation of limited genetic exchange. Divergence is high across every presence class: population-specific loci carry the most divergent flanks (mean Dxy = 0.145 and 0.171 for Population 1- and Population 2-specific loci, respectively), whereas loci shared between populations are generally less divergent (mean Dxy = 0.117; Mann–Whitney U, $p < 0.02$ against both population-specific classes) (Figure 3.4b). Thus, the shared introners are most consistent with ancestral insertions retained at a comparatively less-diverged subset of orthologous sites rather than with ongoing gene flow between the populations.

Introner sequences themselves display much different variation patterns from their surrounding genomic context. Even at ancestral, shared loci, the introner body is more diverged between populations than the genomic flanking regions. Across all callable ancestral loci, mean body Dxy reached 0.197, nearly double the divergence of their flanking regions (mean flank Dxy = 0.117; Mann–Whitney U, $p = 6.8 \times 10^{-8}$). This contrast indicates that although ancestral insertion sites have been retained since the population split, introner sequences have evolved substantially faster than their surrounding genomic context (primarily exonic sequence), consistent with weaker evolutionary constraint within the inserted element.

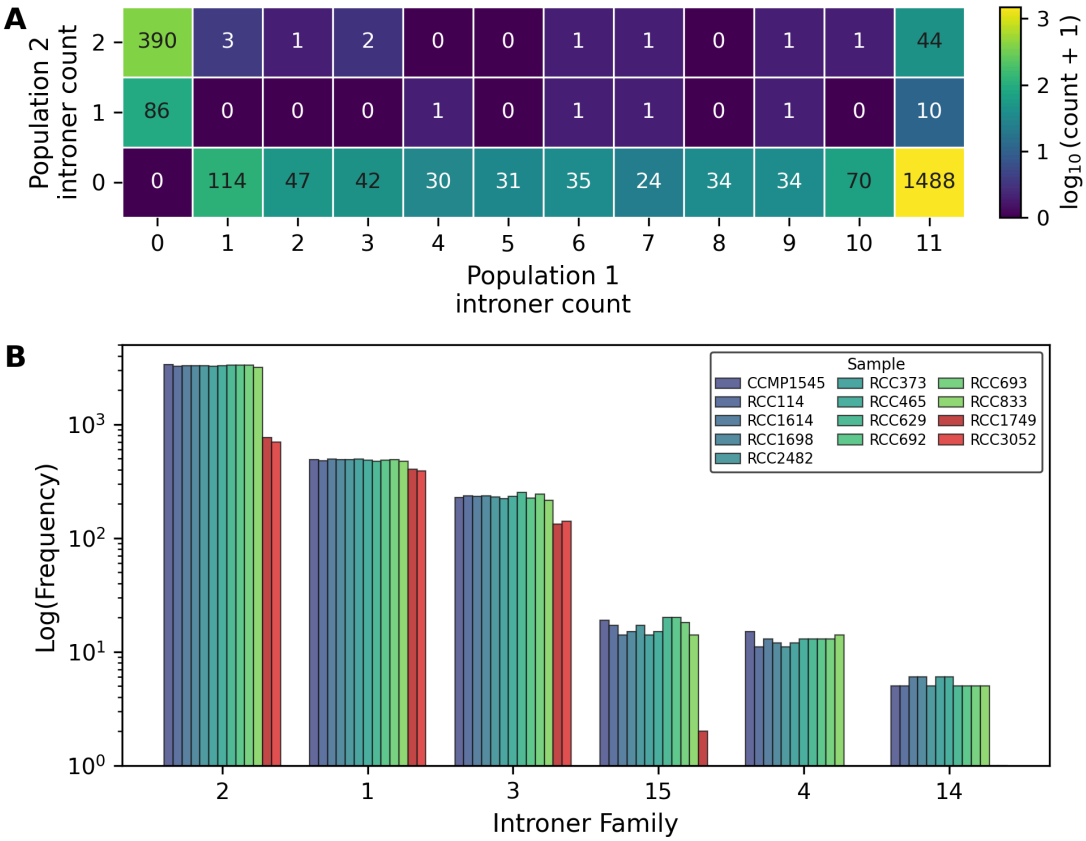


Figure 3.3 Introner allele-frequency spectra and family distributions
A.) Introner allele-frequency spectrum. Joint distribution of intron-present counts across Population 1 and Population 2 for loci that are callable in all samples. Cell labels indicate the number of loci; color represents $\log_{10}(\text{count} + 1)$. B.) Introner family distribution. Numbers of present introners from each family across sampled strains, displayed on a logarithmic scale. Colors identify individual strains.

3.4.5 Population 1 polymorphic intron loci are reflective of both gains and losses

Polymorphic introners within Population 1 display an allele frequency spectrum distinct from that expected under the standard infinite-sites model of sequence evolution. Polarizing presence/absence alleles toward the “absent” state, the intron SFS is strongly U-shaped, enriched for both rare-present and rare-absent alleles, whereas genome-wide synonymous and nonsynonymous SNPs follow the

expected L-shaped decline (Figure 3.4a). The two tails map onto the two halves of the introner life-cycle such that rare-present alleles are the signature of recent gains not yet risen in frequency, while rare-absent alleles (near-fixed for presence) are consistent with introners on a trajectory toward loss. This spectrum is therefore consistent with a dynamic process of recurrent introner gain and loss, rather than the accumulation of variants under a mutation process dominated by forward mutation with negligible back mutation.

Sequence diversity within introner bodies independently supports this gain-and-loss interpretation of the site-frequency spectrum. If rare-present alleles largely represent recent gains, their introner bodies should exhibit less diversity because less time has elapsed for variation to accumulate following insertion. Consistent with this expectation, polymorphic Population 1 introners had lower median π than fixed introners (0.0015 versus 0.0034; Mann–Whitney U test, two-sided $p = 4.9 \times 10^{-16}$) and fewer segregating sites per callable base (median 0 versus 0.0116; $p = 1.2 \times 10^{-85}$) (Figure S2). These patterns are consistent with recent insertions comprising a substantial fraction of polymorphic introners.

A molecular evolution model fit directly to the presence/absence spectrum supports the conclusion that segregating introner polymorphism has been shaped by both recurrent gain and loss. We fit a gain/loss mixture model to the full Population 1 introner SFS, conditioned on the demographic history inferred from 4D sites, so that the fitted rates are corrected for the population's expansion history (see methods). Under the best-fitting 4D-site demographic model, the introner SFS was best

explained by both gain and loss components, where gain outpaced loss (per-locus scaled rates 0.035 vs. 0.023), and the model resolved the U-shape into its two origins. Rare-present alleles are overwhelmingly gains (P(gain) = 0.96 for present doubletons), while high-presence alleles are predominantly losses (P(loss) = 0.73 for absent doubletons). These results demonstrate that the unusual introner frequency spectrum arises from the joint action of recurrent gain and loss processes operating on a dynamically changing set of insertion loci.

Applying the same gain/loss model to canonical introns shows that the U-shaped spectrum is a specific feature of introners, suggesting that unique evolutionary processes are acting on introners. In this class of presumably ancestral introns, the gain component was effectively absent ($\theta_{gain} \approx 0$), indicating that detectable intron acquisition does not contribute measurably to their frequency spectrum over this evolutionary timescale. In contrast, loss showed comparable dynamics between the two classes, with a per-locus loss rate for canonical introns (0.014) within the same order of magnitude as that estimated for introners (0.023). Thus, while introner polymorphism is distinguished by recurrent gain, both introner and canonical introns appear subject to similar loss processes, consistent with a shared mechanism of intron deletion or genomic erosion.

After accounting for the relative numbers of loci subject to gain and loss, the aggregate rate of introner acquisition ($\theta_{gain} = 202.7$) is broadly comparable to the aggregate rate of introner loss ($\theta_{loss} = 131.1$). Canonical introns demonstrate a similar overall loss rate per locus ($\theta_{loss\_canonical} = 0.014$), and we note that the

sum of canonical intron and introner loss rates ($\theta_{loss\_overall} = 193.3$) is quite similar to the rate of introner gain. This balance suggests that Population 1 introns may be maintained near a short-term dynamic equilibrium, where recurrent gains replenish insertions removed by ongoing loss. Though the pronounced differences in overall introner content between populations 1 and 2 indicates that longer-term dynamics are not close to an equilibrium. Furthermore, short-term stability in total introner abundance does not imply stability in intron composition. Continued turnover of individual loci will gradually replace ancestral unique introns with introner-derived sequences, progressively reshaping the genomic intron repertoire even in the absence of net intron expansion.

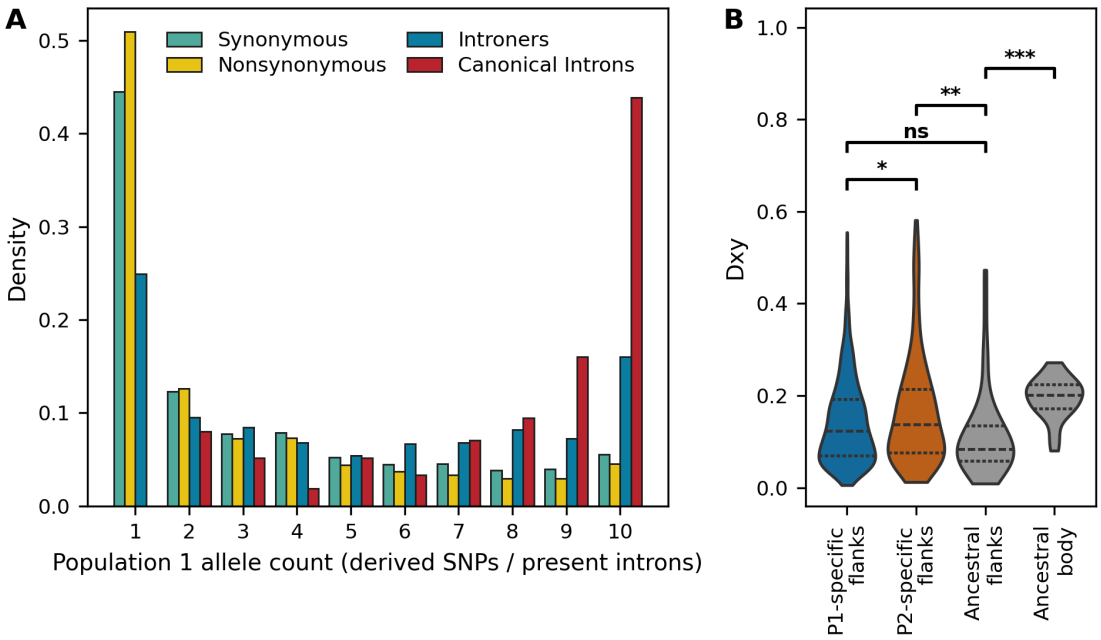


Figure 3.4 Allele-frequency spectra and introner-locus sequence diversity
A.) Population 1 allele-frequency spectra. Normalized distributions of derived synonymous and nonsynonymous SNPs, introner presence, and regular intron presence across segregating allele-count classes. SNPs were polarized using Population 2 as the outgroup. B.) Introner locus sequence diversity. Violin plots show Dxy for population-specific 200bp flanks and ancestral introner flanks and bodies. Horizontal lines indicate quartiles; significance bars show two-sided Mann–Whitney U tests (*($p<0.05$), **($p<0.01$), *** ($p<0.001$)).

