UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THE EVOLUTIONARY CONSEQUENCES OF INTROGRESSION AMONG STRONGYLOCENTROTID SEA URCHINS

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

DOCTOR OF PHILOSOPHY

in

ECOLOGY AND EVOLUTIONARY BIOLOGY

by

Matthew R. Glasenapp

June 2024

The Dissertation of Matthew R. Glasenapp
is approved:

_____
Professor Emeritus Grant Pogson, Chair

_____
Professor Peter Raimondi

_____
Associate Professor Russ Corbett-Detig

_____
Associate Professor Kord Kober

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables

List of Figures

Abstract

The Evolutionary Consequences of Introgression

Among Strongylocentrotid Sea Urchins

by

Matthew R. Glasenapp

Understanding the genomic architecture of speciation remains a challenge in evolutionary biology. Among broadcast-spawning marine invertebrates, reproductive isolation is thought to be established and maintained by the divergence of gamete recognition proteins located on the surfaces of sperm and egg cells. However, it remains unclear whether gametic isolation has been an effective barrier to gene flow during and/or following speciation. In this dissertation, I characterized the history of introgression among the North Pacific sea urchin species of the family *Strongylocentrotidae* to deepen our understanding of their diversification and evaluate the importance of gametic isolation in speciation. Using whole-genome sequencing data from each strongylocentrotid species and cutting-edge phylogenomic approaches, I documented widespread introgression in both extant taxa and ancestral lineages, demonstrating that gametic isolation did not effectively limit introgression. I implemented a phylogenetic hidden Markov model to locate the specific regions of the genome affected by introgression, finding evidence of strong selection against introgression across much of the genome. Although introgressed variation has predominantly persisted in slowly evolving, low-divergence genomic regions,

numerous protein-coding genes showed both introgression and historical positive selection, suggesting an adaptive role for introgression. Finally, I showed that the two gamete recognition proteins responsible for species-selectivity in sea urchin fertilization, sperm protein bindin and its egg receptor, EBR1, have experienced historical adaptive introgression, a pattern inconsistent with expectations for barrier loci. My findings contribute to the body of literature evaluating the biological consequences of introgression and question the importance of gamete recognition proteins in the evolution of reproductive isolation among incipient strongylocentrotid sea urchin species.

Acknowledgments

This work would not have been possible without the guidance and support I received from my mentors, peers, and family. My advisor, Grant Pogson, has had a profound influence on the way I think about biology and approach science. Grant's love of nature is infectious and propelled me through my degree. My dissertation committee members Kord Kober, Pete Raimondi, and Russ Corbett-Detig provided invaluable feedback on my research ideas and manuscripts. I benefitted tremendously from access to the UCSC Hummingbird HPC cluster and thank Rion Parsons and Jeffrey Weekly for their support with storage and computing. I thank Luay Nakleh, Qiqige Wuyun, and Zhi Yan for their help with the PhyloNet software package, and Matthew Kustra for his help with data analysis and visualization.

I am forever grateful for the support I've received from my parents, Bob and Nancy, my partner, Megha, and her family, Radhika, Ambarish, and Uma. They have been nothing but encouraging and uplifting from day one and tremendously increased my quality of life as a graduate student.

Funding for this research was provided by the National Science Foundation (DEB-1011061), the STEPS Foundation, Friends of Long Marine Lab, and the Myers Trust. The funding bodies did not participate in research design, sample collection, data analysis, or manuscript writing.

The text of this dissertation includes a reprint of the following previously published material. The co-author listed in this publication directed and supervised the research which forms the basis for the dissertation.

Glasenapp, M. R., & Pogson, G. H. (2023). Extensive introgression among strongylocentrotid sea urchins revealed by phylogenomics. *Ecology and Evolution*, 13, e10446. https://doi.org/10.1002/ece3.10446

Introduction

The origin of species remains a central and elusive problem in evolutionary biology (Coyne and Orr 2004; Schluter and Rieseberg 2022). To link the continuous microevolutionary processes operating within species to the emergence of new genetically discrete groups, speciation research centers around identifying barriers to reproduction between young species and characterizing the evolutionary forces behind their emergence. While species boundaries are often defined by pronounced phenotypes, as seen in ecomorphs or trophic specialists, the prevalence of cryptic species demonstrates that reproductive isolation commonly evolves without significant changes in ecology, physiology, or morphology. Cryptic species typically remain undetected until genetic identification, and it remains unclear whether the loci underlying reproductive barriers are involved in adaptive diversification or non-adaptive processes such as intragenomic conflict and divergent gene duplication resolution.

The recent availability of genomics data has expanded the taxonomic representation in speciation studies to include cryptic species and other non-model groups that are difficult to observe in nature or rear in the lab. In cases where phenotypic divergence between species is minimal, or the underlying reproductive barriers are unknown, nucleotide differentiation and divergence can be profiled across entire aligned genomes to identify outlier loci that may represent early diverging genes involved in reproductive isolation. This “landscape” of divergence or differentiation is primarily a product of variable selection acting across the genome. However, the

presence of gene flow between incipient species while reproductive isolation remains incomplete can also cause heterogeneity in divergence, where barrier loci resistant to gene flow continue to diverge while much of the rest of the genome is homogenized. Therefore, characterizing the genomic landscapes of divergence and introgression among hybridizing species has become a popular approach for reconstructing the genetic architecture of speciation.

The "genomic landscape of introgression" approach has benefitted from the discovery that hybridization is much more common than previously thought, occurring in at least 25% of flowering plants and 10% of animal species (Mallet 2005). Historically, hybridization was thought to impede speciation because it can break down prezygotic reproductive barriers and homogenize accumulated genetic differences between incipient species. However, the genetic differences distinguishing populations are often maintained in the face of hybridization and may even be strengthened by selection for increased prezygotic isolation. While reproductive isolation remains incomplete, hybridization can lead to the introgression of alleles between species and is now known to be an important source of variation fueling adaptive radiations. Our ability to sequence and analyze entire genomes has made it increasingly clear that hybridization and introgression are key sources of molecular and phenotypic diversity. Therefore, in addition to identifying barrier loci resistant to introgression, quantifying introgression across the genome also allows the characterization of introgressed loci involved in adaptive diversification.

For my dissertation, I characterized the history of introgression among the North Pacific sea urchin species of the family *Strongylocentrotidae* to deepen our understanding of their diversification. Like other broadcast-spawning marine invertebrate groups, sea urchins have massive fecundities and highly dispersive larvae, resulting in enormous population sizes and broad geographic ranges. The rarity of absolute geographic barriers in the sea and the high amount of gene flow across large distances should restrict opportunities for population differentiation and the evolution of reproductive isolation (Palumbi 1992). However, species diversity in sea urchins and other broadcast spawners appears high. One proposed mechanism accounting for this higher-than-expected diversity is that the rapid evolution of a small number of reproductive proteins may establish gametic isolation between differentiated populations and facilitate speciation (Swanson and Vacquier 2002a; S. R. Palumbi 2009). If speciation proceeded predominantly through gametic isolation, evidence of post-speciation gene flow should be rare, especially between non-sister taxa, as gametic incompatibilities should grow with time since separation. However, introgression might be more common if speciation was initiated by geographic, habitat, or temporal isolating barriers, as changes in range, habitat preference, or spawning time following speciation could lead to secondary contact.

Here, I demonstrate the biological and evolutionary consequences of introgression among the strongylocentrotid sea urchins at multiple scales, ranging from the whole-genome to the individual gene. In Chapter 1, I tested for introgression using several complementary phylogenomic approaches and evaluated whether introgression

was associated with gamete recognition protein or phylogenetic distances. In my second chapter, I located the specific genomic regions affected by introgression and tested whether these regions showed nonrandom patterns relative to the genome-wide background. In Chapter 3, I characterized the signatures of selection and introgression at the two genes responsible for species-selectivity in fertilization, bindin and egg bindin receptor 1 (EBR1), and asked whether their molecular evolutionary histories are consistent with a role in reproductive isolation. Taken together, my dissertation chapters provide a unique perspective on the origin, maintenance, and diversification of the strongylocentrotid urchins, improving our understanding of these processes in broadcast-spawning marine invertebrates.

Chapter 1 Extensive Introgression among Strongylocentrotid Sea Urchins Revealed by Phylogenomics

Abstract

Gametic isolation is thought to play an important role in the evolution of reproductive isolation in broadcast-spawning marine invertebrates. However, it's unclear whether gametic isolation commonly develops early in the speciation process or only accumulates after other reproductive barriers are already in place. It is also unknown whether gametic isolation is an effective barrier to introgression following speciation. Here, we used whole-genome sequencing data and multiple complementary phylogenomic approaches to test whether the well-documented gametic incompatibilities among the strongylocentrotid sea urchins have limited introgression. We quantified phylogenetic discordance, inferred reticulate phylogenetic networks, and applied the Δ statistic using gene tree topologies reconstructed from multiple sequence alignments of protein-coding single-copy orthologs. In addition, we conducted ABBA-BABA tests on genome-wide single nucleotide variants and reconstructed a phylogeny of mitochondrial genomes. Our results revealed strong mito-nuclear discordance and considerable nonrandom gene tree discordance that cannot be explained by incomplete lineage sorting alone. Eight of the nine species examined demonstrated a history of introgression with at least one other species or ancestral lineage, indicating that introgression was common during the diversification of the strongylocentrotid urchins. There was strong support for introgression between four

extant species pairs (*Strongylocentrotus pallidus* ⇔ *S. droebachiensis*, *S. intermedius* ⇔ *S. pallidus*, *S. purpuratus* ⇔ *S. fragilis*, and *Mesocentrotus franciscanus* ⇔ *Pseudocentrotus depressus*) and additional evidence for introgression on internal branches of the phylogeny. Our results suggest that the existing gametic incompatibilities among the strongylocentrotid urchin species have not been a complete barrier to hybridization and introgression following speciation. Their continued divergence in the face of widespread introgression indicates that other reproductive isolating barriers likely exist and may have been more critical in establishing reproductive isolation early in speciation.

1.1 Introduction

The new availability of genome-scale data has stimulated considerable investigation into the genomic architecture of speciation - the number, kind, location, and relative effect size of loci underlying reproductive isolation. Understanding the genetic basis of speciation requires identifying these so-called "barrier loci" and characterizing the selective agents responsible for their divergence (Orr 2005). Although it is well established that reproductive isolation often evolves as a by-product of diversifying selection (Coyne and Orr 2004), the link between phenotypic divergence and the specific genetic changes underlying reproductive isolation remains weak (Schluter and Rieseberg 2022). One of the major outstanding questions concerns whether reproductive incompatibilities evolve more commonly from adaptive divergence or nonadaptive processes such as intragenomic conflict and divergent gene duplication resolution (Schluter and Rieseberg 2022). Contrary to the recent enthusiasm for ecological speciation, hybrid incompatibility loci are often associated with nonadaptive processes (Presgraves 2010; Maheshwari and Barbash 2011; Campbell et al. 2018). However, research seeking to identify barrier loci has been historically biased towards postzygotic isolation, which may be less likely to evolve from ecological selection than prezygotic isolation (Campbell et al. 2018). Broader taxonomic representation is needed because most conclusions have been drawn from a limited number of taxa (Campbell et al. 2018).

Studying speciation in the sea offers a unique opportunity to characterize the evolution of reproductive isolation in settings where geographic barriers are less

common. Especially compelling are the broadcast-spawning marine invertebrates, whose life histories and reproductive ecologies differ drastically from most animal speciation models. Broadcast spawners typically have massive fecundities and highly dispersive larvae, resulting in large population sizes and broad geographic ranges. Their high levels of gene flow across large distances and the rarity of absolute geographic barriers should limit opportunities for population differentiation and the evolution of reproductive isolating barriers (Palumbi 1994). Furthermore, broadcast spawners such as sea urchins lack pre-mating mechanical and behavioral drivers of reproductive isolation, and incipient species often show little morphological, ecological, or physiological divergence. Despite these constraints, species diversity in broadcast spawners appears high. One explanation for the high species richness observed in the absence of obvious physical barriers and ecological divergence is that the rapid evolution of a small number of reproductive proteins may establish reproductive isolation (Palumbi and Metz 1991; Palumbi 1992; Metz et al. 1994; Swanson and Vacquier 2002a; S. R. Palumbi 2009; Levitan et al. 2019).