3.4.6 Introner insertions are positively associated with recombination but not recent positive selection

Polymorphic introners show no evidence of association with recent selective sweeps, but instead behave more similarly as nearly neutral variants shaped by local genomic context. A SweepFinder2 scan of Population 1 SNPs did not find fixed or polymorphic introners enriched in any candidate sweep regions (0.43% overlap versus 0.94% expected by permutation; P = 0.210), while a supplementary scan of all SNPs also found no such enrichment (0.43% versus 0.91%; P=0.272), thus indicating that strong positive selection does not appear to be a major determinant of their

genomic distribution. This implies that strong positive selection does not disproportionately favor introner insertions.

Introners are associated with higher recombination rates. We used pyrho (Spence and Song 2019) to estimate a recombination landscape for *M. pusilla* using polymorphic SNPs (see methods). Fixed introners in population 1 are positively correlated with increasing recombination rates (17.1% elevated; P = 0.005; MWU), and polymorphic introners are even more strongly associated with elevated location recombination rates (55.4% elevated; $P = 6.89 \times 10^{-6}$; MWU). We also observe a similar enrichment for all introners in population 2 after lifting over recombination rate estimates (34.5% elevated; $P = 7.11 \times 10^{-5}$; MWU, see Methods). The excess in polymorphic introners likely reflects more ancient coalescence in highly recombining regions due to the effect of linked selection (R. B. Corbett-Detig et al. 2015; Charlesworth et al. 1993; Begun and Aquadro 1993), as SNPs are also similarly enriched compared to monomorphic 4D sites (22.4% elevated; $P = 1.91 \times 10^{-78}$; MWU). The elevated recombination rates observed for fixations and polymorphisms may indicate that recombination enhances the rate of insertion of new introner elements, and may help explain why no introners are observed in the non-recombining mating-type region (reported previously in XXX-XXX).

3.4.7 The mechanistic basis of introner loss is not related to reverse transcription and recombination

Further analysis did not identify a specific recombination-dependent mechanism of introner loss. One proposed mechanism of intron loss is 3′-biased reverse-transcriptase-mediated intron loss, whereby a spliced mRNA is reverse

transcribed and the resulting intron-free cDNA replaces the homologous genomic sequence through recombination (Derr and Strathern 1993). Because reverse transcription begins at the 3′ end of the transcript and may terminate before reaching the 5′ end, this mechanism predicts preferential loss of introns near gene 3′ ends. We compared recent-loss loci with fixed loci using two-sided Mann–Whitney U tests of CDS-normalized distance from the gene 3′ end. Neither recent-loss introners ($P = 0.422$) nor recent-loss conventional introns ($P = 0.368$) showed significant 3′ enrichment, providing no evidence that this mechanism is a major source of the observed loss polymorphisms. More broadly, genes carrying polymorphic introners did not differ significantly in expression from genes containing only fixed introners in any of the three strains examined (permutation tests, $P = 0.219$–0.923), and fixed and polymorphic introners showed similar splice-site usage, with nearly all RNA-supported junctions employing canonical GT–AG or GC–AG boundaries. The only detectable structural difference was a modest reduction in introner length, with polymorphic introners having a median length 4 bp shorter than that of fixed introners (permutation test, $P = 0.0002$). Taken together, these results indicate that recent introner losses lack a consistent positional, transcriptional, or splice-site signature and are unlikely to be driven predominantly by a single reverse-transcription-mediated mechanism.

The parallel absence of a 3′ positional bias among recent losses of both introners and canonical introns, together with their broadly comparable per-locus loss intensities inferred from the site-frequency spectrum, suggests that the two classes

may be removed by overlapping evolutionary processes. Although these observations do not identify a specific molecular mechanism or demonstrate that the loss processes are similar, they imply that introner loss may largely draw on the same deletion, repair, or genomic-erosion processes that remove canonical introns. Introner-specific mobility may therefore distinguish the gain dynamics of these elements, whereas their loss may be governed primarily by more general mechanisms of intron deletion.

3.4.8 Introner Presence Affects both Transcript Isoform Diversity and Expression

Within genes, differences in introner polymorphism are associated with corresponding differences in transcript isoform diversity. Using long-read R2C2 sequencing across three *M. pusilla* strains (CCMP1545 and RCC1614 from Population 1 and RCC1749 from Population 2), we quantified transcript isoforms without relying on short-read inference and modeled introner number within orthologous genes while accounting for strain and sequencing depth. A gene-conditioned Poisson model found that one additional introner was associated with 24.0% more alternative isoforms beyond the baseline transcript (rate ratio = 1.24, 95% CI: 1.12–1.38, $p = 5.7 \times 10^{-5}$)(Figure 3.5a, Figure 3.6). Conversely, one fewer introner corresponded to 19.3% fewer alternative isoforms (rate ratio = 0.81, 95% CI: 0.73–0.90) (Figure 3.5a). Thus, within-gene changes support a relationship between introner number and isoform richness, whereas the cross-gene burden association appears sensitive to underlying gene architecture.

Polymorphic and recent introner gains are associated with lower gene expression. In a negative-binomial GLM of short-read counts with log-library size as

an offset, each current introner was associated with a 36.0% reduction in expected expression after accounting for strain, gene span, GC content, and transcript isoform number (Figure 3.5b). Genes carrying reference-relative introner gains exhibited an additional reduction in expression beyond that explained by current introner burden, whereas genes with introner losses showed the opposite trend. Together, these results indicate that introner turnover is concentrated among genes with relatively low expression. Importantly, because the strongest association is observed for polymorphic introner gains within orthologous genes, these findings are more consistent with introner acquisition contributing to reduced expression than with an intrinsic tendency for some genes to simply harbor more introners. Although this analysis cannot prove causality or exclude the possibility that introners preferentially insert into genes that are already weakly expressed, the observed pattern supports the hypothesis that introner insertion itself can reduce gene expression.

Polymorphic introners splice as efficiently as fixed introners, indicating both that they are broadly compatible with host splicing and that a baseline level of splicing efficiency is likely required for their persistence in the population. After accounting for gene expression and replicate effects using a beta-binomial model, we find that polymorphic and fixed-present introners had nearly identical retention rates (0.122 and 0.118, respectively; marginal-contrast $P = 0.39$). Canonical introns had a lower retention rate (0.08, ($p = 0.015$)), suggesting greater apparent splicing efficiency than either introner class (Figure 3.5d). Thus, although introners overall may splice somewhat less efficiently than conventional introns, we find no evidence

that polymorphic introners impose a greater splicing-associated burden than fixed introners, arguing against inefficient splicing—and consequent purifying selection—as an explanation for their segregating frequencies in Population 1.

Introner-containing genes are enriched for isoforms that contain premature stop codons and are expected to be targeted by nonsense-mediated decay (Kurosaki et al. 2019), offering a route by which introners could depress host expression. *M. pusilla* maintains an unusually balanced transcriptome in which 70.1% of multi-isoform genes show high Shannon evenness ($J' > 0.7$; median, 0.82), with secondary isoforms often maintained near the abundance of the primary transcript (Condon et al. 2026). Isoform diversity scales with isoform count ($r = 0.72$), so introner-associated isoform gains translate directly into greater transcript diversity. A fraction of that diversity is non-productive: 3.7% of multi-exon isoforms were predicted NMD substrates. In a binomial GLM clustered by gene, isoforms from introner-containing genes had 2.44-fold higher odds of containing a premature termination codon, consistent with potential NMD targeting (95% CI, 1.61–3.70; $p = 2.4 \times 10^{-5}$) (Figure 3.5c). Introners are therefore associated not only with greater isoform diversity but also with a larger proportion of transcripts predicted to undergo NMD, providing a potential mechanistic link between introner presence and apparently reduced gene expression.

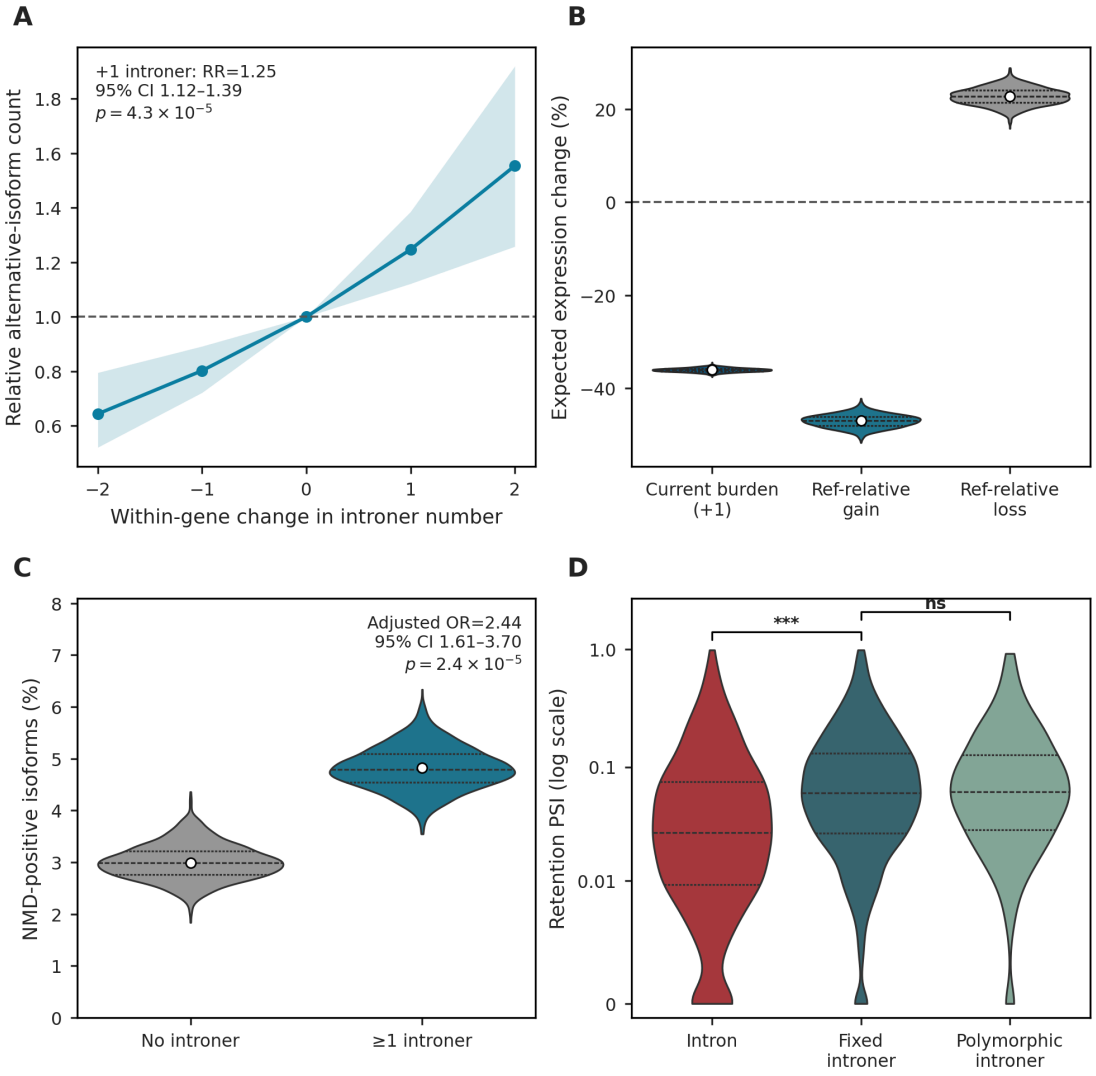
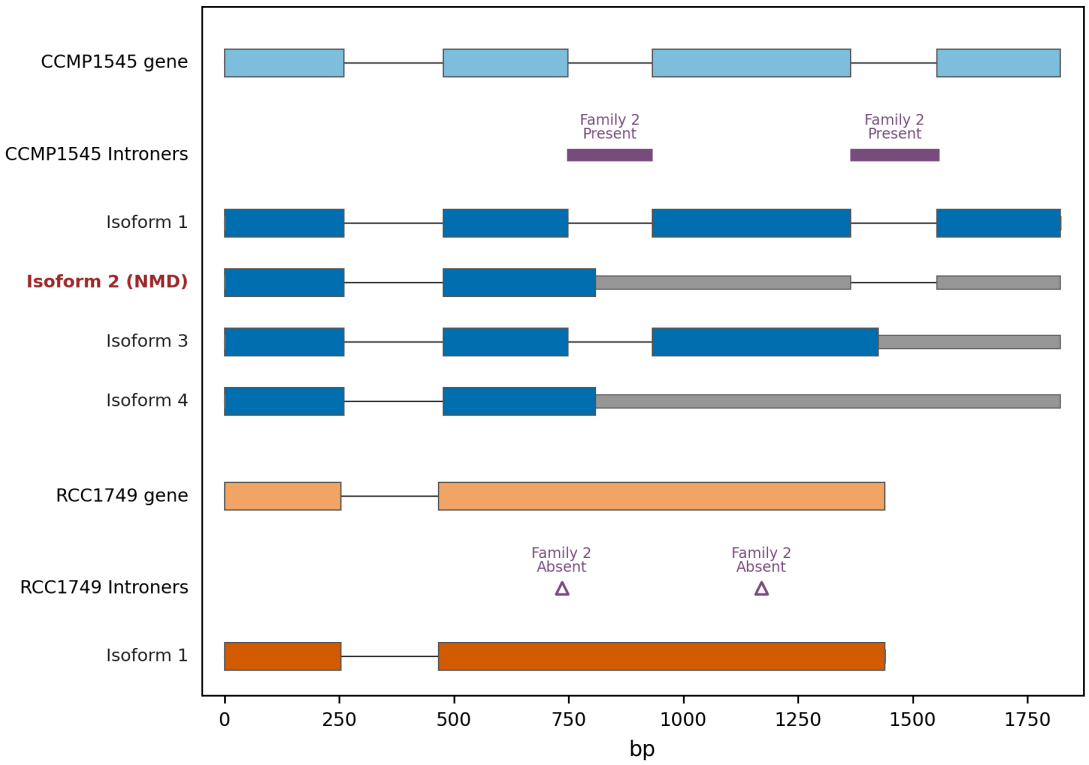


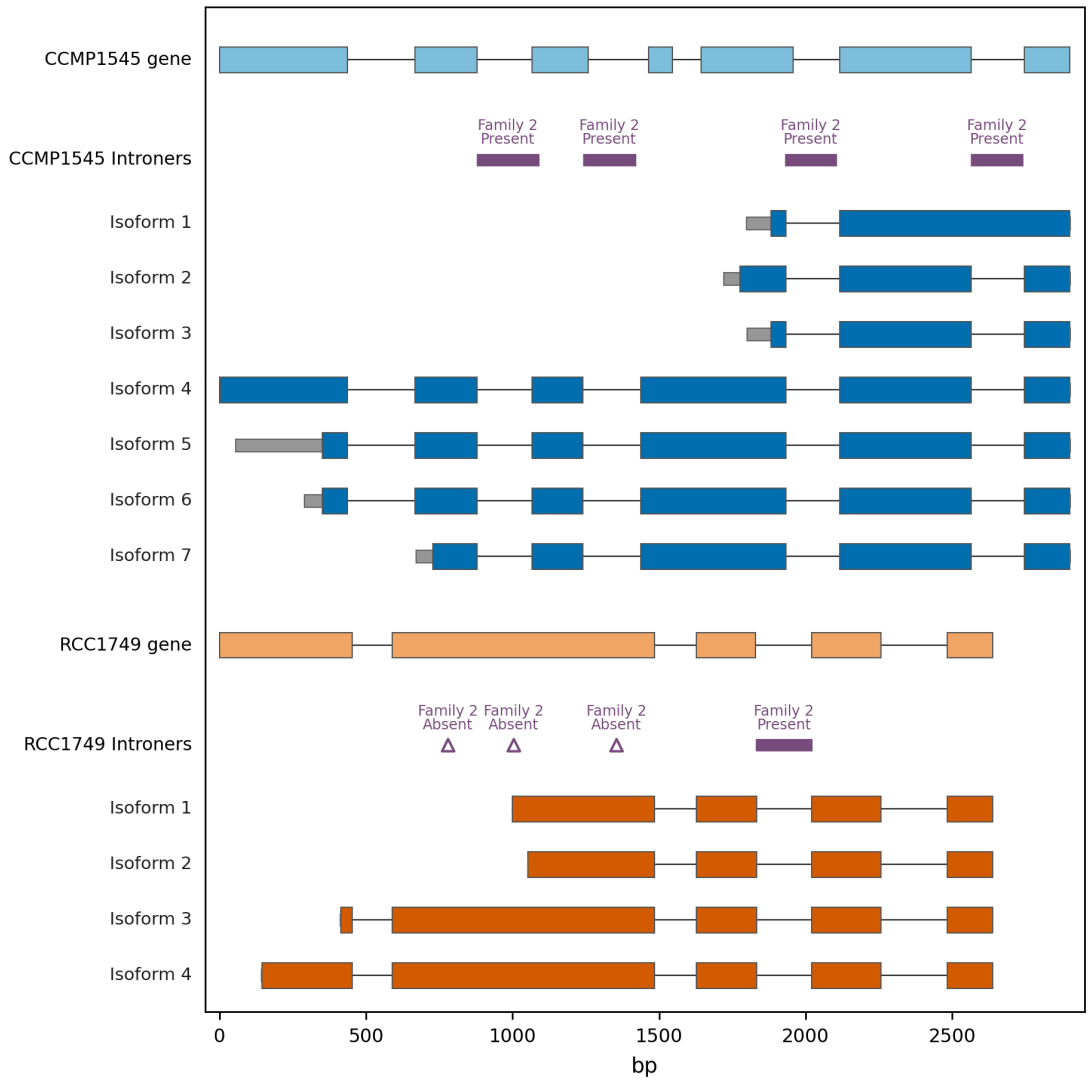
Figure 3.5 Functional associations of introners with transcript processing and gene expression

(A) Predicted relative alternative-isoform count as a function of within-gene changes in intron number; shading indicates the 95% confidence interval. (B) Estimated expression changes associated with current intron burden and reference-relative gains or losses; error bars show 95% confidence intervals. (C) Observed fraction of predicted NMD-targeted isoforms in genes with and without introners; error bars show bootstrap 95% confidence intervals, with the adjusted odds ratio indicated. (D) Intron-retention PSI distributions for conventional introns and fixed or polymorphic introners. Higher PSI indicates lower splicing efficiency; internal lines show quartiles. Polymorphic and fixed introners did not differ significantly in retention PSI.

Figure 3.6 Introner polymorphism and transcript isoform diversity at representative genes

Reference gene models and high-confidence R2C2 transcript isoforms are shown for CCMP1545 and RCC1749. Coding regions are colored by strain (CCMP1545 blue, RCC1749 orange), untranslated regions are grey, and introns are represented by connecting lines. Plum bars mark present introners, whereas triangles represent absent introner loci in RCC1749.


CCMP1545 gene
CCMP1545 Introners
Family 2
Present
Family 2
Present
Isoform 1
Isoform 2 (NMD)
Isoform 3
Isoform 4
RCC1749 gene
RCC1749 Introners
Family 2
Absent
Family 2
Absent
Isoform 1
0
250
500
750
1000
1250
1500
1750
bp



3.4.9 Introners Display Significant Gene Enrichment and Depletion Patterns

Introner-containing genes are depleted for core ribosomal and transcription-associated functions. GO enrichment analyses of introner-containing genes, evaluated against a CDS-length-weighted permutation null, revealed strong and consistent depletion of ribosomal subunits and structural constituents, ribonucleoprotein complexes, DNA-templated transcription, transcription regulator activity, and RNA polymerase II-associated terms (empirical FDR < 0.001). This

depletion was observed across the full introner set and independently among Population 1-fixed, Population 2-fixed, and family-2 introners, indicating that the functional bias is shared across lineages and introner classes. Together with the observed association between polymorphic introner gains and reduced gene expression, these patterns support a model in which introners are preferentially retained in genes where modest reductions in expression are tolerated or even beneficial, while being selectively eliminated in genes requiring tightly constrained transcriptional output.

Functional enrichment among introner-containing genes differs markedly between introner families. After CDS-length correction, introner-containing genes were enriched for catalytic and oxidoreductase activity, small-molecule metabolism, vacuolar transport, and protein localization to the vacuole. However, much of this signal was family-specific: family-1 introners were enriched in cytokinesis, cell septum assembly, Ras signaling, and mitotic-exit regulation terms, whereas family-2 introners were enriched in SNARE-complex assembly and vesicle-mediated transport. These distinct functional associations suggest that the genomic distribution of introners is not governed by a single accumulation process, but instead reflects the independent evolutionary histories of individual introner families.

Population-specific introner distributions may contribute to ecological differentiation by disproportionately affecting genes involved in vitamin-dependent and redox metabolism. Genes containing introners fixed in both population 2 isolates were enriched for water-soluble vitamin metabolism, including genes associated with

thiamine, riboflavin, biotin, and ascorbate pathways (3.47-fold enrichment; $p = 7.87 \times 10^{-5}$). Because these compounds function as cofactors or redox metabolites in central metabolism, this enrichment raises the possibility that introner turnover has contributed to population differences in the regulation or plasticity of cofactor-dependent metabolism (McRose et al. 2014). Such differences could influence responses to environmental vitamin availability, microbial metabolite exchange, or light- and oxidative-stress regimes. However, we caution that introner presence does not itself demonstrate altered gene function, and the small number of population 2 isolates make this an ecological hypothesis rather than evidence of adaptive differentiation.