Many species of broadcast spawners exhibit species-specific fertilization mediated by gamete recognition proteins (GRPs) located on the surfaces of sperm and egg cells (Summers and Hylander 1975; Vacquier and Moy 1977; Metz et al. 1994). These proteins often evolve rapidly under positive selection and have been implicated in the establishment of reproductive isolation (Lee and Vacquier 1992; Lee et al. 1995; Metz and Palumbi 1996; Biermann 1998a; Yang et al. 2000; Swanson and Vacquier 2002b; Swanson and Vacquier 2002a). Furthermore, gametic compatibility among sea

urchin species was found to be negatively correlated with sequence divergence of the sperm GRP bindin (Zigler et al., 2005), suggesting that bindin sequence similarity determines gametic compatibility. These discoveries reinforced the hypothesis that speciation in broadcast spawners may occur when diversifying selection at GRPs produces gametic incompatibility, leading to the classification of bindin and its egg receptor protein (EBR1) as speciation genes (Noor and Feder 2006; Nei and Nozawa 2011; Blackman 2016). Several mathematical models have shown that both allopatric and sympatric speciation are theoretically possible when sexual conflict mediated by polyspermy risk drives a coevolutionary chase between the sexes and causes GRP divergence (Gavrilets 2000; Van Doorn et al. 2001; Gavrilets and Waxman 2002; Gavrilets and Hayashi 2005). However, it remains unclear whether divergence at reproductive proteins caused speciation or instead accumulated after significant reproductive isolation had already evolved.

The strongylocentrotid sea urchin family is an ideal group for studying the evolution of reproductive isolation. Due to their translucent embryos, sea urchins became model organisms for fertilization studies during the late 19th century. Like many other marine species, sea urchins have large effective population sizes, broad geographic ranges, and limited population structure. The purple sea urchin, *Strongylocentrotus purpuratus* (Stimpson), is a member of the strongylocentrotid family and has a well-annotated reference genome in its fifth major revision. It is currently believed that the strongylocentrotid species are strongly reproductively isolated and have not shared alleles through introgression due to well-documented

gametic incompatibilities and the rarity of natural hybrids (Strathmann 1981; Lessios 2007). However, recent studies indicate that reproductive isolation may be incomplete, evidenced by introgression between *S. pallidus* (Sars) and *S. droebachiensis* (O. F. Müller) in the Northeast Pacific (Addison and Hart 2005a; Harper et al. 2007a; Addison and Pogson 2009a; Pujolar and Pogson 2011) and Northwest Atlantic (Addison and Hart 2005a; Harper et al. 2007a). Whether other strongylocentrotid taxa have experienced introgression remains unknown.

If gametic isolation were an important isolating barrier early in strongylocentrotid speciation events, evidence of introgression should be rare and negatively correlated with phylogenetic distances and gametic incompatibilities. We tested these predictions using whole-genome sequencing data from the strongylocentrotid urchin species and multiple complementary phylogenomic approaches. Given the documented susceptibility of *S. droebachiensis* eggs to heterospecific sperm (Levitan 2002a) and the previous finding of *S. pallidus* alleles in *S. droebachiensis* individuals (Addison and Pogson 2009a), we expected to find a signal of introgression between *S. droebachiensis* and other congeners. Further predictions about introgression were challenging for several reasons. First, heterospecific cross data only exists for a few strongylocentrotid species pairs. Second, although fertilization is more efficient in conspecific crosses of strongylocentrotid urchins (Strathmann 1981; Minor et al. 1991; Levitan 2002a), heterospecific fertilizations readily occur in no-choice experiments between highly divergent species (Newman 1923; Moore 1957a; Zhao et al. 2021). Furthermore, whether hybrid matings

occur *naturally* depends heavily upon the distance between a female urchin and the nearest conspecific male (Levitan 2002a), and little is known about the fitness of hybrid offspring in most heterospecific crosses.

Contrary to our expectation of limited introgression, we found widespread introgression across the strongylocentrotid family at multiple time scales, suggesting that gametic incompatibilities have not been an effective barrier to introgression. The existing gametic incompatibilities either weren't strong enough to prevent significant introgression or evolved after significant introgression had already occurred, both of which are inconsistent with gametic isolation establishing reproductive isolation and causing speciation. Our findings indicate that additional reproductive barriers must have been in place for the establishment and maintenance of species barriers.

1.2 Materials and Methods

1.2.1 Study System

The strongylocentrotid phylogeny comprises two major clades: Clade S includes *Strongylocentrotus* and *Hemicentrotus*; Clade M includes *Mesocentrotus* and *Pseudocentrotus*. Both *Hemicentrotus* and *Pseudocentrotus* are monotypic genera. The phylogeny is parsimoniously consistent with a Western Pacific common ancestor and at least two independent Eastern Pacific colonizations (Kober & Bernardi, 2013). Four species are limited to the Northwest Pacific: *P. depressus* (A. Agassiz), *M. nudus* (A. Agassiz), *H. pulcherrimus* (A. Agassiz), and *S. intermedius* (A. Agassiz). An additional two species, *S. pallidus* and *S. droebachiensis*, are found in, but not limited to, the

Northwest Pacific. Five species co-occur in the East Pacific with overlapping geographic ranges, depth preferences, and spawning seasons: *S. droebachiensis*, *S. fragilis* (Jackson), *S. pallidus*, *S. purpuratus*, and *M. franciscanus* (A. Agassiz). *S. droebachiensis* and *S. pallidus* have further expanded their ranges, crossing the Bering Sea to colonize the Arctic Ocean and the West and East Atlantic. These two species show little differentiation between the Pacific and Atlantic Oceans, likely due to stepping-stone populations that facilitate gene flow (Palumbi and Kessing 1991).

1.2.2 Whole Genome Resequencing and Data Pre-Processing

The genomes of all strongylocentrotid species had been previously sequenced at high coverage depth with the Illumina HiSeq 2500 (Kober & Bernardi, 2013; Kober & Pogson, 2017). The raw sequencing reads were deposited in the NCBI Sequence Read Archive under BioProject PRJNA391452. Metadata for the genome samples is available in Table S1. The sequencing reads were pre-processed with Picard (Broad Institute 2018) and GATK v4.2.6.1 following GATK's Best Practices (Van der Auwera et al. 2013). Adapter sequences were marked using Picard MarkIlluminaAdapters, sequencing reads were mapped to the *S. purpuratus* reference genome (Spur_5.0) using bwa-mem2 v2.2.1 (Vasimuddin et al. 2019), and duplicate reads were marked with Picard MarkDuplicates. Reference mapping and alignment were evaluated using samtools flagstat (Danecek et al. 2021) and mosdepth v0.3.3 (Pedersen and Quinlan 2018).

Variant calling and joint genotyping were performed using GATK's HaplotypeCaller and GenotypeGVCFs. Variant quality filtering was performed independently for each subset of species used in downstream analyses. Vcf files were hard-filtered for variants with skewed values across all samples following GATK recommendations. Single nucleotide variants (SNVs) were filtered that had low quality (QUAL<30), low map quality (MQ<40), low quality by depth scores (QD<2), high fisher strand scores (FS > 60), high strand odds ratios (SOR>3), low mapping quality rank sum scores (MQRankSum<-12.5), or low read position rank sum scores (ReadPosRankSum<-8). Indels were filtered that had low quality (QUAL<30), low quality by depth scores (QD<2), high fisher strand scores (FS>200), or low read position rank sum scores (ReadPosRankSum<-20.0). Furthermore, individual genotypes with low quality (GQ<20) or low read depth (DP<3) were set to missing, and SNVs within three base pairs of an indel were filtered.

1.2.3 Phylogenetic Relationships and Concordance Factor Statistics

For phylogenetic inference, multiple sequence alignments were created for protein-coding single-copy orthologs inferred by filtering *S. purpuratus* nuclear gene models by coverage depth. Genes were filtered if any sample had a mean depth lower than 10x, a mean depth greater than double the sample's mean depth for *S. purpuratus* exons, or fewer than 75% of the bases in the gene covered by ten reads. To account for nonindependence among loci, genes were filtered so that there was a minimum of 20kb between included loci. Multiple sequence alignments of concatenated CDS were

created for each gene passing filter by applying the hard-filtered SNVs and deletions to the *S. purpuratus* reference sequence using vcf2fasta (Sanchez-Ramirez, 2017). Insertions were ignored to keep gene coordinates consistent with the *S. purpuratus* reference. After creating the fasta alignments, genes were excluded if they had no parsimony informative sites or if their length was not a multiple of three.

A maximum-likelihood species tree was inferred using the edge-linked partition model of IQ-TREE (Nguyen et al. 2015; Chernomor et al. 2016) on the concatenated single-copy ortholog fasta alignments. Branch supports were obtained using ultrafast bootstrap with 1,000 replicates (Hoang et al. 2018). Single locus trees were reconstructed for each single-copy ortholog fasta alignment using IQ-TREE's ModelFinder (Kalyaanamoorthy et al. 2017).

Gene concordance factor (gCF) and site concordance factor (sCF) statistics (Minh et al. 2020) were calculated for each branch in the species tree to quantify the amount of phylogenetic discordance present in the data. For each branch in the species tree, the gCF measures the proportion of gene trees containing that branch, while the sCF measures the proportion of informative sites concordant with that branch. The sCFs were calculated by randomly sampling 300 quartets around each internal branch in the phylogeny using an updated version of sCF based on maximum likelihood implemented in IQ-TREE v2.2.2 (Mo et al. 2023). In addition to the gCF and sCF values, IQ-TREE also calculates the frequencies of the two discordant trees produced by nearest-neighbor interchanges (NNI) around each branch. Coalescent theory predicts that the two discordant trees should be equally observed if the discordance is

caused by incomplete lineage sorting (ILS) only. However, one tree may become more frequent than the other if introgression has occurred. To test for introgression, chi-square tests were used to compare counts of the two discordant NNI trees for each branch in the species tree.

1.2.4 Mitochondrial Phylogenetics

To investigate the relationships between mitochondrial genomes and look for signs of introgression, mitochondrial genomes were assembled for the same samples used in the species tree inference (Kober and Bernardi, 2013; Kober and Pogson, 2017). Metadata for the mitochondrial genomes is available in Table S2. The *S. purpuratus* sample used was from the original reference genome assembly (NC_001453.1; Jacobs et al., 1998). The sequences were aligned with Clustal Omega v1.2.3 (Sievers et al., 2011; Sievers & Higgins, 2018), and a maximum likelihood tree was created with IQ-TREE using ModelFinder. Branch supports were obtained using ultrafast bootstrap with 10,000 replicates.

1.2.5 Tests for Introgression

Recent powerful phylogenomic approaches for characterizing introgression based on the multi-species coalescent (MSC) model make it possible to detect introgression with just a single genome sample per species (Hibbins and Hahn 2022). Due to limited *a priori* hypotheses about which species may have experienced introgression, we implemented several independent tests for introgression based on

gene tree discordance that use different inference methods. Patterson's *D* statistic uses genome-wide counts of biallelic site patterns (Green et al. 2010; Durand et al. 2011), the Δ statistic uses genome-wide counts of gene genealogies (Huson et al. 2005), and PhyloNet uses maximum likelihood to estimate reticulate phylogenies using distributions of gene genealogies (Than et al. 2008; Nguyen et al. 2015).

1.2.5.1 Patterson's *D* Statistic

Patterson's *D* statistic, or the ABBA-BABA test, is the most widely used summary statistic in introgression studies and is robust in a wide parameter space (Zheng and Janke 2018; Kong and Kubatko 2021). Patterson's *D* statistic tests for a genome-wide imbalance in the counts of the biallelic site patterns consistent with the two possible discordant topologies in a rooted triplet (Green et al. 2010; Durand et al. 2011). Significance for *D* is calculated using a block jackknife approach that accounts for nonindependence among sites in the data. Patterson's *D* statistic was calculated for all phylogenetically relevant triplets using the genome-wide genotype call set and the Dsuite Dtrios program (Malinsky et al. 2021) with a block-jackknife size of 1 Mb. For comparisons within the S clade, separate tests were run with *M. nudus*, *M. franciscanus*, and *P. depressus* as outgroups. For the test within the M clade, *S. purpuratus* and *S. fragilis* were used as the outgroup. A recent addition to Patterson's *D*, D_p, can approximate the genome-wide introgression proportion (Hamlin et al. 2020) and was calculated for each triplet using the Dsuite output. To determine whether introgression is correlated with phylogenetic distance or GRP divergence, we performed linear

regressions of mean Patterson's *D* and D_p by overall phylogenetic distance, bindin distance, and EBR1 distance (Appendix A).