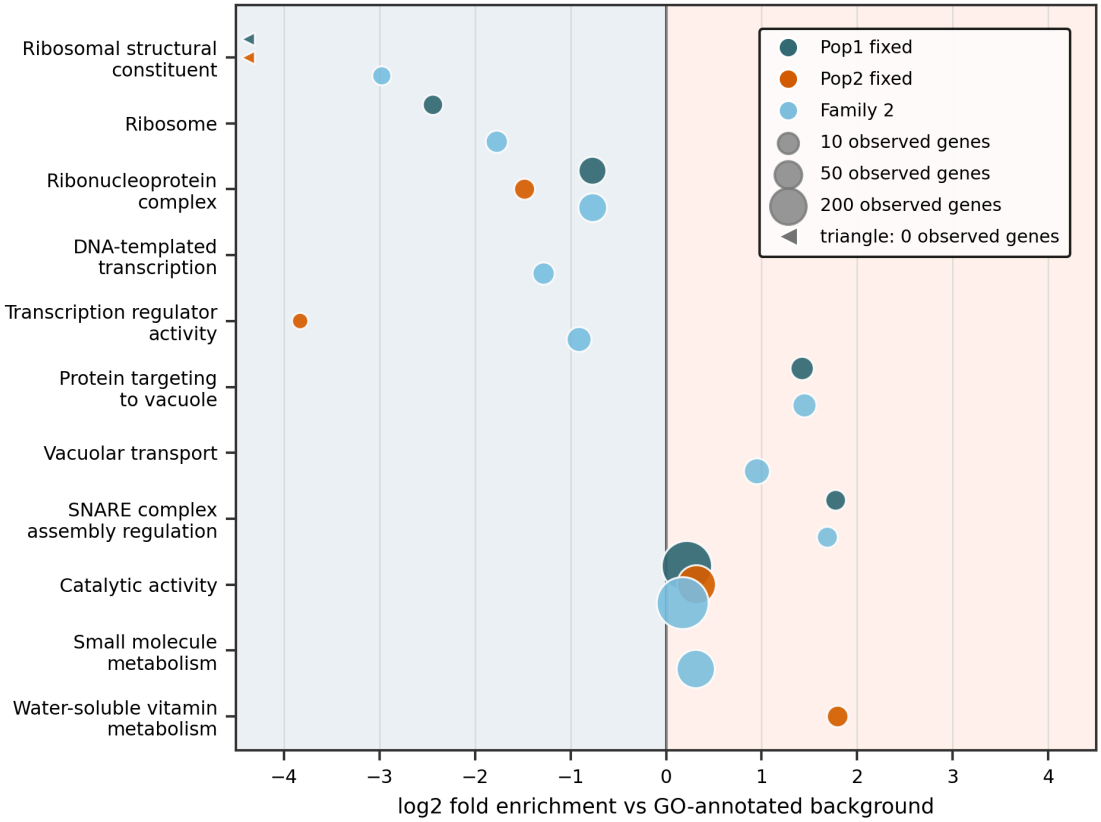


Figure 3.7 Gene Ontology enrichment among introner-associated gene groups
Representative GO terms are shown for genes containing Population 1 fixed introners, Population 2 fixed introners, or Family 2 introners. The x-axis gives $\log_2$ fold enrichment relative to the GO-annotated background; positive and negative values indicate enrichment and depletion, respectively. Point area scales with the number of observed genes, while triangles denote terms with no observed genes.

3.5 Conclusion

Taken together, our results show that introner evolution in *M. pusilla* is shaped by recurrent mobility filtered through genomic context and host fitness. The two deeply diverged populations contain markedly different and mostly non-overlapping orthologous loci, yet introners accumulate in similar broad-scale genomic neighborhoods. Within Population 1, the introner allele-frequency spectrum and demographic modeling reveal ongoing gain and loss. However, the absence of

selective-sweep signatures and the similar recombination environments of putative gains and losses argue against recombination as the direct driver of turnover. Instead, open, recombinogenic, and genetically diverse regions appear to provide permissive environments in which introner polymorphisms can arise and persist.

The persistence of introners is constrained by their functional consequences. Introners increase transcript isoform diversity while reducing gene expression and enriching for transcripts likely targeted by nonsense-mediated decay. Their depletion from core ribosomal and transcription-associated genes further indicates that purifying selection removes introner insertions from genes that are especially sensitive to transcriptional or splicing disruption. The introner landscape therefore records both element activity and host filtering, whereby insertions compatible with efficient splicing and located in more tolerant functional contexts are retained, and more costly insertions are eliminated.

By capturing introner gain and loss within a single species, M. pusilla offers a tractable system for observing spliceosomal-intron evolution in progress. Broader sampling of the introner-poor population, direct measurement of new insertions across generations, and experimental tests of chromatin and recombination effects will be needed to distinguish insertion bias from selective retention and to identify the factors that trigger lineage-specific introner expansions.

Supplementary

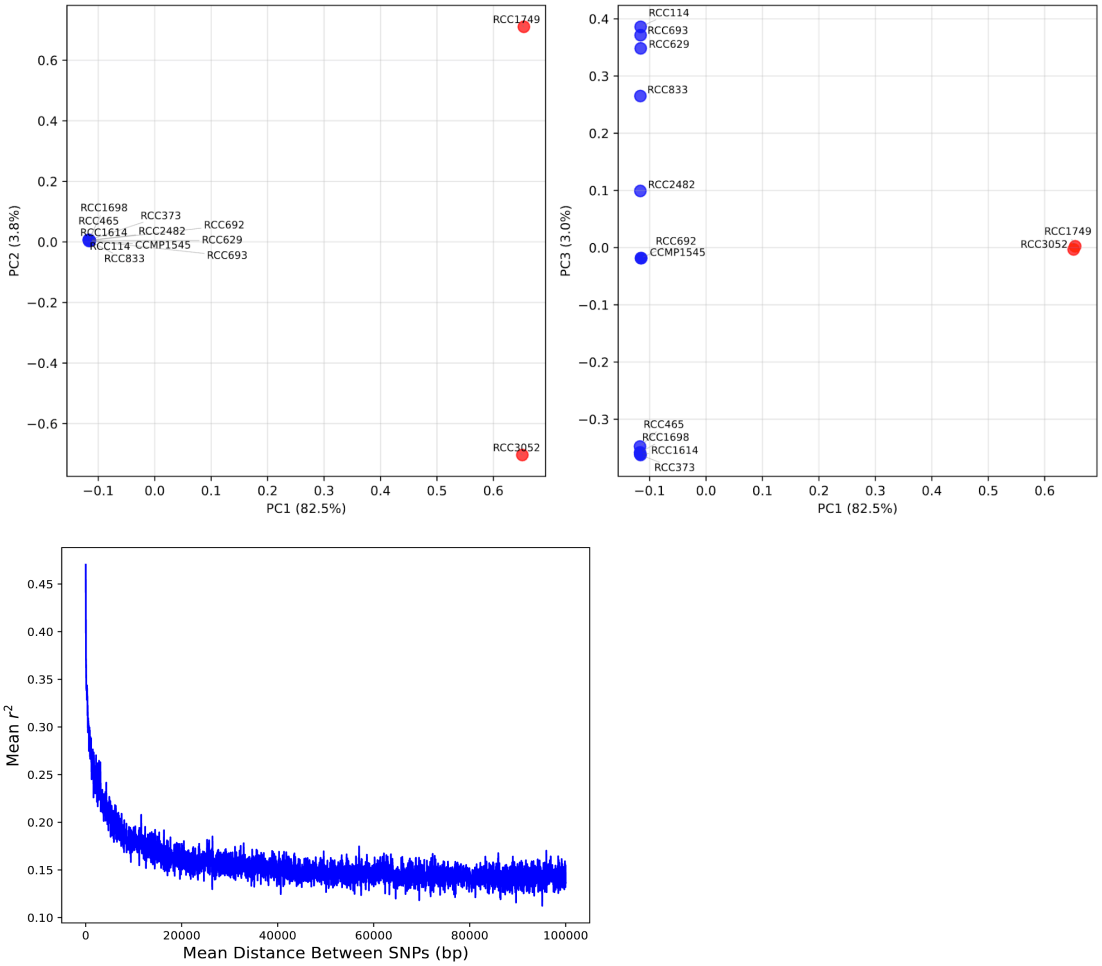


Figure S1 Population structure and linkage disequilibrium in *M. pusilla*
(A) Principal component analysis of *M. pusilla* isolates. PCA across all 13 isolates using 11,085 LD-pruned SNPs at fourfold-degenerate sites, excluding the mating-type region. Panels show PC1 versus PC2 and PC1 versus PC3; axes indicate the proportion of variance explained. Blue and red points represent Population 1 and Population 2, respectively. (B) Linkage disequilibrium decay within Population 1. Mean pairwise linkage disequilibrium (r^2) between callable, polymorphic 4D SNPs separated by up to 100 kb across the 11 Population 1 isolates, excluding the mating-type region. Values were averaged in 10-kb distance bins, revealing a decline from a mean (r^2) of approximately 0.23 in the first bin to a background of approximately 0.14 at longer distances.

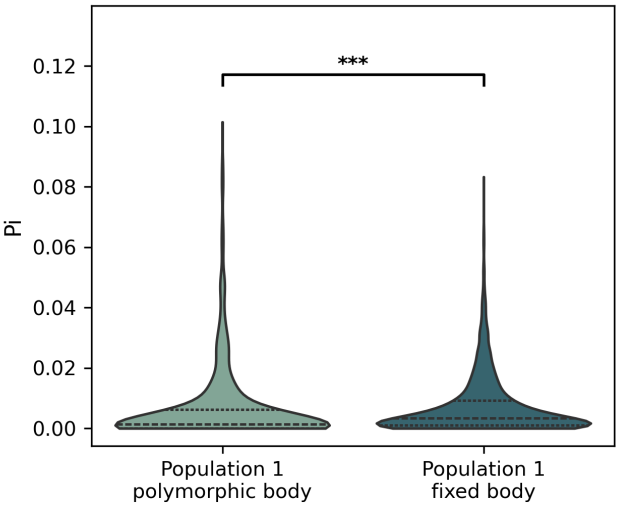


Figure S2 Sequence diversity in polymorphic and fixed introner bodies
Violin plots compare nucleotide diversity (π) between polymorphic and fixed introner bodies in Population 1. Polymorphic introners have lower diversity than fixed introners (two-sided Mann–Whitney U test, $p = 4.9 \times 10^{-16}$).