1.2.5.2 Δ Statistic

The Δ statistic is an alternative approach to Patterson's *D* that uses counts of discordant gene tree topologies rather than site patterns (Huson et al. 2005). Δ is less sensitive to the assumption of Patterson's *D* that there have not been multiple substitutions per site (Hahn 2018) and was used as a secondary measure to confirm significant Patterson's *D* statistic tests where introgression must have occurred between extant taxa. Δ was estimated using gene tree topologies reconstructed from multiple sequence alignments of single-copy orthologs for three different quartets: ((((*M. nudus*, *M. franciscanus*), *P. depressus*), *S. purpuratus*); ((((*S. droebachiensis*, *S. pallidus*), *S. intermedius*), *M. franciscanus*); (((*S. fragilis*, *S. droebachiensis*), *S. pallidus*), *M. franciscanus*). Significance was assessed by calculating Δ for 10,000 pseudoreplicate datasets created by resampling the gene tree topologies with replacement (Vanderpool et al. 2020).

1.2.5.3 PhyloNet

The PhyloNet software package implements a powerful set of likelihood methods based on the multispecies network coalescent (MSNC) model (Meng and Kubatko 2009) that can be used to formally test for introgression (Than et al. 2008; Wen et al. 2018). PhyloNet programs can identify introgression on the internal

branches of a phylogeny and reliably infer the direction of introgression (Hibbins and Hahn 2022). To further characterize the history of introgression within the strongylocentrotid family, we ran PhyloNet's InferNetwork_ML program (Yu et al. 2014) with reconstructed gene tree topologies to infer phylogenetic networks with reticulation edges representing discrete introgression events. A smaller subset of species was used in the PhyloNet analysis due to computational constraints and the requirement that the gene trees be rooted. A new set of single-copy orthologs was inferred for *M. franciscanus*, *H. pulcherrimus*, and the five *Strongylocentrotus* taxa (Table S10). Gene trees were estimated with IQ-TREE2, and 100 bootstrap trees were generated for each gene using standard nonparametric bootstrap to account for uncertainty in gene tree reconstruction. InferNetwork_ML was run to infer phylogenetic networks with 0, 1, 2, and 3 reticulations.

1.3 Results

1.3.1 Data Pre-processing

The results of the reference genome mapping are summarized in Table 1.1. The read mapping percentage per sample ranged from 76% to 98%. Mean genome-wide coverage depth typically ranged from 18x - 32x, except for *S. purpuratus* and *S. pallidus*. Coverage depth for *S. pallidus* (12x) was lower because of a reduced library complexity resulting from the early developmental phase of automated library preparation protocols (Kober and Pogson 2017). *S. purpuratus* was sequenced at a higher depth (91x) for reference genome assembly. Mean coverage depth increased to

>38x for protein-coding single-copy orthologs, except for *S. pallidus* (15x). Additional coverage metrics are presented in tables S3-S5.

1.3.2 Phylogenetic Discordance among Strongylocentrotids

Although the inferred maximum likelihood species tree topology agreed with the topology produced by Kober & Bernardi (2013), the gene and site concordance factor statistics revealed extensive phylogenetic discordance on most species tree branches (Figure 1.1a, Table S6). The three internal branches relating the *Strongylocentrotus* species had very low gCF and sCF values. These branches are short, and the lower gCF values than sCF values signal that error in gene tree reconstruction likely contributed to the observed signal of phylogenetic discordance. However, the low sCF values suggest that there is not overwhelming support for any single resolution of these branches, implying considerable ILS and introgression. Although the low gCF values may be partially explained by error in gene tree reconstruction, biases in the frequencies of the discordant topologies are suggestive of introgression (Figure 1.1b, Table S6). For the branch in the species tree placing *S. purpuratus* as the outgroup to the rest of the *Strongylocentrotus* species (Branch C), the discordant resolution placing *S. intermedius* as the first diverging member of *Strongylocentrotus* (15.9% gene trees, 34.5% sites) was observed more frequently than the other NNI discordant resolution (13.3% gene trees, 29.7% sites, p=0.0015), indicating introgression between *S. purpuratus* and one or more of *S. pallidus*, *S. droebachiensis*, *S. fragilis*, or an ancestral lineage. Three other branches also had a discordant topology that was significantly

overrepresented (Branches D, E, F), implying introgression between *S. intermedius* ⇔ *S. pallidus*, *S. pallidus* ⇔ *S. droebachiensis*, and *P. depressus* ⇔ *M. franciscanus* (Figure 1.1b).

1.3.3 Mitochondrial Introgression

The phylogeny of the mitochondrial genome accessions did not recover the true species relationships, showing several discordant patterns consistent with introgression (Figure 1.2). *M. franciscanus* clustered with *P. depressus* with 99 percent bootstrap support rather than with its sister taxon, *M. nudus*. Similarly, *S. droebachiensis* clustered with *S. pallidus* with 99 percent bootstrap support rather than its sister taxon, *S. fragilis*. The last source of discordance was the placement of *S. purpuratus* and *S. intermedius*. In the mitochondrial tree, the positions of *S. purpuratus* and *S. intermedius* are swapped relative to the species tree, consistent with gene flow between *S. purpuratus* and one or more of *S. pallidus*, *S. droebachiensis*, *S. fragilis,* or an ancestral lineage. All three of these discordant topologies were also overrepresented in the gene concordance factor analysis, indicating that the mito-nuclear discordance observed was caused by introgression.

1.3.4 Introgression Tests

1.3.4.1 Patterson's *D* Statistic

Seventeen of the twenty-one Patterson's *D* tests were significant, implicating ten independent species pairs in introgression (Figure 1.3, Table 1.2). For simplicity,

only the results with *M. nudus* and *S. purpuratus* as the outgroup are displayed (Figure 1.3, Table 1.2). However, the results were consistent regardless of the outgroup choice, and the full results are provided in Tables S7-9. In the M clade, there was support for introgression between *P. depressus* and *M. franciscanus*. In the S clade, there was evidence for introgression between *H. pulcherrimus* and each of *S. intermedius*, *S. pallidus*, *S. droebachiensis*, and *S. fragilis*. There was also support for introgression between *S. purpuratus* and each of *S. pallidus*, *S. fragilis*, and *S. droebachiensis*. Two additional species pairs were implicated in introgression: *S. intermedius* and *S. pallidus*, and *S. pallidus* and *S. droebachiensis*. In cases where a taxon shows introgression with several species that form a monophyletic group, it may be more parsimonious to assume that introgression occurred between that taxon and the MRCA of the monophyletic group, an internal branch in the phylogeny (Suvorov et al. 2022). For example, it's likely that *H. pulcherrimus* experienced introgression with the common ancestor of the four youngest *Strongylocentrotus* taxa rather than with each of them independently. Similarly, the significant tests involving *S. purpuratus* could have been produced by a single introgression event between *S. purpuratus* and the MRCA of *S. pallidus*, S. *droebachiensis,* and *S. fragilis*. This would reduce the total number of introgression events from ten to five, a conservative number because introgression could have occurred both on the internal and terminal branches.

We found no significant correlations between Patterson's *D* and overall phylogenetic distance, bindin distance, and EBR1 distance (Appendix A) Furthermore, when only including *Strongylocentrotus* species, we found a significant, positive

correlation between introgression (Patterson's D, D_p) and overall phylogenetic distance. The two *Strongyloentrotus* species pairs with the highest overall phylogenetic distances also had the highest mean values of Patteron's D and D_p. (*S. purpuratus* - *S. fragilis*, *S. purpuratus* - *S. droebachiensis*).

1.3.4.2 Δ Statistic

Δ was significantly positive for each of the three quartets tested, signaling introgression between *P. depressus* and *M. franciscanus*, *S. intermedius* and *S. pallidus*, and *S. pallidus* and *S. droebachiensis* (Table 1.3). All three test results were consistent with the estimated Patterson's D statistics (Figure 1.3, Table 1.2).

1.3.4.3 PhyloNet

The PhyloNet analysis revealed similar patterns of introgression to the Patterson's D and Δ statistics. Conditioning on the species tree backbone, the one-reticulation edge phylogenetic network with the highest likelihood implied introgression from *S. purpuratus* into *S. fragilis* (Figure 1.4a). The D statistic with the highest magnitude also demonstrated introgression between *S. purpuratus* and *S. fragilis* (Figure 1.3, Table 1.2). The network with the next highest likelihood implied introgression between *S. purpuratus* and the *S. droebachiensis*-*S. fragilis*-*S. pallidus* MRCA (Figure 1.4b), consistent with the gene concordance factor analysis and mitochondrial phylogeny. The best network with two reticulation edges had an additional edge implying introgression from *S. intermedius* into *S. pallidus* (Figure

1.4c), and the network with three reticulation edges added a third edge indicating introgression from the MRCA of *S. intermedius, S. pallidus*, *S. droebachiensis*, and *S. fragilis* into *H. pulcherrimus* (Figure 1.4d).

1.4 Discussion

1.4.1 Widespread Introgression among the Strongylocentrotid Urchins

Our study is the first to describe genome-wide patterns of introgression among sea urchins. It is currently believed that only limited introgression has occurred among sea urchins, but the results of our study indicate that it may be common, at least within *Strongylocentrotidae*. The ubiquity of introgression among the strongylocentrotid taxa suggests that gametic isolation has not been an effective barrier to introgression and may not have played a major role in speciation.

Our tests for introgression revealed that eight out of the nine species included in the study experienced introgression with at least one other species or ancestral lineage. The introgression patterns are clear and consistent regardless of the methodology used (Table 1.4). A minimum of six introgression events is supported by the data and is a conservative estimate for several reasons. First, we collapsed all tests where a species showed introgression with multiple species forming a monophyletic group. Second, it was not possible to test for introgression between the two pairs of sister taxa as methods relying on phylogenetic discordance cannot detect introgression between sister taxa. Third, we could not rule out introgression in the one species that did not show introgression (*M. nudus*) because the only taxa triplet we could test in the

M clade, ((*M. nudus*, *M. franciscanus*), *P. depressus*), implied significant introgression between *P. depressus* and *M. franciscanus*. Finally, we could not test for introgression between the M and S clade members without high-quality sequence data from a close outgroup to the family. We stress that these are historical introgression events in which the genomic signal has been preserved for millions of years in most cases. Given (i) the methods employed here test for ancient introgression, (ii) introgression is likely not ongoing in most cases, and (iii) only a single diploid genome per species was sampled, we find it likely that the observed signal of introgression was driven by introgressed variation that has been fixed. Furthermore, given that population structure is nearly non-existent in these sea urchin species (Palumbi and Wilson 1990; Palumbi and Kessing 1991), it is likely that most populations and individuals of introgressed taxa would show a similar signal of introgressed ancestry.

Despite considerable phylogenetic discordance in the underlying data, there was strong support for all branches in the strongylocentrotid species tree. This is unsurprising given that these species are well-diverged, with the youngest pair of sister taxa evolving 4-6 million years ago (Kober & Bernardi, 2013). Incomplete lineage sorting is expected to be pervasive in species with high levels of polymorphism, and the five *Strongylocentrotus* taxa speciated relatively rapidly 4-9 mya (Kober & Bernardi, 2013), resulting in short internal branches. However, incomplete lineage sorting alone is insufficient to explain the observed discordance patterns.

The *D,* Δ, and gCF/sCF statistics implied introgression between at least three pairs of extant taxa: *S. pallidus* ⇔ *S. droebachiensis*, *S. intermedius* ⇔ *S. pallidus*, and

P. depressus ⇔ *M. franciscanus*. Introgression between *S. purpuratus* and *S. fragilis* also likely occurred, but the signal could also be explained by introgression on an internal branch. The mitochondrial phylogeny supported two of these introgression events (*S. pallidus* ⇔ *S. droebachiensis, P. depressus* ⇔ *M. franciscanus*), and the PhyloNet analysis supported introgression between *S. intermedius* and *S. pallidus*, and *S. purpuratus* and *S. fragilis*.