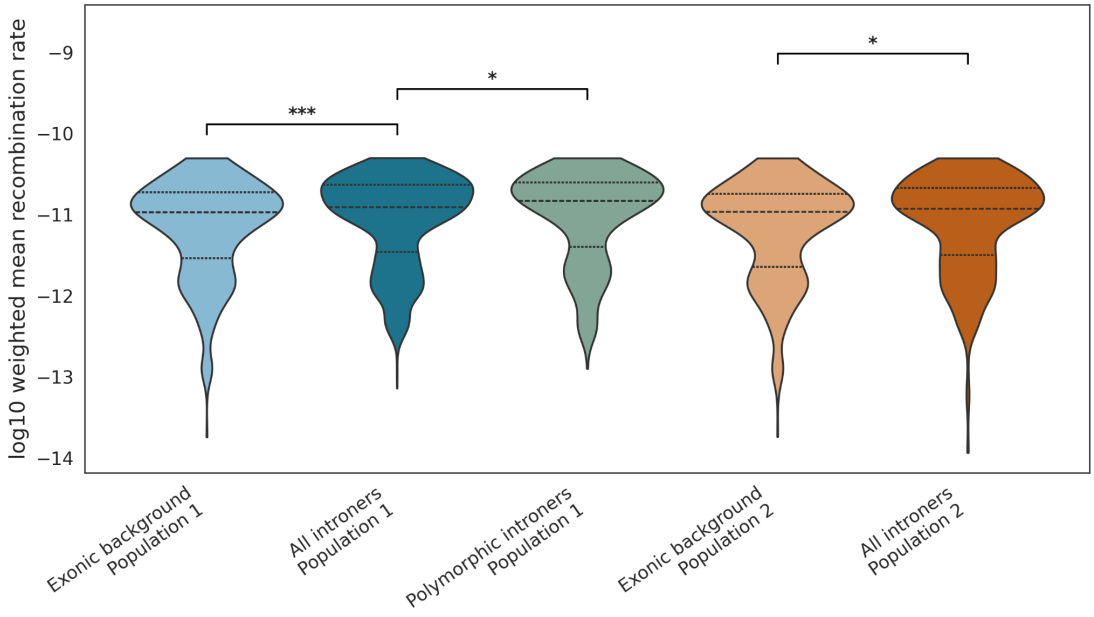


Figure S3 Recombination rates across introner-associated genomic categories
Violin plots show $\log_{10}$ weighted mean recombination rates in selected 5-kb exonic-background and introner-associated windows from Populations 1 and 2. Horizontal lines indicate quartiles; significance bars show Bonferroni-corrected two-sided Mann–Whitney U tests (*($p<0.05$), **($p<0.01$), *** ($p<0.001$)).

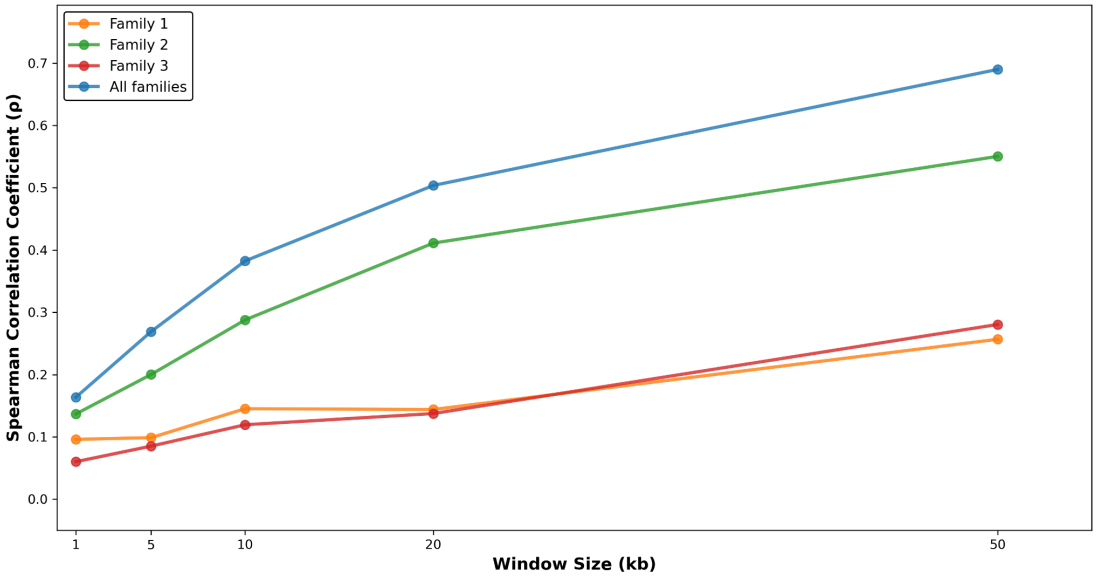


Figure S4 Correlation of introner density between reference genomes
Spearman's rank correlations between introner densities in corresponding MUMmer-aligned genomic windows of CCMP1545 and RCC1749 at window sizes of 1, 5, 10, 20, and 50 kb. Correlations are shown for all introner families combined and separately for families 1–3; families with insufficient data were omitted.

Table S1: *M. pusilla* isolate metadata.

Strain	Sampling Ocean	Sampling regional sea	Sampling site	Sampling country	lat.	lon.	Sampling date	Date entered catalog	Group
RCC2482	Mediterranean Sea	Tyrrhenian Sea	Gulf of Naples	Italy	40.79	14.34	1993-06-01	2011-04-14	1
RCC833	Atlantic Ocean	Gulf of Mexico	USA	USA	27.7	-91.3	2004-04-20	2005-03-10	1
RCC114	Atlantic Ocean	NA	East coast USA	USA	41.52	-70.67	1964-06-18	1998-11-01	1
RCC693	Atlantic Ocean	North Sea	Germany	Germany	54.18	7.9	2001-01-17	2003-10-06	1
RCC629	Atlantic Ocean	North Sea	Germany	Germany	54.18	7.9	2001-01-17	2002-10-22	1
CCMP1545	Atlantic Ocean	English Channel	near Plymouth	UK	50.36	-4.17	1950-01-01	2005-07-04	1
RCC692	Atlantic Ocean	North Sea	Germany	Germany	54.18	7.9	2001-06-19	2003-10-06	1
RCC373	Atlantic Ocean	Baltic Sea	Norway	Norway	58.18	9.1	NA	2001-03-01	1
RCC465	Atlantic Ocean	English Channel	Brtianny coast	France	48.75	-3.95	2001-06-13	2002-03-04	1
RCC1614	Atlantic Ocean	North Sea	Norway	Norway	56.98	3.98	2007-07-14	2008-07-31	1
RCC1698	Atlantic Ocean	North Sea	Norway	Norway	56.32	4.32	2007-07-24	2009-01-23	1
RCC3052	Pacific Ocean	South East Pacific	Chile upwelling	Chile	-33.38	-71.7	2012-01-17	2013-04-15	2
RCC1749	Atlantic Ocean	English Channel	Brittany Coast	France	48.72	-3.97	2007-07-18	2009-05-26	2

Table S2: Significant GO terms enriched or depleted in *M. pusilla* groups or in introner family groups.

Term name	Introner class	Observed genes	log2 fold enrichment	Empirical p	Empirical FDR
structural constituent of ribosome	Pop1 fixed	0		1.00E-06	1.06E-05
	Pop2 fixed	0		1.64E-02	2.33E-02
	Family 2	5	-2.97	1.60E-04	3.66E-04
ribosome	Pop1 fixed	8	-2.47	2.70E-05	1.07E-04
	Family 2	15	-1.77	1.48E-03	2.51E-03
ribonucleoprotein complex	Pop1 fixed	56	-0.78	7.01E-04	1.28E-03
	Pop2 fixed	10	-1.50	8.47E-04	1.53E-03
	Family 2	65	-0.76	2.54E-04	5.25E-04
DNA-templated transcription	Family 2	15	-1.28	1.00E-06	1.06E-05
transcription regulator activity	Pop2 fixed	1	-3.84	1.90E-05	8.20E-05
	Family 2	30	-0.91	9.30E-05	2.27E-04
protein targeting to vacuole	Pop1 fixed	20	1.39	3.30E-05	1.15E-04
	Family 2	24	1.45	1.00E-06	1.06E-05
vacuolar transport	Family 2	37	0.95	1.10E-05	5.81E-05
regulation of SNARE complex assembly	Pop1 fixed	9	1.74	5.10E-05	1.45E-04
	Family 2	10	1.69	3.00E-05	1.08E-04
catalytic activity	Pop1 fixed	877	0.21	1.60E-05	7.07E-05
	Pop2 fixed	281	0.33	1.30E-05	6.33E-05
	Family 2	989	0.18	3.15E-04	6.11E-04
small molecule metabolic process	Family 2	256	0.31	1.00E-06	1.06E-05

Term name	Introner class	Observed genes	log2 fold enrichment	Empirical p	Empirical FDR
water-soluble vitamin metabolic process	Pop2 fixed	13	1.90	5.00E-06	3.80E-05
oxidoreductase activity	Pop2 fixed	18	1.42	5.40E-05	1.51E-04
regulation of cytokinetic process	Family 1	4	3.78	1.00E-05	5.43E-05
regulation of cell septum assembly	Family 1	4	3.78	1.00E-05	5.43E-05
division septum assembly	Family 1	4	3.78	2.20E-05	9.09E-05
regulation of actomyosin structure organization	Family 1	4	3.52	3.40E-05	1.15E-04
Ras protein signal transduction	Family 1	5	3.10	4.40E-05	1.34E-04
regulation of exit from mitosis	Family 1	4	3.52	5.70E-05	1.57E-04

Table S3: Oligonucleotides used in cDNA library construction. Index corresponds to a random 8 nucleotide sequence unique to each replicate library.