Due to limitations in the fossil record, little is known about the geography of strongylocentrotid urchin speciation and the historical ranges of its extant taxa. However, the patterns of introgression help fill in some of these gaps by demonstrating that some currently allopatric species showing signals of introgression must have had overlapping ranges in the past. For example, the strong signal of introgression between *P. depressus* and *M. franciscanus* was unexpected, given that the ranges of these two species are currently separated by an ocean basin. The M clade phylogeny of the strongylocentrotid family is consistent with a West Pacific common ancestor (Kober & Bernardi, 2013), followed by the colonization of the East Pacific by *M. franciscanus*. Therefore, introgression must have occurred at a time of range overlap in the distant past, implying that *M. franciscanus* speciated in the West Pacific, interbred with sympatric *P. depressus* before colonizing the East Pacific, and later became locally extinct in the West Pacific.

It was similarly unexpected to find support for introgression between *S. intermedius* and *S. pallidus,* given their current distributions. Although *S. intermedius* and *S. pallidus* co-occur in the Sea of Japan, the *S. pallidus* sample used in this study

was from coastal Washington State, indicating that the signal of introgression is ancient. The net direction of gene flow inferred by PhyloNet was from *S. intermedius* into *S. pallidus*, implying that introgression must have occurred before *S. pallidus* expanded its range into the East Pacific. Whether introgression is ongoing between *S. intermedius* and *S. pallidus* in the Sea of Japan is unknown.

In addition to introgression between extant taxa, it also likely occurred between extant taxa and ancestral lineages (i.e., internal branches). While the optimal phylogenetic network with one reticulation edge implied introgression from *S. purpuratus* into *S. fragilis*, a second network with a similar likelihood supported introgression from the *S. droebachiensis*-*S. fragilis*-*S. pallidus* MRCA into *S. purpuratus*. Both networks are consistent with the Patterson's *D* statistic results as there was support for introgression between *S. purpuratus* and each of *S. droebachiensis*, *S. fragilis*, and *S. pallidus*. Both the mitochondrial phylogeny and the concordance factor analysis were also consistent with introgression on an internal branch. In the mitochondrial phylogeny, *S. purpuratus* is pulled down as a sister to the *S. droebachiensis*-*S. fragilis*-*S. pallidus* MRCA and the concordance factor analysis revealed that this topology was overrepresented. A similar potential case of introgression on an internal branch was evidenced by the optimal phylogenetic network with three reticulation edges, which implied introgression between *H. pulcherrimus* and the MRCA of *S. intermedius*, *S. pallidus*, *S. fragilis*, and *S. droebachiensis*. The results of the phylogenetic network analyses underscore the importance of sampling all species of the focal genus or family when testing for introgression. By only sampling a

subset of the taxa, introgression may be incorrectly attributed to extant taxa in cases where it occurred on internal branches of the phylogeny. If introgression did occur on an internal branch, there should be considerable overlap in the location of introgressed DNA in each species descendent from that branch.

There are several limitations in the approaches we used to test for introgression. First, it is difficult to quantify the proportion of the genome that is introgressed in each scenario without polymorphism data or populations that are known *a priori* to have not experienced introgression. However, the D_p statistic and the PhyloNet reticulation edge weights provide reasonable estimates. Second, the geographic history of speciation, hybridization, and introgression is challenging to interpret given the old divergence times of this group, its limited fossil record, and the fact that current ranges of the extant taxa may not be representative of their past distributions. This limitation applies to many other marine invertebrate clades due to limitations in the fossil record and shifting ranges due to cycles of sea level rise-and-fall (S. R. Palumbi 2009). Furthermore, the geographic pattern of hybridization and introgression may be especially complex for marine organisms with high dispersal potential because hybrid zones are more ambiguous.

Our study adds further representation of marine invertebrates to the rapidly growing evidence for hybridization and introgression and will facilitate investigations into how patterns of introgression vary across different organismal groups. Introgression had long been recognized as a significant evolutionary force in plants (Anderson and Hubricht 1938; Anderson and Stebbins 1954) but was only recently appreciated in

animals (Hedrick 2013). Historically, it was thought that introgression between marine taxa was rare (Arnold and Fogarty 2009) and had not occurred among sea urchins (Lessios 2007). However, reticulate evolution in marine systems may be as common as that of non-marine taxa (Gardner 1997), but the difficulty in collecting and observing marine organisms has limited its detection (Arnold and Fogarty 2009). Although hybridization has been detected in at least five genera of sea urchins (*Diadema*: Lessios & Pearse, 1996, *Lytechinus*: Zigler & Lessios, 2004, *Strongylocentrotus*: Addison & Pogson, 2009, *Pseudoboletia*: Zigler et al., 2012, *Arbaci*a: Lessios et al., 2012), this is the first study that has tested for introgression among sea urchins using genome-scale data. Among other broadcast spawners, introgression has been detected in *Acropora* corals (Mao et al. 2018), *Mytilus* mussels (Saarman and Pogson 2015; Fraïsse et al. 2016; Vendrami et al. 2020; Popovic et al. 2021; Simon et al. 2021), *Ophioderma* brittle stars (Weber et al. 2019), *Asterias* sea stars (Harper and Hart 2007), Western Pacific *Haliotis* abalones (Hirase et al. 2021), and *Ciona* sea squirts (Nydam and Harrison 2011; Nydam et al. 2017).

1.4.2 On the Relative Importance of Gametic Isolation

It is currently believed that the rapid evolution of gamete recognition proteins (GRPs) is a major contributor to reproductive isolation among broadcast spawners. Although reproductive proteins evolve rapidly under positive selection in a wide variety of taxa (Swanson and Vacquier 2002a), it remains unclear how often this rapid evolution establishes reproductive isolation and causes speciation (Turner and

Hoekstra 2008). Among sea urchins, gametic compatibility can sometimes be maintained for up to five million years and is rarely a bi-directional barrier to hybridization (McCartney & Lessios, 2004; Zigler et al., 2005). Asymmetric gametic incompatibilities may be the rule rather than the exception (Zigler et al., 2005) and are incapable of preventing gene flow between incipient species (McCartney and Lessios 2004; Addison and Pogson 2009a; Lessios 2011), suggesting the importance of additional barriers. Furthermore, bindin is not one of the fastest-evolving sea urchin genes and only shows evidence of positive selection in three of the seven sea urchin genera studied to date (Geyer et al. 2020a). The drivers of selection at bindin are poorly understood and vary across the three genera showing positive selection (*Echinometra*: Metz & Palumbi, 1996; Geyer & Palumbi, 2003; McCartney & Lessios, 2004, *Heliocidaris*: Zigler et al., 2003, *Strongylocentrotus*: Biermann, 1998; Pujolar & Pogson, 2011). In some cases, the selective agent appears to be reinforcement, while in others, it's not clear that the selection at bindin has established sufficient reproductive isolation for the formation of new species.

Within *Strongylocentrotidae*, gametic compatibility *between species* is likely determined by variation in the selective pressures acting on gamete traits *within species* because intraspecific density-dependent selection acting on gamete traits to maximize fecundity and limit polyspermy also influences susceptibility to heterospecific fertilization (Levitan 2002b; Levitan 2002a; Levitan et al. 2007). Species that more commonly experience sperm-limiting conditions are selected for high fertilization rates and produce eggs that are more readily fertilized by both conspecific and heterospecific

sperm. Conversely, species with higher population densities and high sperm availability likely evolve under sexual conflict and produce faster, more competitive sperm and more sperm-resistant eggs. This density-dependent selection has likely led to the asymmetric gametic incompatibilities observed between *S. droebachiensis* and other congeners (Hagström and Lönning 1967; Strathmann 1981; Levitan 2002a) and may have also resulted in asymmetric introgression (Addison and Pogson 2009a). Under the scenario of density-dependent selection on sperm and egg traits, reproductive isolation between populations should only be strengthened in times or locations of high spawning density. When spawning density is low and populations experience sperm limitation, purifying selection to maximize mating opportunities should favor more easily fertilized eggs and prevent divergence of GRPs.

Field experiments on *S. droebachiensis* in the Barkley Sound have demonstrated that gametic isolation is not an effective barrier to hybrid matings when spawning females are closer to heterospecific males than conspecific males (Levitan 2002a). Hybrid fertilizations readily occur when *S. droebachiensis* eggs are swamped by heterospecific sperm, suggesting that some spatial or temporal isolation during spawning is required to prevent hybridization. Work in other broadcast spawner groups has shown that reproductive isolation can evolve without gamete recognition barriers. For example, ecological divergence evolved before GRP divergence in the Western Pacific abalones and maintains species barriers despite ongoing hybridization and introgression (Hirase et al. 2021). In another case, strong reproductive isolation has

evolved between the Australian sea urchin species *Pseudoboletia indiana* and *P. maculata* despite only a single amino acid substitution at bindin (Zigler et al. 2012).

The extensive introgression observed among the strongylocentrotid urchins and the lack of a significantly negative correlation between introgression signal and phylogenetic distance, bindin distance, or EBR1 distance indicates that gametic incompatibilities either weren't strong enough to prevent significant introgression or evolved after significant introgression had already occurred. Both scenarios are inconsistent with gametic isolation commonly establishing reproductive isolation and causing speciation, suggesting that the GRPs bindin and EBR1 are not speciation genes in the strongylocentrotid family. Other isolating barriers were likely in place and should be investigated further to understand the genetic basis of speciation in strongylocentrotid urchins and other broadcast spawners. Lessios (2007) reviewed isolating barriers in sea urchins and concluded that each prezygotic barrier alone appeared incapable of preventing gene flow between sympatric species. Unfortunately, the relative strength of different isolating barriers has rarely been quantified in pairs of sea urchin sister taxa (S. R. Palumbi 2009).

1.4.3 Possible Alternative Isolating Mechanisms

1.4.3.1 Postzygotic Isolation

How does speciation proceed in high gene flow marine invertebrates with minimal population structure and ecological divergence when geographic barriers are seemingly limited? One possibility is that some postzygotic isolation evolves in

allopatry before the evolution of gametic isolation. There are well-documented cases of hybrid sterility and inviability in interspecific crosses of strongylocentrotid urchins. For example, the *M. nudus* ♀ x *S. intermedius* ♂ cross is lethal (Ding et al. 2007). Although the reciprocal cross produces viable offspring, hybrid larval survival, metamorphosis rates, and juvenile survival are significantly lower than conspecific controls. Furthermore, the surviving juveniles produce very few or no mature gamete cells, a pattern also observed in the *Hemicentrotus pulcherrimus* ♀ x *S. intermedius* ♂ cross (Liu et al. 2020).

In crosses of *S. droebachiensis* x *S. pallidus*, Hagström & Lönning (1967) found that chromosomal abnormalities were frequent during mitosis in embryos of F1 hybrids. Strathmann (1981) performed ten separate reciprocal crosses between *S. droebachiensis* and *S. pallidus*, but only four hybrids survived to the three-year mark when spawning was induced, and all were female. The female hybrids were successfully backcrossed in both directions, although backcross fertilization success was much higher with *S. pallidus* males than with *S. droebachiensis* males. Reduced survival of hybrid juveniles has also been found in crosses of female *S. droebachiensis* with male *S. purpuratus* and *M. franciscanus* (Levitan 2002a) and the cross between *S. purpuratus* and *M. franciscanus* (Newman 1923). Postzygotic isolation may be even stronger than these studies suggest because intrinsic postzygotic isolation may not appear until generations beyond the F1 if the alleles that cause intrinsic postzygotic isolation are partially recessive in hybrids (Coyne and Orr 2004). Reproductive barriers

may also result from extrinsic (i.e., ecological) postzygotic isolation produced by a mismatch between hybrid individuals and their environment.

1.4.3.2 Chemical Barriers and Carbohydrate-Based Gamete Recognition

The possibility that chemical barriers contribute to reproductive isolation has received limited attention. The egg jelly of broadcast spawners often serves as a chemoattractant to guide conspecific sperm towards the egg, a process called sperm chemotaxis. Conspecific chemoattractant preference has been demonstrated in the abalone species *H. rufescens* and *H. fulgens* (Riffell et al. 2004), although the interaction of gamete recognition proteins is a better predictor of fertilization success in these species (Evans and Sherman 2013). Sperm chemotaxis has also been described in the sea urchins *Arbacia puctulata* (Ward et al. 1985), *Lytechinus pictus* (Guerrero et al. 2010), and *S. purpuratus* (Ramírez-Gómez et al. 2020).

In sea urchin fertilization, the acrosome reaction is a precondition for the binding of sperm to the egg and may also be species-specific in some cases. Alves et al. (1997) found that sulfated polysaccharides in the egg jelly induce the acrosome reaction in a conspecific manner, although the three species tested were quite divergent (*Echinometra lucunter*, *Arbacia lixula*, and *Lytechinus variegatus*). Biermann et al. (2004) similarly found that the jelly coat of *S. droebachiensis* eggs only induces the acrosome reaction in conspecific sperm due to the rapid evolutionary change in the *S. droebachiensis* egg-jelly fucan. Furthermore, *S. droebachiensis* sperm react with *S. pallidus* and *S. purpuratus* eggs at considerably lower rates than with conspecific eggs.

However, the acrosome reaction is not species-specific between *S. purpuratus*, *M. franciscanus*, and *S. pallidus* (Biermann et al. 2004) or between *Echinometra mathaei* and *Echinometra oblonga* (Metz et al. 1994).

1.4.3.3 Habitat and Temporal Isolation

While differences in habitat preference or spawning time could prevent most heterospecific gamete encounters, sea urchin species' ranges commonly overlap, and it is believed that the cues of spawning cycles are too spatially or temporally variable for spawning asynchrony to be an effective barrier (Lessios 2007). However, species often show depth zonation in areas of range overlap (Lessios 2007), and slight differences in the timing and location of gamete release among congeners could prevent heterospecific fertilization as sperm rapidly age, disperse, and become diluted following release (Pennington 1985; Levitan 1993; Levitan et al. 2004). A short gap in peak spawning times is an effective reproductive barrier for a pair of Panamanian *Montastraea* reef-building corals (Knowlton et al. 1997) and a pair of Australian subspecies of *Heliocidaris erythrogramma* (Binks et al. 2012). Furthermore, genetic differences in habitat preference were shown to isolate two *Mytilus* mussel species in a contact zone in southern France (Bierne et al. 2003).

1.5 Conclusions

Although gametic incompatibilities may help maintain species boundaries in strongylocentrotid urchins, gametic isolation does not appear to have been an effective

barrier to introgression. The long persistence of gametic compatibility between divergent taxa and evidence of extensive introgression within the family are inconsistent with the rapid evolution of gametic isolation being an important mode of speciation in this family. Additional isolating barriers likely evolved earlier and were more critical in establishing reproductive isolation. The continued divergence of the strongylocentrotid species in the face of significant introgression emphasizes the importance of postzygotic isolation in maintaining species integrities.

1.6 Tables and Figures

Table 1.1. Summary of genomic DNA sequencing, reference mapping, and coverage.

Species	Reference Mapping			% Bases Covered			Mean Coverage Depth		
	Raw Reads	Mapped%	Proper Pair %	Whole Genome[a]	Coding[b]	Single-Copy Orthologs 10x[c]	Whole Genome[d]	Coding[e]	Single-Copy Orthologs[f]
Sdro	3.04E+08	91.74%	78.11%	78%	92%	0.97	24.7x	41.5x	42.5x
Sfra	3.97E+08	89.87%	78.21%	81%	93%	0.97	32.1x	46.8x	48.2x
Spal	1.50E+08	91.82%	72.39%	78%	91%	0.97	11.9x	15x	15.5x
Sint	4.01E+08	84.24%	73.06%	77%	91%	0.97	28.3x	44.2x	50.3x
Spur	6.21E+08	98.11%	89.04%	99%	100%	0.99	91.3x	100.3x	108.2x
Hpul	3.76E+08	82.71%	68.67%	69%	86%	0.95	24.5x	44.3x	53.3x
Mnud	3.82E+08	77.00%	63.08%	58%	82%	0.92	21.1x	40.5x	45.3x
Mfra	3.39E+08	80.36%	64.30%	60%	84%	0.93	19.9x	33.8x	38.3x
Pdep	3.28E+08	76.17%	60.79%	50%	77%	0.89	18.1x	47.5x	53.5x

[a]Percentage of bases in the *S. purpuratus* reference genome covered with at least one read

[b]Percentage of coding bases in the *S. purpuratus* reference genome covered with at least one read

[c]Percentage of single-copy ortholog coding bases covered at 10x depth

[d]Mean genome-wide coverage depth of the *S. purpuratus* reference genome

[e]Mean coverage depth for 246,202 unique exons in the *S. purpuratus* genome assembly

[f]Mean coverage depth of coding bases for 4,497 single-copy orthologs

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*; *Mnud - M. nudus*; *Mfra - M. franciscanus*; *Pdep - P. depressus*.

Table 1.2. Results of ABBA-BABA tests with Dsuite. The tests are organized by P3 taxon. Equal numbers of ABBA and BABA sites are expected under the null hypothesis of no introgression ($D = 0$). A positive D statistic indicates introgression between P3 and P2.

Samples			Dsuite						
P1	**P2**	**P3**	**D**	**z**	**p**	**D_p**	**BBAA**	**ABBA**	**BABA**
Mnud	Mfra	Pdep	0.076	33.8	0.000	0.040	240,218	144,747	124,331
Sfra	Sdro	Spal	0.025	11.8	0.000	0.013	319,896	185,499	176,591
Sfra	Sdro	Sint	0.001	0.3	0.735	0.000	427,693	185,058	184,824
Sdro	Spal	Sint	0.010	5.5	0.000	0.006	249,986	187,513	183,693
Sfra	Spal	Sint	0.012	6.7	0.000	0.007	250,248	194,472	189,743
Sdro	Sfra	Spur	0.059	28.9	0.000	0.026	490,027	200,788	178,420
Sint	Sfra	Spur	0.099	51.5	0.000	0.062	289,884	271,623	222,678
Spal	Sfra	Spur	0.096	47.9	0.000	0.055	292,707	210,001	173,050
Sint	Sdro	Spur	0.052	27.5	0.000	0.032	278,541	239,301	215,697
Spal	Sdro	Spur	0.050	25.7	0.000	0.028	297,221	189,217	171,072
Sint	Spal	Spur	0.008	4.0	0.000	0.005	251,450	194,590	191,590
Spur	Sfra	Hpul	0.013	6.1	0.000	0.005	443,234	162,520	158,463
Spur	Sdro	Hpul	0.020	9.6	0.000	0.009	406,457	159,147	152,805
Spal	Sdro	Hpul	0.006	2.5	0.013	0.002	411,339	115,830	114,528
Sfra	Sdro	Hpul	0.008	3.5	0.000	0.002	608,640	119,046	117,138
Spur	Spal	Hpul	0.017	7.5	0.000	0.007	342,870	139,011	134,494
Sfra	Spal	Hpul	0.003	1.3	0.206	0.001	414,614	118,974	118,304
Spur	Sint	Hpul	0.022	10.5	0.000	0.010	406,767	172,255	164,957
Spal	Sint	Hpul	0.010	4.4	0.000	0.004	370,005	128,140	125,634
Sfra	Sint	Hpul	0.011	5.4	0.000	0.005	436,461	156,898	153,403
Sdro	Sint	Hpul	0.006	2.8	0.006	0.002	417,256	149,052	147,317

Species abbreviations: Sfra - *S. fragilis*; Sdro - *S. droebachiensis*; Spal - *S. pallidus*; Sint - *S. intermedius*; Spur - *S. purpuratus*; Hpul - *H. pulcherrimus*; Mnud - *M. nudus*; Mfra - *M. franciscanus*; Pdep - *P. depressus*.

Table 1.3. Results of Δ analysis

Samples	Δ Analysis						
Quartet	**Trees**[†]	**Concordant**[‡]	**Discordant 1**[§]	**Discordant 2**[¶]	**Δ**	**SE**	**z**
(((Sfra,Sdro),Spal),Mfra)	2,085	974	639	472	0.15	0.03	5.04
(((Sdro,Spal,),Sint),Mfra)	2,107	1,104	550	453	0.10	0.03	3.06
(((Mnud,Mfra),Pdep),Spur)	2,416	1,187	683	546	0.11	0.03	3.94

[†]Total number of gene trees reconstructed from single-copy orthologs

[‡]Number of gene trees that were concordant with the species tree relationships (((P1,P2),P3),O)

[§]Number of gene trees that had the discordant relationship (((P2,P3),P1),O)

[¶]Number of gene trees that had the discordant relationship (((P1,P3),P2),O)

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Mnud - M. nudus*; *Mfra - M. franciscanus*; *Pdep - P. depressus*.

Table 1.4. Summary of the phylogenomic methods supporting different introgression events. *nt* – not tested; *SNVs* – single nucleotide variants.

Taxa	**Analysis**				
	gCF/sCF	**mtDNA**	**Patterson's *D***	**Δ**	**PhyloNet**
	Input Data				
	4,497 Single-Copy Orthologs	Mitochondrial Genome Assemblies	Genome-Wide SNVs	Single-Copy Orthologs†	2,224 Single-Copy Orthologs
Mfra - Pdep	x	x	x	x	nt
Spal - Sdro	x	x	x	x	
Sint - Sdro				nt	
Sint - Spal	x		x	x	x
Spur - Sfra			x	nt	x
Spur - Sdro			x	nt	
Spur - Spal			x	nt	
Hpul - Sfra			x	nt	
Hpul - Sdro			x	nt	
Hpul - Spal			x	nt	
Hpul - Sint			x	nt	
				nt	
Hpul - Sdro/Spal/Sfra/Sint MRCA			x	nt	x
Spur - Sdro/Sfra/Spal MRCA	x	x	x	nt	x

†The number of single-copy orthologs varied depending on the taxa triplet tested. See Table 1.3 for counts.

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*; *Mnud - M. nudus*; *Mfra - M. franciscanus*; *Pdep - P. depressus*.

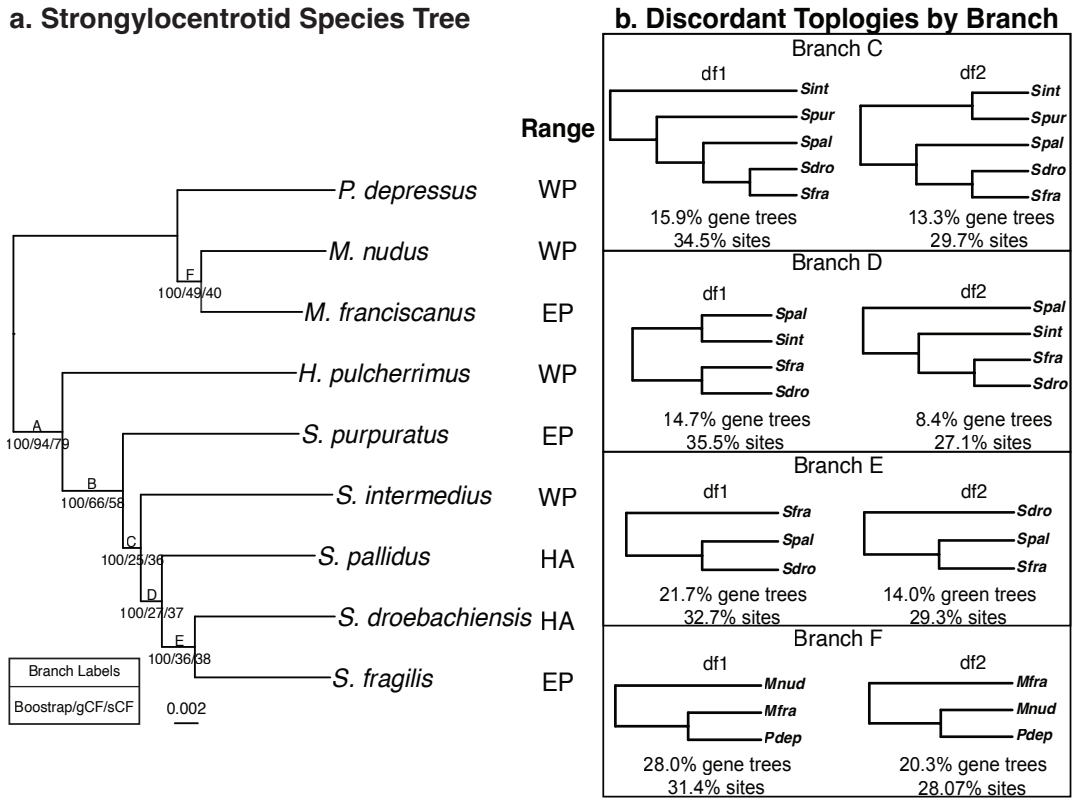


Figure 1.1. a. Phylogeny of the nine strongylocentrotid sea urchin species included in the study. A maximum-likelihood species tree was inferred using the edge-linked partition model of IQ-TREE (Nguyen et al. 2015; Chernomor et al. 2016) on 4,497 concatenated single-copy ortholog alignments. Branch supports were obtained using ultrafast bootstrap (Hoang et al., 2018) with 1,000 replicates. Gene concordance factor (gCF) and site concordance factor (sCF) statistics (Minh et al., 2020; Mo et al., 2022) were calculated using IQ-TREEv2.2.2. For each branch in the species tree, the gCF measures the proportion of gene trees containing that branch, while the sCF measures the proportion of informative sites concordant with that branch (Minh et al., 2020). **b**. Extended output from the gene concordance factor statistics, showing the most frequent discordant topologies (df1, df2) for branches in the species tree with significant imbalances in the frequencies of df1 and df2. The frequencies of the df1 and df2 topologies are expected to be equal under incomplete lineage sorting alone.