Oligo ID	Sequence
TSO-Smart-seq2	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG
Oligo_dT_Index1	AAGCAGTGGTATCAACGCAGAGT [Index] ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index2	AAGCAGTGGTATCAACGCAGAGT TATCTGACCT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index3	AAGCAGTGGTATCAACGCAGAGT ATATGAGACG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index4	AAGCAGTGGTATCAACGCAGAGT CTTATGGAAT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index5	AAGCAGTGGTATCAACGCAGAGT TAATCTCGTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index6	AAGCAGTGGTATCAACGCAGAGT GCGCGATGTT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index7	AAGCAGTGGTATCAACGCAGAGT AGAGCACTAG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index8	AAGCAGTGGTATCAACGCAGAGT TGCCTTGATC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index9	AAGCAGTGGTATCAACGCAGAGT CTACTCAGTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index10	AAGCAGTGGTATCAACGCAGAGT TCGTCTGACT ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index11	AAGCAGTGGTATCAACGCAGAGT GAACATACGG ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
Oligo_dT_Index12	AAGCAGTGGTATCAACGCAGAGT CCTATGACTC ACTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
ISPCR primer	AAGCAGTGGTATCAACGCAGAGT
ISPCR_Anneal_Splint_1_F	ACTCTGCGTTGATACCACTGCTT GCACGCACCGACAAACTCTGGCCGATTGTACTTTCTTATAAGG CGTAACACTAGACCATATCGTTTCTATAGATTAATTATGGCGA GAATTGCTAGCGAACTAGACAATTTTCGAAATAATCCTTTTTA TAT AAGCAGTGGTATCAACGCAGAGT
ISPCR_Anneal_Splint_1_R	ACTCTGCGTTGATACCACTGCTT ATATAAAAAGGATTATTTCGAAAATTGTCTAGTTCGCTAGCAA TTCTCGCCATAATTAATCTATAGAAACGATATGGTCTAGTGTT ACGCCTTATAAGAAAGTACAATCGGCCAGAGTTTGTCGGTGCG TGC AAGCAGTGGTATCAACGCAGAGT

Conclusion

In this dissertation, I explored how genomic conflict shapes evolution through two related routes: biased inheritance and the proliferation of mobile elements. In the first chapter, I showed that pollen-acting segregation distortion is already widespread among diverging *Arabidopsis* populations and becomes more frequent and pronounced as parental divergence increases. The absence of distortion in the least-diverged crosses, together with its occurrence within as well as between species, indicates that transmission defects can arise before complete reproductive isolation. Moreover, the tendency for maternal- and paternal-direction effects to offset one another, including repeated antiparallel signals involving chromosomes 4 and 5, is more consistent with negative epistatic incompatibilities than with a collection of independent single-locus drivers. These findings identify the haploid male gametophyte as an early arena in which incompatibilities can reduce hybrid fertility and potentially accelerate speciation.

In the second chapter, I examined the consequences of long-term recombination suppression in the UV mating-type chromosomes of Mamiellales. Across four deeply diverged species, genes within these regions showed elevated intron retention and multiple other forms of alternative splicing, demonstrating that the defect is broad and evolutionarily persistent. Long-read isoforms from *M. pusilla* further showed that mating-type genes generate unusually diverse transcripts, but a smaller fraction retain most of the annotated coding sequence in frame and their predicted protein domains are more frequently disrupted. At the same time, many

genes retain at least one likely productive transcript, explaining how essential coding capacity can persist despite pervasive molecular dysfunction. These results support a model in which altered nucleotide composition and chromatin environment progressively erode RNA-processing fidelity, revealing transcript-level degeneration as an alternative to the familiar gene-loss trajectory of non-recombining chromosomes.

In the third chapter, I studied introners as a direct example of transposable-element proliferation reshaping gene structure. Natural *M. pusilla* populations differed approximately fourfold in introner abundance and shared few orthologous insertion sites, yet introners accumulated in similar broad genomic neighborhoods. Within the introner-rich population, the allele-frequency spectrum and demographic modeling revealed ongoing gain and loss, with recombinogenic regions acting as permissive environments rather than evidence that recent positive selection drives insertion turnover. Introner number was associated with increased isoform diversity, reduced gene expression, and a higher probability of premature termination and nonsense-mediated decay, while depletion from core ribosomal and transcriptional genes showed that purifying selection removes insertions from especially dosage-sensitive contexts. The resulting landscape records both element mobility and host filtering: introners can proliferate, but their persistence depends on where they insert and how strongly they disrupt gene function.

Overall, this work shows that genomic conflict cannot be understood solely from the molecular machinery of a selfish element. Its evolutionary impact emerges

from an interaction among transmission mechanism, recombination, genome architecture, and host fitness. In *Arabidopsis*, divergence converts ordinary allelic differences into gamete-level incompatibilities that bias inheritance; in algal mating-type regions, recombination suppression permits a cryptic accumulation of transcript-processing defects; and in *M. pusilla*, mobile introners continually create new variation that is filtered by local genomic context and functional cost. Together, these studies bridge the short timescale of segregation and insertion with the long timescale of chromosome evolution. They demonstrate that selfish processes can contribute both to the formation of reproductive barriers and to lasting changes in genome and transcriptome organization, while also highlighting the host constraints that prevent unchecked spread.

List of Supplemental Files

Supplemental file 1: Dataset_S1_PO_Categories.xlsx

Supplemental file 2: Dataset_S2_chr4_genes.xlsx

Supplemental file 3: Dataset_S3_chr5_genes.xlsx

Supplemental file 4: Dataset_S4_pollen_related_loci.xlsx

References

Bachtrog, Doris. 2013. “Y-Chromosome Evolution: Emerging Insights into Processes of Y-Chromosome Degeneration.” *Nature Reviews Genetics* 14 (2): 113–24.

Bailey, Timothy L. 2011. “DREME: Motif Discovery in Transcription Factor ChIP-Seq Data.” *Bioinformatics (Oxford, England)* 27 (12): 1653–59.

Beeman, R. W., K. S. Friesen, and R. E. Denell. 1992. “Maternal-Effect Selfish Genes in Flour Beetles.” *Science (New York, N.Y.)* 256 (5053): 89–92.

Begun, D. J., and C. F. Aquadro. 1993. “African and North American Populations of Drosophila Melanogaster Are Very Different at the DNA Level.” *Nature* 365 (6446): 548–50.

Ben-David, Eyal, Alejandro Burga, and Leonid Kruglyak. 2017. “A Maternal-Effect Selfish Genetic Element in Caenorhabditis Elegans.” *Science (New York, N.Y.)* 356 (6342): 1051–55.

Benites, Luis Felipe, François Bucchini, Sophie Sanchez-Brosseau, Nigel Grimsley, Klaas Vandepoele, and Gwenaël Piganeau. 2021. “Evolutionary Genomics of Sex-Related Chromosomes at the Base of the Green Lineage.” *Genome Biology and Evolution* 13 (10): evab216.

Bénitière, Florian, Anamaria Necsulea, and Laurent Duret. 2024. *Random Genetic Drift Sets an Upper Limit on mRNA Splicing Accuracy in Metazoans*. March 12. https://doi.org/10.7554/eLife.93629.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. “Retroposed New Genes out of the X in Drosophila.” *Genome Research* 12 (12): 1854–59.

Blum, Matthias, Hsin-Yu Chang, Sara Chuguransky, et al. 2021. “The InterPro Protein Families and Domains Database: 20 Years On.” *Nucleic Acids Research* 49 (D1): D344–54.

Bomblies, Kirsten, Janne Lempe, Petra Epple, et al. 2007. “Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants.” *PLoS Biology* 5 (9): e236.

Bull, James J. 1978. “Sex Chromosomes in Haploid Dioecy: A Unique Contrast to Muller’s Theory for Diploid Dioecy.” *The American Naturalist*, ahead of print, January 1. https://doi.org/10.1086/283267.

Burt, A., and R. Trivers. 1998. “Selfish DNA and Breeding System in Flowering Plants.” *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265 (1391): 141–46.

Burt, Austin, and Robert Trivers. 2008. *Genes in Conflict: The Biology of Selfish Genetic Elements*. Belknap Press.

Camacho, Christiam, George Coulouris, Vahram Avagyan, et al. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (1): 421.

Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. “EggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.” *Molecular Biology and Evolution* 38 (12): 5825–29.

Carioscia, Sara A., Kathryn J. Weaver, Andrew N. Bortvin, et al. 2022. “A Method for Low-Coverage Single-Gamete Sequence Analysis Demonstrates Adherence to Mendel’s First Law across a Large Sample of Human Sperm.” *eLife* 11 (December): e76383.

Carmel, Ido, Saar Tal, Ida Vig, and Gil Ast. 2004. “Comparative Analysis Detects Dependencies among the 5’ Splice-Site Positions.” *RNA (New York, N.Y.)* 10 (5): 828–40.

Castric, V., J. S. Bechsgaard, S. Grenier, R. Noureddine, M. H. Schierup, and X. Vekemans. 2010. “Molecular Evolution within and between Self-Incompatibility Specificities.” *Molecular Biology and Evolution* 27 (1): 11–20.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. “The Effect of Deleterious Mutations on Neutral Molecular Variation.” *Genetics* 134 (4): 1289–303.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, et al. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3: SNPs in the Genome of Drosophila Melanogaster Strain w ^{1118} ; Iso-2; Iso-3.” *Fly* 6 (2): 80–92.

Coelho, S. M., J. Gueno, A. P. Lipinska, J. M. Cock, and J. G. Umen. 2018. “UV Chromosomes and Haploid Sexual Systems.” *Trends in Plant Science* 23 (9): 794.

Condon, Chris, Andrea Galvez, Alexander Kramer, et al. 2026. “Splicing Deficiency Is Driven by Genomic Erosion in Non-Recombining Algal Mating-Type Chromosomes.” *PLoS Biology* 24 (6): e3003823.

Corbett-Detig, Russell B., Daniel L. Hartl, and Timothy B. Sackton. 2015. “Natural Selection Constrains Neutral Diversity across a Wide Range of Species.” *PLoS Biology* 13 (4): e1002112.

Corbett-Detig, Russell B., Jun Zhou, Andrew G. Clark, Daniel L. Hartl, and Julien F. Ayroles. 2013. “Genetic Incompatibilities Are Widespread within Species.” *Nature* 504 (7478): 135–37.

Corbett-Detig, Russell, Paloma Medina, Hélène Frérot, Christelle Blassiau, and Vincent Castric. 2019. “Bulk Pollen Sequencing Reveals Rapid Evolution of Segregation Distortion in the Male Germline of Arabidopsis Hybrids.” *Evolution Letters* 3 (1): 93–103.

Cowell, Finn. 2023. “100 Years of Haldane’s Rule.” *Journal of Evolutionary Biology* 36 (2): 337–46.

Coyne, Jerry A. 1992. “Genetics and Speciation.” *Nature* 355 (6360): 511–15.

Coyne, Jerry A., and H. Allen Orr. 2004. *Speciation*. Sinauer Associates.

Csuros, Miklos, Igor B. Rogozin, and Eugene V. Koonin. 2011. “A Detailed History of Intron-Rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes.” *PLoS Computational Biology* 7 (9): e1002150.

Dagilis, Andrius J., Mark Kirkpatrick, and Daniel I. Bolnick. 2019. “The Evolution of Hybrid Fitness during Speciation.” *PLoS Genetics* 15 (5): e1008125.

Danecek, Petr, Adam Auton, Goncalo Abecasis, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics (Oxford, England)* 27 (15): 2156–58.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2): giab008.

Davey, Colin S., Sari Pennings, Carmel Reilly, Richard R. Meehan, and James Allan. 2004. “A Determining Influence for CpG Dinucleotides on Nucleosome Positioning in Vitro.” *Nucleic Acids Research* 32 (14): 4322–31.

DeGiorgio, Michael, Christian D. Huber, Melissa J. Hubisz, Ines Hellmann, and Rasmus Nielsen. 2016. “SweepFinder2: Increased Sensitivity, Robustness and Flexibility.” *Bioinformatics (Oxford, England)* 32 (12): 1895–97.