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*; *Mnud - M. nudus*; *Mfra - M. franciscanus*; *Pdep - P. depressus*.

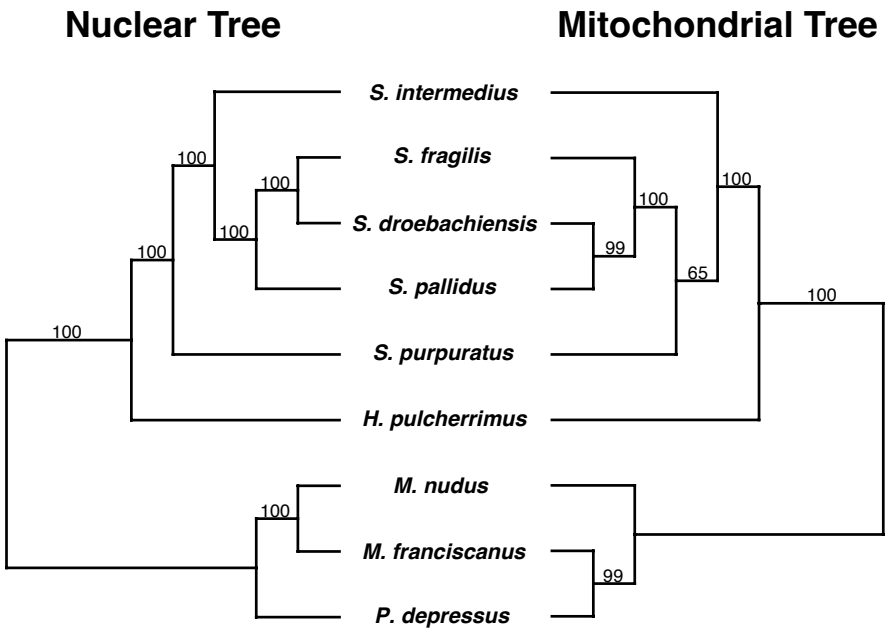


Figure 1.2. A maximum likelihood tree of mitochondrial genome assemblies was inferred from the same samples used in the nuclear species tree shown in Figure 1.1a. Both nuclear and mitochondrial trees were rooted at the midpoint. The mitochondrial genomes were aligned using Clustal Omega v1.2.3, and a maximum likelihood tree was constructed using IQ-TREE (Nguyen et al., 2015) and ModelFinder (Kalyaanamoorthy et al., 2017). Branch supports were obtained using ultrafast bootstrap (Hoang et al., 2018) with 1,000 replicates. Relative to the true species relationships (Figure 1.1a), the placements of the following are swapped: (i) *M. nudus* and *P. depressus*, (ii) *S. purpuratus* and *S. intermedius*, and (iii) *S. pallidus* and *S. fragilis*.

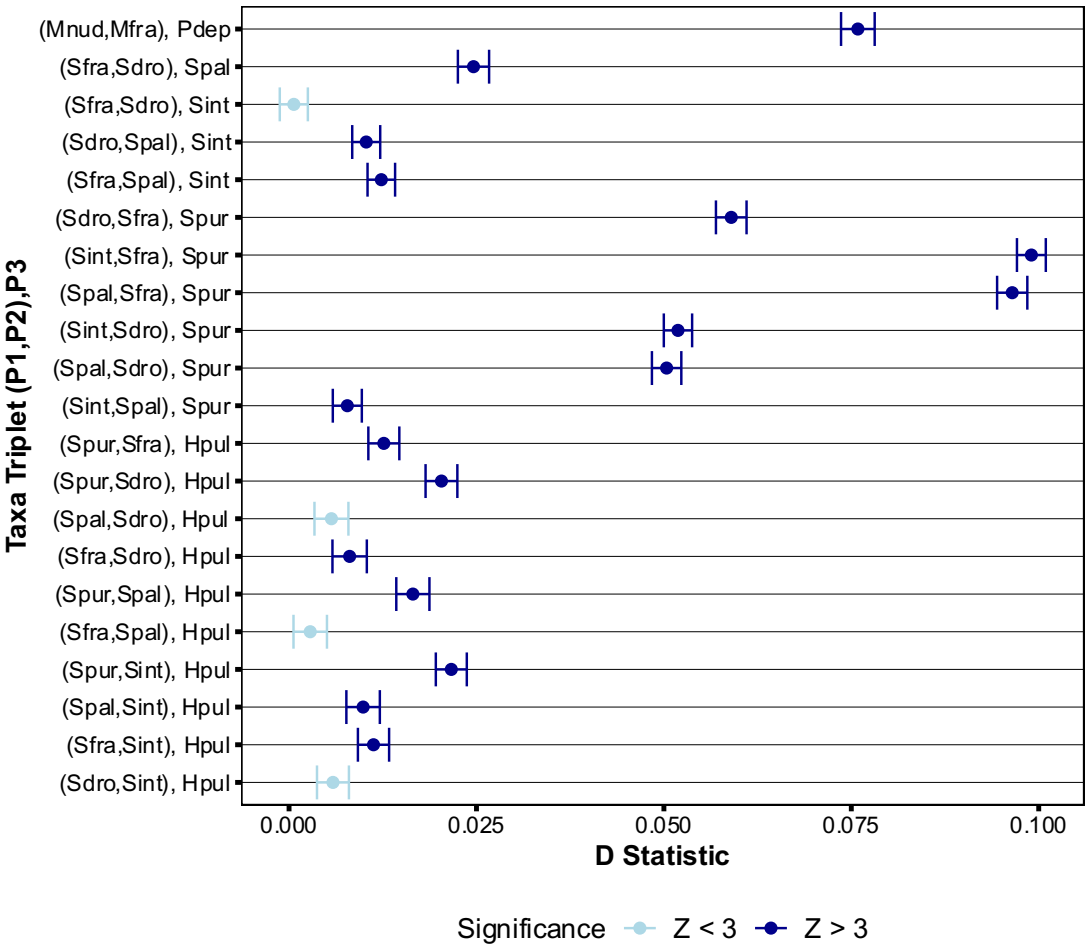


Figure 1.3. Results of ABBA-BABA tests for all phylogenetically relevant triplets. Equal numbers of ABBA and BABA sites are expected under the null hypothesis of no introgression ($D = 0$). A positive D statistic indicates introgression between P3 and P2. Significance was assessed using a block jackknife size of 1Mb. Error bars represent the standard error.

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*; *Mnud - M. nudus*; *Mfra - M. franciscanus*; *Pdep - P. depressus*.

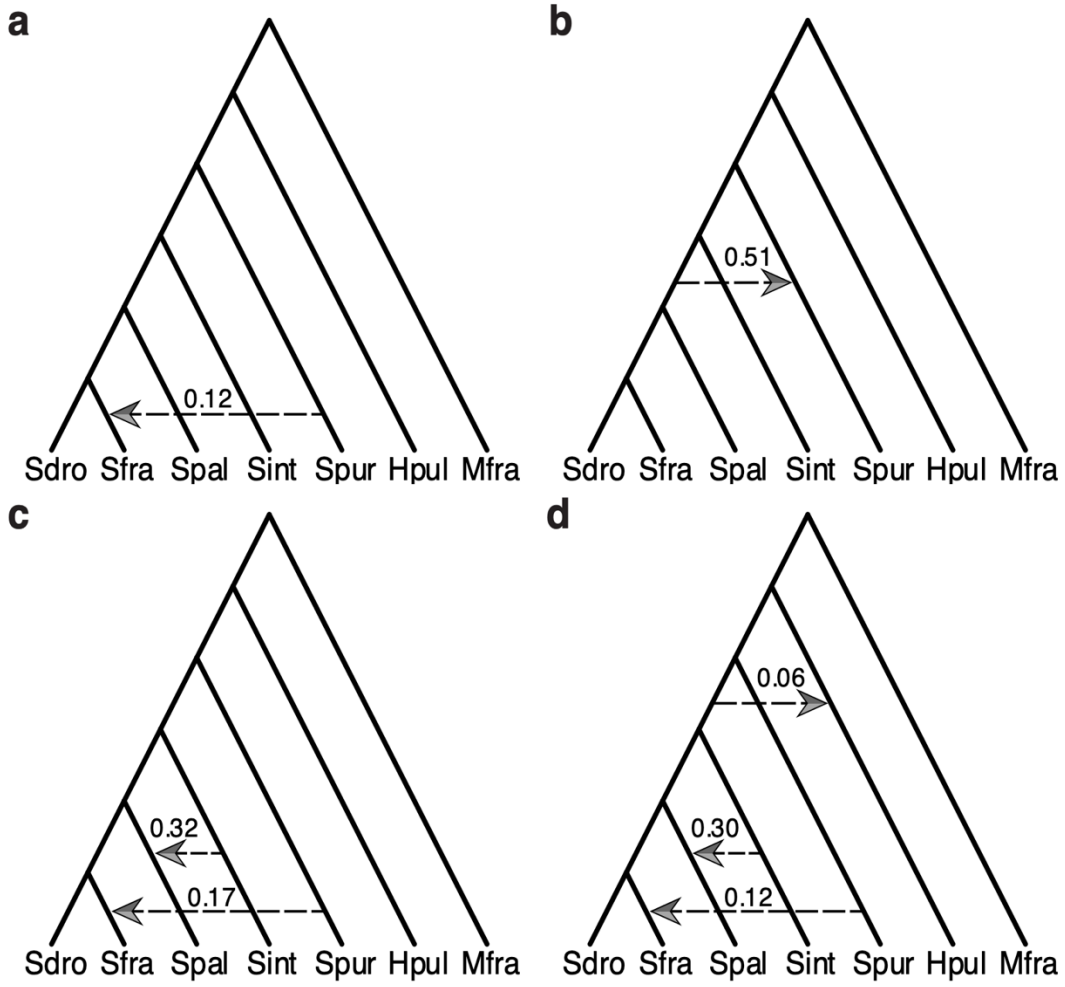


Figure 1.4. Phylogenetic networks with reticulation edges and inheritance probabilities inferred by PhyloNet InferNetwork_ML. The inheritance probabilities represent the proportion of sampled genes inherited through gene flow. The network with zero reticulation edges recovered the species relationships and had a log-likelihood of -11,054 (not shown). a. The best network with one reticulation edge (log-likelihood: -10,966). b. The second-best network with one reticulation edge (log-likelihood: -10,976). c. The network inferred with two reticulation edges (log likelihood: -10,929) d. The network inferred with three reticulation edges (log-likelihood: -10,903).

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*; *Mfra - M. franciscanus*.

Chapter 2 Selection shapes the genomic landscape of introgressed ancestry in a pair of sympatric sea urchin species

Abstract

A growing number of recent studies have demonstrated that introgression is common across the tree of life. However, we still have a limited understanding of the fate and fitness consequence of introgressed variation at the whole-genome scale across diverse taxonomic groups. Here, we implemented a phylogenetic hidden Markov model to identify and characterize introgressed genomic regions in a pair of well-diverged, non-sister sea urchin species: *Strongylocentrotus pallidus* and *S. droebachiensis*. Despite the old age of introgression, a sizable fraction of the genome (1% - 5%) exhibited introgressed ancestry, including numerous genes showing signals of historical positive selection that may represent cases of adaptive introgression. One striking result was the overrepresentation of hyalin genes in the identified introgressed regions despite observing considerable overall evidence of selection against introgression. There was a negative correlation between introgression and chromosome gene density, and two chromosomes were observed with considerably reduced introgression. Relative to the non-introgressed genome-wide background, introgressed regions had significantly reduced nucleotide divergence (d_{XY}) and overlapped fewer protein-coding genes, coding bases, and genes with a history of positive selection.

Additionally, genes residing within introgressed regions showed slower rates of evolution (d_N, d_S, d_N/d_S) than random samples of genes without introgressed ancestry. Overall, our findings are consistent with widespread selection against introgressed ancestry across the genome and suggest that slowly evolving, low-divergence genomic regions are more likely to move between species and avoid negative selection following hybridization and introgression.

2.1 Introduction

Advances in genome sequencing have revealed that many species hybridize with close relatives and share alleles through introgression. However, relatively little is known about the fate and fitness consequence of introgressed variation at the whole-genome scale. Although it is well established that introgression often facilitates adaptation (Song et al. 2011; The Heliconius Genome Consortium 2012; Huerta-Sánchez et al. 2014; Lamichhaney et al. 2015; Arnold et al. 2016), studies documenting adaptive introgression often focus on phenotypes that were known *a priori* to have been involved in adaptation (Martin and Jiggins 2017). Despite enthusiasm about the potential for introgression to introduce new alleles already tested by selection at a frequency higher than mutation (Hedrick 2013; Martin and Jiggins 2017), there remain few estimates of the proportion of introgressed ancestry fixed by positive selection across entire genomes.

A promising approach to understanding the overall fitness consequence of introgression involves identifying and characterizing genomic regions showing introgressed ancestry. Genome-wide scans for introgression have demonstrated that the distribution of introgressed haplotypes across the genome, termed the genomic landscape of introgression, is highly heterogeneous due to the combined effects of natural selection and recombination. Most introgressed variation is thought to be deleterious and removed by selection because it either (i) is maladapted to the recipient's ecology (i.e., ecological selection; McBride and Singer 2010; Arnegard et al. 2014; Cooper et al. 2018), (ii) causes negative epistasis in the genomic background

of the recipient (i.e., hybrid incompatibilities; Orr 1995), or (iii) imposes a genetic load on the recipient if the donor has a smaller effective population size (i.e., hybridization load; Harris and Nielsen 2016; Juric et al. 2016; but see Kim et al. 2018). Selection against introgression should lead to a positive correlation between introgression and recombination because recombination weakens the strength of selection acting on introgressed variation by breaking up long introgression tracts with multiple linked, deleterious alleles and distributing introgressed ancestry more evenly among individuals (Barton 1983; Harris and Nielsen 2016; Veller et al. 2023). Numerous empirical studies have demonstrated a positive correlation between introgression and the local recombination rate (Brandvain et al. 2014; Schumer et al. 2016; Ravinet et al. 2018; Martin et al. 2019; Calfee et al. 2021; Ravinet et al. 2021). However, few organismal groups have been studied, and some notable exceptions exist, including a negative correlation between local recombination rate and admixed ancestry in *Drosophila melanogaster* (Pool 2015; Corbett-Detig and Nielsen 2017).

Gene density is also thought to influence the distribution of introgressed ancestry, as the strength of selection depends on the density of selected sites (Barton 1983; Barton and Bengtsson 1986; Martin and Jiggins 2017). Reduced rates of introgression near functionally important elements have been found in diverse groups, such as humans (Sankararaman et al. 2014; Juric et al. 2016; Sankararaman et al. 2016; Petr et al. 2019), house mice (Teeter et al. 2008; Janoušek et al. 2015), *Histoplasma* fungi (Maxwell et al. 2018), wild strawberries (Feng et al. 2023), and *Xiphophorus* swordtails (Schumer et al. 2016). It is important to note that gene density and

recombination rate are often not independent, further complicating the interpretation of introgression patterns (Martin and Jiggins 2017). For example, a positive correlation between recombination rate and gene density may lead to higher retention of introgression in gene-dense regions than expected (Schumer et al. 2016; Baker et al. 2017; Schumer et al. 2018; Moran et al. 2021).

Within the protein-coding portion of the genome, introgressed variation should be less common at genes with high divergence or faster rates of adaptive evolution because they are more likely to underlie locally adapted phenotypes or hybrid incompatibilities. However, few empirical studies have tested this prediction. Among modern humans, introgressed Neanderthal ancestry is negatively correlated with fixed differences between humans and Neanderthals, consistent with introgressed variation having negative fitness consequences in high-divergence regions (Vernot and Akey 2014). Conversely, Schumer et al. (2016) found that introgressed loci in *Xiphophorus* swordtails had higher divergence than non-introgressed loci due to reduced selective constraint. Characterizing patterns of introgression across a broader range of taxonomic groups is needed to better understand the factors influencing the distribution of introgression along genomes.

The strongylocentrotid family of sea urchins is a compelling group for characterizing the genomic landscape of introgression. Extensive introgression has occurred among the strongylocentrotid urchins (Glasenapp and Pogson 2023), and the purple sea urchin, *Strongylocentrotus purpuratus* (Stimpson), has a well-annotated reference genome and a long history of use as a model organism in fertilization and

development studies. The selective histories of the single-copy protein-coding genes in this family have been formally characterized (Kober and Pogson 2017), providing valuable context for interpreting introgression patterns. The massive effective population sizes of sea urchins should lead to considerably more efficient selection on introgressed variation than most model systems studied thus far. Additionally, their high amounts of recombination (Brennan et al. 2019) and lack of population structure (Palumbi and Wilson 1990; Palumbi and Kessing 1991) should promote the retention of introgressed ancestry. Several studies have documented introgression between a pair of recently-diverged, non-sister taxa that co-occur and hybridize in both the North Pacific and North Atlantic: *S. pallidus* and *S. droebachiensis* (Addison and Hart 2005b; Harper and Hart 2007; Addison and Pogson 2009a; Pujolar and Pogson 2011; Glasenapp and Pogson 2023). However, the genomic regions showing signals of introgression have yet to be characterized.

To better understand the fitness consequence of introgression in strongylocentrotid sea urchins, we characterized genomic regions exhibiting introgressed ancestry between *S. pallidus* and *S. droebachiensis* and asked whether these regions show any nonrandom patterns compared to the genome-wide background. We predicted that introgressed regions would have lower gene density, divergence, and rates of evolution than the genome-wide background and that genes with a history of positive selection would show reduced rates of introgression. We further looked for potential cases of adaptive introgression and tested whether introgressed genes were enriched for any gene families or functional categories.

2.2 Results

To identify genomic regions supporting introgression between non-sister taxa *S. pallidus* and *S. droebachiensis*, we applied the phylogenetic hidden Markov model PhyloNet-HMM (Liu et al. 2014) to pseudo-haploid multi-species multiple sequence alignments of the 21 largest scaffolds in the *S. purpuratus* reference genome assembly (Figure 2.1). The multiple sequence alignments were constructed from hard-filtered genotypes of each species in the rooted triplet (*Hemicentrotus pulcherrimus*, (*Strongylocentrotus pallidus*, (*S. droebachiensis*, *S. fragilis*))). The genotypes were obtained by mapping paired-end sequencing reads of each species to the *S. purpuratus* reference genome with bwa-mem2 (Vasimuddin et al. 2019) and calling and genotyping variants following GATK's Best Practices (Van der Auwera et al. 2013). The 21 largest scaffolds in the Spur_5.0 assembly correspond to the 21 *S. purpuratus* chromosomes (2n=42) and represent 90% of the bases in the 922 Mb assembly. The PhyloNet-HMM model walks along each chromosome, identifies changes in the underlying genealogy, and outputs posterior probabilities of having evolved by each parent tree (i.e., species tree, introgression tree) for each site in the multiple sequence alignment (Liu et al. 2014). The model accounts for both convergence and incomplete lineage sorting (ILS) by employing a finite-sites model and allowing for changes in gene trees within each parent tree (Liu et al. 2014; Liu et al. 2015; Schumer et al. 2016). We ran PhyloNet-HMM 100 times on each scaffold and averaged the posterior probabilities across runs to avoid the effects of reaching local optima during hill climbing (per suggestion by PhyloNet-HMM developer Qiqige Wuyun). To infer

introgression tracts, we applied a posterior probability threshold for introgression of 90% and recorded the genomic coordinates of consecutive sites with posterior probabilities at or above this threshold. We also proceeded with a less stringent dataset at the 80% posterior probability threshold for comparison, given the conservative nature of our test (see Discussion) and the small size of the dataset identified at the 90% threshold. In both datasets, the inferred introgression tracts were filtered if they had a mean coverage depth less than 5x or greater than 100x and trimmed if they overlapped a gap of more than 25kb between adjacent genotypes (including invariant sites).

At the 90% posterior probability threshold, we identified 4,855 introgression tracts (≥2 bases), with 164 exceeding 10kb. Tracts greater than 10kb in length had mean and median lengths of 22,850 and 16,595 base pairs, respectively (Supplementary Figure S1). The coverage depth and breadth metrics for introgression tracts were similar to the genome-wide averages (Table 2.1, Supplementary Table S12). The percent of bases introgressed, both overall and in coding regions, was 1%.

When the posterior probability threshold was lowered to 80%, the total number of tracts increased to 17,037, with 953 exceeding 10kb. The mean and median tract lengths for 10kb tracts (22,897, 17,236 base pairs) remained similar to those at the 90% threshold. The percentage of bases introgressed rose to 5% overall and 6% in coding regions. These estimates of the proportion of the genome introgressed (1 - 5 %) align with previous estimates for *S. pallidus* and *S. droebachiensis* (Glasenapp and Pogson, 2023). Summary statistics for the introgression tracts at both probability thresholds are provided in Supplementary Table S13, and information on all introgression tracts can

be found in Supplementary Tables S14 and S15. Additionally, Figure 2.2 and Supplementary Figure S2 depict the locations of the 10 kb introgression tracts along scaffolds.

There was a significantly negative association between the percent scaffold introgressed and scaffold-wide gene density (Supplementary Figure S3). The scaffold with the highest percent introgressed and most 10kb introgression tracts (NW_022145606.1) also had the lowest gene density (Supplementary Table S16). Unexpectedly, two scaffolds (NW_022145615.1, NW_022145595.1) did not have any sites that crossed the 90% posterior probability threshold for introgression (Supplementary Table S16). There were no discernible features of these two chromosomes that would lead to reduced power to detect introgression. Both had high site density (Supplementary Table S16) and although NW_022145595.1 is the shortest scaffold at 30 Mb, fifteen of the 21 scaffolds were between 30 and 40 Mb in length. To determine the probability of having two chromosomes without 10 kb introgression tracts due to chance, we divided the genome into non-overlapping 10kb blocks and selected 164 blocks at random 10,000 times, recording the frequency of one or more chromosomes not being represented. The number of times a single chromosome was not represented was 154 in 10,000 (0.015). The number of times two chromosomes were not represented was 1 in 10,000 (0.0001). At the 80% posterior probability threshold, all chromosomes had 10kb introgression tracts, ranging from 9-127 (Supplementary Table S17). Consistent with the results at the 90% threshold level, scaffolds NW_022145615.1 and NW_022145595.1 had the fewest percent of

introgressed sites (1.4%, 1.5%) and 10kb introgression tracts (9, 10), while scaffold NW_022145606.1 again had the highest proportion of introgressed sites (11.5%) and the most 10kb introgression tracts (127) (Supplementary Table S17).