Derr, L. K., and J. N. Strathern. 1993. “A Role for Reverse Transcripts in Gene Conversion.” *Nature* 361 (6408): 170–73.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, et al. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics (Oxford, England)* 29 (1): 15–21.

Dobzhansky, Theodosius. 1982. *Genetics and the Origin of Species: Columbia Classics Edition*. Columbia University Press.

Emerson, J. J., Henrik Kaessmann, Esther Betrán, and Manyuan Long. 2004. "Extensive Gene Traffic on the Mammalian X Chromosome." *Science (New York, N.Y.)* 303 (5657): 537–40.

Enevoldsen, Henrik O., and Unesco. 2003. *Manual on Harmful Marine Microalgae*. UNESCO.

Farlow, Ashley, Eshwar Meduri, Marlies Dolezal, Liushuai Hua, and Christian Schlötterer. 2010. "Nonsense-Mediated Decay Enables Intron Gain in Drosophila." *PLoS Genetics* 6 (1): e1000819.

Fishman, Lila, and John H. Willis. 2005. "A Novel Meiotic Drive Locus Almost Completely Distorts Segregation in Mimulus (Monkeyflower) Hybrids." *Genetics* 169 (1): 347–53.

Fong, Nova, Hyunmin Kim, Yu Zhou, et al. 2014. "Pre-mRNA Splicing Is Facilitated by an Optimal RNA Polymerase II Elongation Rate." *Genes & Development* 28 (23): 2663–76.

Freed, Donald, Rafael Aldana, Jessica Weber, and Jeremy Edwards. 2017. *The Sentieon Genomics Tools - A Fast and Accurate Solution to Variant Calling from next-Generation Sequence Dat*. https://doi.org/10.1101/115717.

Fu, Xiang-Dong, and Manuel Ares Jr. 2014. "Context-Dependent Control of Alternative Splicing by RNA-Binding Proteins." *Nature Reviews. Genetics* 15 (10): 689–701.

Gadau, J., R. E. Page, and J. H. Werren. 1999. "Mapping of Hybrid Incompatibility Loci in Nasonia." *Genetics* 153 (4): 1731–41.

Gelfman, Sahar, Noa Cohen, Ahuvi Yearim, and Gil Ast. 2013. "DNA-Methylation Effect on Cotranscriptional Splicing Is Dependent on GC Architecture of the Exon-Intron Structure." *Genome Research* 23 (5): 789–99.

Gozashti, Landen, Scott W. Roy, Bryan Thornlow, Alexander Kramer, Manuel Ares Jr, and Russell Corbett-Detig. 2022. "Transposable Elements Drive Intron Gain in Diverse Eukaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 119 (48): e2209766119.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics (Oxford, England)* 29 (8): 1072–75.

Haldane, J. B. 1922. “Sex Ratio and Aunisexual Sterility in Hybrid Animals.” *Journal of Genetics* 12: 101–9.

Hämälä, Tuomas, Tiina M. Mattila, Päivi H. Leinonen, Helmi Kuittinen, and Outi Savolainen. 2017. “Role of Seed Germination in Adaptation and Reproductive Isolation in Arabidopsis Lyrata.” *Molecular Ecology* 26 (13): 3484–96.

Hickey, Glenn, Jean Monlong, Jana Ebler, et al. 2024. “Pangenome Graph Construction from Genome Alignments with Minigraph-Cactus.” *Nature Biotechnology* 42 (4): 663–73.

Hu, Tina T., Pedro Pattyn, Erica G. Bakker, et al. 2011. “The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change.” *Nature Genetics* 43 (5): 476–81.

Huang, Yao-Ting, Po-Yu Liu, and Pei-Wen Shih. 2021. “Homopolish: A Method for the Removal of Systematic Errors in Nanopore Sequencing by Homologous Polishing.” *Genome Biology* 22 (1): 95.

Huff, Jason T., and Daniel Zilberman. 2014. “Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes.” *Cell* 156 (6): 1286–97.

Huff, Jason T., Daniel Zilberman, and Scott W. Roy. 2016. “Mechanism for DNA Transposons to Generate Introns on Genomic Scales.” *Nature* 538 (7626): 533–36.

Irimia, Manuel, and Scott William Roy. 2014. “Origin of Spliceosomal Introns and Alternative Splicing.” *Cold Spring Harbor Perspectives in Biology* 6 (6): a016071.

J, Muller H. 1940. “Bearing of the Drosophila Work on Systematics.” *The New Systematics,* 185–268.

Jimeno-González, Silvia, Laura Payán-Bravo, Ana M. Muñoz-Cabello, et al. 2015. “Defective Histone Supply Causes Changes in RNA Polymerase II Elongation Rate and Cotranscriptional Pre-mRNA Splicing.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (48): 14840–45.

Jones, Philip, David Binns, Hsin-Yu Chang, et al. 2014. “InterProScan 5: Genome-Scale Protein Function Classification.” *Bioinformatics* 30 (9): 1236–40.

Jouganous, Julien, Will Long, Aaron P. Ragsdale, and Simon Gravel. 2017. “Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation.” *Genetics* 206 (3): 1549–67.

Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. 2025. “KEGG: Biological Systems Database as a Model of the Real World.” *Nucleic Acids Research* 53 (D1): D672–77.

Katju, Vaishali, and Ulfar Bergthorsson. 2019. “Old Trade, New Tricks: Insights into the Spontaneous Mutation Process from the Partnering of Classical Mutation Accumulation Experiments with High-Throughput Genomic Approaches.” *Genome Biology and Evolution* 11 (1): 136–65.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15.

Krasovec, Marc, Adam Eyre-Walker, Sophie Sanchez-Ferandin, and Gwenael Piganeau. 2017. “Spontaneous Mutation Rate in the Smallest Photosynthetic Eukaryotes.” *Molecular Biology and Evolution* 34 (7): 1770–79.

Kurosaki, Tatsuaki, Maximilian W. Popp, and Lynne E. Maquat. 2019. “Quality and Quantity Control of Gene Expression by Nonsense-Mediated mRNA Decay.” *Nature Reviews. Molecular Cell Biology* 20 (7): 406–20.

Lahn, B. T., and D. C. Page. 1999. “Four Evolutionary Strata on the Human X Chromosome.” *Science (New York, N.Y.)* 286 (5441): 964–67.

Lahn, Bruce T., Nathaniel M. Pearson, and Karin Jegalian. 2001. “The Human Y Chromosome, in the Light of Evolution.” *Nature Reviews Genetics* 2 (3): 207–16.

Leppälä, J., J. S. Bechsgaard, M. H. Schierup, and O. Savolainen. 2008. “Transmission Ratio Distortion in Arabidopsis Lyrata: Effects of Population Divergence and the S-Locus.” *Heredity* 100 (1): 71–78.

Leppälä, Johanna, Folmer Bokma, and Outi Savolainen. 2013. “Investigating Incipient Speciation in Arabidopsis Lyrata from Patterns of Transmission Ratio Distortion.” *Genetics* 194 (3): 697–708.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics (Oxford, England)* 34 (18): 3094–100.

Li, Heng. 2023. “Protein-to-Genome Alignment with Miniprot.” *Bioinformatics (Oxford, England)* 39 (1): btad014.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60.

Li, Heng, Bob Handsaker, Alec Wysoker, et al. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2013. “The Subread Aligner: Fast, Accurate and Scalable Read Mapping by Seed-and-Vote.” *Nucleic Acids Research* 41 (10): e108.

Lin, S. Y., H. Ikehashi, S. Yanagihara, and A. Kawashima. 1992. “Segregation Distortion via Male Gametes in Hybrids between Indica and Japonica or Wide-Compatibility Varieties of Rice (Oryza Sativa L).” *Theoretical and Applied Genetics* 84 (7): 812–18.

Lipinska, Agnieszka P., Nicholas R. T. Toda, Svenja Heesch, Akira F. Peters, J. Mark Cock, and Susana M. Coelho. 2017. “Multiple Gene Movements into and out of Haploid Sex Chromosomes.” *Genome Biology* 18 (1): 104.

Luo, Xuming, Ning Xu, Junkai Huang, et al. 2017. “A Lectin Receptor-Like Kinase Mediates Pattern-Triggered Salicylic Acid Signaling.” *Plant Physiology* 174 (4): 2501–14.

Maayan Amit, Maya Donyo, Dror Hollander, Amir Goren, Eddo Kim, Sahar Gelfman, Galit Lev-Maor, David Burstein, Schraga Schwartz, Benny Postolsky, Tal Pupko, Gil Ast. 2012. “Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.” *Cell Reports* 1 (5): 543–56.

Mantel, Samuel J., and Andrea L. Sweigart. 2024. “Postzygotic Barriers Persist despite Ongoing Introgression in Hybridizing Mimulus Species.” *Molecular Ecology* 33 (4): e17261.

Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. 2018. “MUMmer4: A Fast and Versatile Genome Alignment System.” *PLoS Computational Biology* 14 (1): e1005944.

Martin, Marcel, Murray Patterson, Shilpa Garg, et al. 2016. “WhatsHap: Fast and Accurate Read-Based Phasing.” In *bioRxiv*. November 14. bioRxiv. https://doi.org/10.1101/085050.

Matute, Daniel R., and Jerry A. Coyne. 2010. “INTRINSIC REPRODUCTIVE ISOLATION BETWEEN TWO SISTER SPECIES OF DROSOPHILA.” *Evolution; International Journal of Organic Evolution* 64 (4): 903–20.

McDaniel, Stuart F., John H. Willis, and A. Jonathan Shaw. 2007. “A Linkage Map Reveals a Complex Basis for Segregation Distortion in an Interpopulation Cross in the Moss Ceratodon Purpureus.” *Genetics* 176 (4): 2489.

McDermott, Shannon R., and Mohamed A. F. Noor. 2012. “Mapping of Within-Species Segregation Distortion in D. Persimilis and Hybrid Sterility between D. Persimilis and D. Pseudoobscura.” *Journal of Evolutionary Biology* 25 (10): 2023.

McRose, Darcy, Jian Guo, Adam Monier, et al. 2014. “Alternatives to Vitamin B1 Uptake Revealed with Discovery of Riboswitches in Multiple Marine Eukaryotic Lineages.” *The ISME Journal* 8 (12): 2517–29.

Mi, Huaiyu, Dustin Ebert, Anushya Muruganujan, et al. 2021. “PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API.” *Nucleic Acids Research* 49 (D1): D394–403.

Mirchandani, Cade D., Allison J. Shultz, Gregg W. C. Thomas, et al. 2024. “A Fast, Reproducible, High-Throughput Variant Calling Workflow for Population Genomics.” *Molecular Biology and Evolution* 41 (1): msad270.

Mirchandani, Cade, Max Genetti, Pingting Wang, Evan Pepper-Tunick, Shelbi Russell, and Russell Corbett-Detig. 2024. “Plate Scale Tn5 Based Tagmentation Library Prep Protocol V1.” In *Protocols.Io*. March 25. https://doi.org/10.17504/protocols.io.4r3l2qmzpl1y/v1.

Monnet, François, Zoé Postel, Pascal Touzet, et al. 2023. “Rapid Establishment of Species Barriers in Plants Compared to Animals.” In *bioRxiv*. October 20. bioRxiv. https://doi.org/10.1101/2023.10.16.562535.

Moran, Benjamin M., Cheyenne Y. Payne, Daniel L. Powell, et al. 2024. “A Lethal Mitonuclear Incompatibility in Complex I of Natural Hybrids.” *Nature* 626 (7997): 119–27.

Muller, H. J. 1942. “Isolating Mechanisms, Evolution, and Temperature.” *Biol. Symp.* 6: 71.

Muralla, Rosanna, Johnny Lloyd, and David Meinke. 2011. “Molecular Foundations of Reproductive Lethality in Arabidopsis Thaliana.” *PloS One* 6 (12): e28398.

Naftelberg, Shiran, Ignacio E. Schor, Gil Ast, and Alberto R. Kornblihtt. 2015. “Regulation of Alternative Splicing through Coupling with Transcription and Chromatin Structure.” *Annual Review of Biochemistry* 84: 165–98.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Noble, Luke M., John Yuen, Lewis Stevens, et al. 2021. “Selfing Is the Safest Sex for Caenorhabditis Tropicalis.” *eLife* 10 (January): e62587.

Nogués, Guadalupe, Sebastián Kadener, Paula Cramer, et al. 2003. “Control of Alternative Pre-mRNA Splicing by RNA Pol II Elongation: Faster Is Not Always Better.” *IUBMB Life* 55 (4–5): 235–41.

Novikova, Polina Yu, Nora Hohmann, Viktoria Nizhynska, et al. 2016. “Sequencing of the Genus Arabidopsis Identifies a Complex History of Nonbifurcating Speciation and Abundant Trans-Specific Polymorphism.” *Nature Genetics* 48 (9): 1077–82.

Orr, H. A. 1996. “Dobzhansky, Bateson, and the Genetics of Speciation.” *Genetics* 144 (4): 1331–35.

Pardo-Palacios, Francisco J., Angeles Arzalluz-Luque, Liudmyla Kondratova, et al. 2024. “SQANTI3: Curation of Long-Read Transcriptomes for Accurate Identification of Known and Novel Isoforms.” *Nature Methods* 21 (5): 793–97.

Parfrey, Laura Wegener, Daniel J. G. Lahr, Andrew H. Knoll, and Laura A. Katz. 2011. “Estimating the Timing of Early Eukaryotic Diversification with Multigene Molecular Clocks.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (33): 13624–29.

Pavan, Flavia, Jacinthe Azevedo Favory, Eléanore Lacoste, et al. 2024. “The Evolutionary History and Functional Specialization of microRNA Genes in Arabidopsis Halleri and A. Lyrata.” In *bioRxiv*. May 7. bioRxiv. https://doi.org/10.1101/2024.05.03.592357.

Phadnis, Nitin, and H. Allen Orr. 2009. “A Single Gene Causes Both Male Sterility and Segregation Distortion in Drosophila Hybrids.” *Science* 323 (5912): 376–79.

Price, T. a. R., R. Verspoor, and N. Wedell. 2019. “Ancient Gene Drives: An Evolutionary Paradox.” *Proceedings. Biological Sciences* 286 (1917): 20192267.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.

Rajaby, Ramesh, Dong-Xu Liu, Chun Hang Au, et al. 2023. “INSurVeyor: Improving Insertion Calling from Short Read Sequencing Data.” *Nature Communications* 14 (1): 3243.

Ravinet, M., R. Faria, R. K. Butlin, et al. 2017. “Interpreting the Genomic Landscape of Speciation: A Road Map for Finding Barriers to Gene Flow.” *Journal of Evolutionary Biology* 30 (8): 1450–77.

Reimer, Kirsten A., Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. 2021. “Co-Transcriptional Splicing Regulates 3’ End Cleavage during Mammalian Erythropoiesis.” *Molecular Cell* 81 (5): 998-1012.e7.

Reiser, Leonore, Erica Bakker, Sabarinath Subramaniam, et al. 2024. “The Arabidopsis Information Resource in 2024.” *Genetics* 227 (1): iyae027.

Roux, Camille, Vincent Castric, Maxime Pauwels, Stephen I. Wright, Pierre Saumitou-Laprade, and Xavier Vekemans. 2011. “Does Speciation between Arabidopsis Halleri and Arabidopsis Lyrata Coincide with Major Changes in a Molecular Target of Adaptation?” *PloS One* 6 (11): e26872.

Roux, Camille, Christelle Fraïsse, Jonathan Romiguier, Yoann Anciaux, Nicolas Galtier, and Nicolas Bierne. 2016. “Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence.” *PLoS Biology* 14 (12): e2000234.

Saudemont, Baptiste, Alexandra Popa, Joanna L. Parmley, et al. 2017. “The Fitness Cost of Mis-Splicing Is the Main Determinant of Alternative Splicing Patterns.” *Genome Biology* 18 (1): 208.

Sawa, H., and J. Abelson. 1992. “Evidence for a Base-Pairing Interaction between U6 Small Nuclear RNA and 5’ Splice Site during the Splicing Reaction in Yeast.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (23): 11269–73.

Schwartz, Schraga, Eran Meshorer, and Gil Ast. 2009. “Chromatin Organization Marks Exon-Intron Structure.” *Nature Structural & Molecular Biology* 16 (9): 990–95.

Seymour, Danelle K., Eunyoung Chae, Burak I. Arioz, Daniel Koenig, and Detlef Weigel. 2019. “Transmission Ratio Distortion Is Frequent in Arabidopsis Thaliana Controlled Crosses.” *Heredity* 122 (3): 294–304.

Shen, Shihao, Juw Won Park, Zhi-Xiang Lu, et al. 2014. “rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (51): E5593-601.

Sheth, Nihar, Xavier Roca, Michelle L. Hastings, Ted Roeder, Adrian R. Krainer, and Ravi Sachidanandam. 2006. “Comprehensive Splice-Site Analysis Using Comparative Genomics.” *Nucleic Acids Research* 34 (14): 3955–67.

Simon, Alexis, Nicolas Bierne, and John J. Welch. 2018. “Coadapted Genomes and Selection on Hybrids: Fisher’s Geometric Model Explains a Variety of Empirical Patterns.” *Evolution Letters* 2 (5): 472–98.

Spence, Jeffrey P., and Yun S. Song. 2019. “Inference and Analysis of Population-Specific Fine-Scale Recombination Maps across 26 Diverse Human Populations.” *Science Advances* 5 (10): eaaw9206.

Stein, Ricardo J., Stephan Höreth, J. Romário F. de Melo, et al. 2017. “Relationships between Soil and Leaf Mineral Composition Are Element-Specific, Environment-Dependent and Geographically Structured in the Emerging Model Arabidopsis Halleri.” *The New Phytologist* 213 (3): 1274–86.

Sterner, D. A., T. Carlo, and S. M. Berget. 1996. “Architectural Limits on Split Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (26): 15081–85.

Takou, Margarita, Tuomas Hämälä, Evan M. Koch, et al. 2021. “Maintenance of Adaptive Dynamics and No Detectable Load in a Range-Edge Outcrossing Plant Population.” *Molecular Biology and Evolution* 38 (5): 1820–36.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, et al. 2018. “SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification.” *Genome Research* 28 (3): 396–411.

Tegenfeldt, Fredrik, Dmitry Kuznetsov, Mosè Manni, Matthew Berkeley, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2025. “OrthoDB and BUSCO Update: Annotation of Orthologs with Wider Sampling of Genomes.” *Nucleic Acids Research* 53 (D1): D516–22.

Tillo, Desiree, and Timothy R. Hughes. 2009. “G+C Content Dominates Intrinsic Nucleosome Occupancy.” *BMC Bioinformatics* 10 (December): 442.

Turelli, M., and H. A. Orr. 1995. “The Dominance Theory of Haldane’s Rule.” *Genetics* 140 (1): 389–402.

Valiadi, Martha, Keith Harrison, Yann Loe-Mie, et al. 2025. “Insulated Outlier Chromosomes Enable Metabolic Innovation in Minimal Eukaryotic Algae.” In *bioRxiv*. April 22. https://doi.org/10.1101/2025.04.21.649811.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Christopher Hartl, et al. 2013. “From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices

Pipeline." *Current Protocols in Bioinformatics* (Hoboken, NJ, USA) 43 (1): 11.10.1-11.10.33.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. "Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads." *Genome Research* 27 (5): 737–46.

Vogl, Claus, Lynette C. Mikula, and Conrad J. Burden. 2020. "Maximum Likelihood Estimators for Scaled Mutation Rates in an Equilibrium Mutation-Drift Model." *Theoretical Population Biology* 134 (August): 106–18.

Volden, Roger, Theron Palmer, Ashley Byrne, et al. 2018. "Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell cDNA." *Proceedings of the National Academy of Sciences* 115 (39): 9726–31.

Wang, Wei-Kuang, Chuan-Wen Ho, Kuo-Hsiang Hung, et al. 2010. "Multilocus Analysis of Genetic Divergence between Outcrossing Arabidopsis Species: Evidence of Genome-Wide Admixture." *The New Phytologist* 188 (2): 488–500.

Wang, Yan, Jan H. G. Cordewener, Antoine H. P. America, Weixing Shan, Klaas Bouwmeester, and Francine Govers. 2015. "Arabidopsis Lectin Receptor Kinases LecRK-IX.1 and LecRK-IX.2 Are Functional Analogs in Regulating Phytophthora Resistance and Plant Cell Death." *Molecular Plant-Microbe Interactions: MPMI* 28 (9): 1032–48.

Wassarman, D. A., and J. A. Steitz. 1993. "A Base-Pairing Interaction between U2 and U6 Small Nuclear RNAs Occurs in > 150S Complexes in HeLa Cell Extracts: Implications for the Spliceosome Assembly Pathway." *Proceedings of the National Academy of Sciences of the United States of America* 90 (15): 7139–43.

Weichenhan, D., W. Traut, B. Kunze, and H. Winking. 1996. "Distortion of Mendelian Recovery Ratio for a Mouse HSR Is Caused by Maternal and Zygotic Effects." *Genetical Research* 68 (2): 125–29.

Willi, Yvonne. 2013. "Mutational Meltdown in Selfing Arabidopsis Lyrata." *Evolution; International Journal of Organic Evolution* 67 (3): 806–15.

Worden, Alexandra Z., Jae-Hyeok Lee, Thomas Mock, et al. 2009. "Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas." *Science (New York, N.Y.)* 324 (5924): 268–72.

Xiong, Yan, Anthony L. Contento, and Diane C. Bassham. 2005. "AtATG18a Is Required for the Formation of Autophagosomes during Nutrient Stress and

Senescence in Arabidopsis Thaliana." *The Plant Journal: For Cell and Molecular Biology* 42 (4): 535–46.