To characterize the properties of introgressed genomic regions, we compared estimates of absolute nucleotide divergence (d_{XY}), gene density, coding base density, the rate of nonsynonymous substitutions (d_N), the rate of synonymous substitutions (d_S), and the nonsynonymous to synonymous substitution rate ratio (d_N/d_S) for introgressed regions and genes to the non-introgressed genome-wide background. To avoid the confounding effect of introgression between *S. pallidus* and *S. droebachiensis* on d_{XY} estimates, we used *H. pulcherrimus* and *S. fragilis*, who have experienced little-to-no introgression. We implemented a bootstrap comparison of means to compare d_{XY} between introgression tracts and the non-introgressed genome-wide background. First, we calculated *H. pulcherrimus* - *S. fragilis* d_{XY} for all introgression tracts, and a random sample of regions of the same number and size confidently called for the species tree. We then pooled all d_{XY} values, bootstrap resampled the pool in pairs 100,000 times, and calculated the difference in mean d_{XY} between bootstrapped pairs to generate the distribution of differences in means expected if there were no difference between mean introgressed d_{XY} and mean non-introgressed d_{XY}. We then compared the true difference in mean d_{XY} between the introgressed and non-introgressed regions to the null distribution to calculate a p-value. We found that introgressed regions had lower divergence (d_{XY}) than the genome-wide background at both posterior probability thresholds ($p<0.0001$; Table 2.2, Figure 2.4, Supplementary Figure S4).

To compare the rate of evolution (d_N, d_S, and d_N/d_S) between introgressed and non-introgressed regions, the same procedure used in the d_{XY} analysis was repeated for genes with more than half of their bases declared introgressed and a random set of the same number of genes with more than half of their bases confidently called for the species tree. Genes were filtered if they had a mean coverage depth less than 10x or greater than 100x, had fewer than 75% of their coding bases covered by one read or fewer than 50% by ten reads, or contained stop codons. We estimated d_N, d_S, and d_N/d_S using codeml M0 of PAML (Yang 2007). At the 90% posterior probability, introgressed genes had lower d_N ($p=0.03$), d_S ($p=0.34$), and d_N/d_S ($p<0.0001$) than non-introgressed genes (Table 2.2, Figure 2.4, Supplementary Table S18). The relationships remained the same at the 80% posterior probability threshold, with the difference in mean d_S becoming significant ($p=0.001$; Supplementary Table S18, Supplementary Figure S5).

We then compared the number of overlapping protein-coding genes, number of overlapping coding bases, and number of overlapping genes with a history of positive selection between the introgression tracts and the non-introgressed genome-wide background. Following the approach of Schumer (2016), we generated distributions of the counts of overlapping genes, coding bases, and positively selected genes for introgression tracts by bootstrap resampling the 10kb introgression tracts with replacement 1,000 times and counting the number of overlapping features. We calculated the mean and standard deviation for each metric. We then compared these means to null distributions created by randomly permuting intervals of the same

number and size as the introgression tracts into the genomic regions confidently called for the species tree 1,000 times and counting the number of overlapping genes, coding bases, and positively selected genes. Introgression tracts overlapped fewer protein-coding genes, coding bases, and genes with a history of positive selection at both posterior probability thresholds (Table 2.2, Figure 2.4, Supplementary Table S18, Supplementary Figure S5).

We further identified all genes overlapping introgression tracts at both posterior probability thresholds (Supplementary Table S19, S20) using gene models from the latest *S. purpuratus* genome assembly (Spur_5.0). At the 90% posterior probability threshold, 50 protein-coding genes had all their bases declared introgressed, and another 102 had more than half of their bases introgressed. A total of 2,055 genes overlapped an introgression tract by at least two bases. One noteworthy pattern was that many different hyalin genes had bases declared introgressed. Hyalin, an extracellular matrix glycoprotein, is the major component of the hyaline layer, an extraembryonic matrix serving as a cell adhesion substrate during development (McClay and Fink 1982; Wessel et al. 1998). At the 90% threshold, five unique hyalin genes had bases introgressed (LOC578156, LOC752152, LOC373362, LOC100891695, LOC100891850), and an additional four hyalin genes showed introgression at the 80% threshold (LOC578713, LOC586885, LOC576524, LOC578967). Coverage depth for all but one of these genes (LOC578967) was in the range expected if they were single-copy across the four species analyzed (Supplementary Table S21). The high number of hyalin genes observed may not be

unexpected, given that there are 21 hyalin genes in the *S. purpuratus* assembly. To test whether there were more occurrences of hyalin than expected by chance, we randomly sampled the same number of genes as the number of introgressed genes from the set of all genes on the 21 chromosomes 1,000 times and recorded the number of hyalin occurrences. There were 3.3x more hyalin genes in the introgressed set than expected due to chance at the 90% threshold (95% confidence interval: 1.4 - 1.6) and 1.8x at the 80% threshold (95% confidence interval: 4.9 - 5.2). More occurrences of hyalin gene introgression may be expected than due to random chance if they are clustered near each other on chromosomes and/or do not segregate independently. However, the introgressed hyalin genes have nonoverlapping coordinates and occur across six different chromosomes, with the shortest gap between genes sharing a chromosome being > 400kb. Furthermore, in no case did a single introgression tract overlap more than one hyalin gene.

To identify other potential examples of adaptive introgression, we looked for overlap between genes with introgressed coding bases and the 1,008 strongylocentrotid single-copy orthologs with a history of positive selection previously identified by Kober and Pogson (2017). The positively selected genes (PSGs) had been previously identified by comparing the codon sites models M7 (Beta) *vs*. M8 (Beta plus ω) (Yang et al. 2000) using the CODEML program of the PAML Package (Yang, 2007). Branch-sites tests of positive selection were also used to identify lineage-specific episodes of adaptive evolution (Kober and Pogson 2017). At the 90% confidence level, three genes with significant sites tests across the family had more than half of their coding bases

declared introgressed: arachidonate 5-lipoxygenase (Supplementary Figure S6), helicase domino, and kinesin-II 95 kDa subunit (Table 2.3, Supplementary Table S22). There were 32 PSGs (3.2%) with at least 10% of their coding bases introgressed. At the 80% posterior probability threshold, there was a total of 24 PSGs (2.4%) with more than half of their coding bases declared introgressed (Supplementary Table S23), including six genes that had all their coding bases declared introgressed: arachidonate 5-lipoxygenase, 5-hydroxytryptamine receptor 6, transcription termination factor 1, MAK16 homolog, glutathione peroxidase-like, and 2',3'-cyclic-nucleotide 3'-phosphodiesterase-like (Table 2.3). The maximum likelihood gene trees for all introgressed PSGs shown in Table 2.3 grouped *S. pallidus* and *S. droebachiensis* as sister taxa, except for glutathione peroxidase-like, which did not have enough high-quality variant sites for gene tree reconstruction. All candidate introgressed PSGs had high coverage depth in the range expected for single-copy orthologs, indicating that genotyping error likely did not contribute to the introgression signal (Table 2.3).

We also looked for overlap between introgressed genes and genes with significant branch-sites tests on the *S. pallidus* and *S. droebachiensis* terminal branches, indicating lineage-specific episodes of adaptive protein evolution (Yang 2005; Zhang et al. 2005). Four genes with significant branch-sites tests on the *S. pallidus* terminal branch had an appreciable proportion of their coding bases introgressed at the 90% posterior probability threshold: kremen protein 1, arylsulfatase, sodium- and chloride-dependent neutral and basic amino acid transporter B(0+), and fibrosurfin-like (Table 2.4). Two genes with significant branch-sites test on the *S. droebachiensis* terminal

branch had coding bases introgressed at the 90% threshold, PHD finger protein 8, and structural maintenance of chromosomes 1A (Table 2.4). All introgressed genes with significant branch-sites tests in Table 2.4 supported an *S. pallidus* and *S. droebachiensis* sister relationship with an average bootstrap support of 85%. Additional genes with significant branch-sites tests and introgressed bases are available in Supplementary Tables S24-S27.

To determine whether any additional classes of genes were over- or underrepresented in the introgressed set, we tested the genes with more than half of their bases introgressed at the 80% posterior probability threshold for gene ontology enrichment using PANTHER18.0 (Mi et al. 2019; Thomas et al. 2022). Only two significant terms remained after applying a false discovery rate (FDR) correction of 5%. There was an under-enrichment of the cellular component terms "plasma membrane" (GO:0005886, $p = 0.014$) and "cell periphery" (GO:0071944, $p = 0.002$). Interestingly, the cellular component term "membrane" (GO:0016020) was enriched in the set of genes with histories of positive selection from Kober and Pogson (2017). Furthermore, the molecular function term "calcium ion binding" (GO:0005509) and biological process term "proteolysis" (GO:0006508) were overrepresented in the set of positively selected genes (Kober and Pogson 2017) and underrepresented in the collection of introgressed genes, though not significant after correction.

2.3 Discussion

Here, we characterized the genomic landscape of introgression between two sea urchin species to gain insight into the factors determining the fate of introgressed variation and the behavior of selection following introgression. Our study is among the first to perform local ancestry inference with whole-genome sequencing data in a high gene flow marine invertebrate group. The strongylocentrotid sea urchin family stands out relative to other population genetic models for their massive effective population sizes, highly efficient selection, and high gene flow across ocean basins. Although the species are well-diverged (4.2 - 19.0 mya), and natural hybrids are rare, many of the species show strong signals of historical introgression (Glasenapp and Pogson 2023). In our analysis of introgression between *S. pallidus* and *S. droebachiensis*, we found strong evidence for genome-wide selection against introgression, including two chromosomes depleted of introgression warranting further examination. Although our results suggest that slowly evolving loci with low divergence are more likely to be able to move between species, introgression has also likely been an important source of adaptive genetic variation. Between 1% and 6% of coding bases supported introgression, and numerous genes with histories of positive selection also had a significant number of introgressed coding bases. A handful of the introgressed genes with histories of selection are involved in defense, including arachidonate 5-lipoxygenase, glutathione peroxidase, and toll-like receptor 3. Additionally, the introgression of many hyalin genes distributed across multiple chromosomes suggests potential functional and adaptive significance, possibly related to defense. Hyalin is a

large glycoprotein and a major component of the hyaline layer, the egg extracellular matrix that serves as a cell adhesion substrate during gastrulation (McClay and Fink 1982; Adelson and Humphreys 1988; Wessel et al. 1998). Kober and Pogson (2017) have suggested that the prevalence of positive selection at membrane or extracellular proteins (such as collagens) might be driven by pathogen defense.

Consistent with theoretical predictions about the retention of introgressed ancestry, we found introgression to be more common in genomic regions expected to be under weaker selection. These regions exhibited lower gene density, reduced divergence, slower rates of evolution, and fewer positively selected genes than the non-introgressed genome-wide background. We find it unlikely that these patterns were driven by increased power to detect introgression in low divergence regions, as PhyloNet–HMM requires sequence divergence to detect introgression (Schumer et al. 2016). Furthermore, a negative correlation between introgressed ancestry and sequence divergence has been observed in humans (Vernot and Akey 2014), and many studies have found depleted introgression in functional regions (Teeter et al. 2008; Sankararaman et al. 2014; Janoušek et al. 2015; Juric et al. 2016; Sankararaman et al. 2016; Schumer et al. 2016; Maxwell et al. 2018; Petr et al. 2019). Reduced introgression in regions with high divergence or functional density is likely explained by divergent regions harboring loci underlying local adaptation or Dobzhansky-Muller incompatibilities (Moran et al. 2021). Unfortunately, limited information about natural hybrids and ecological selection in the strongylocentrotid family precludes distinguishing between the different sources of selection against introgressed variation.

Our findings are at odds with those of Schumer et al. (2016), who characterized introgressed *Xiphophorus cortezi* ancestry in *X. nezahualcoyotl* genomes and found that introgressed regions had higher sequence divergence, gene density, and rates of synonymous and nonsynonymous substitutions than the genome-wide background. They demonstrated that the higher divergence of introgressed regions was likely driven by introgression at genes not under strong selective constraint, which is still consistent with genome-wide selection against introgression. The unique results in the *Xiphophorus* system may be driven by the fact that recombination hotspots are concentrated near promoter-like features in swordtails and occur further from transcription start sites in humans and other species with PRDM9-direction recombination (Myers et al. 2005; Coop et al. 2008; Baker et al. 2017; Moran et al. 2021). Furthermore, the *X. nezahualcoyotl* swordtail genome samples had low genetic diversity ($\theta\pi$: 0.00025 - 0.00082), indicating that low or fluctuating effective population sizes and less efficient selection could have also contributed to the higher-than-expected amount of introgression in gene dense regions. The differences between our findings and those of Schumer et al. (2016) highlight the importance of characterizing admixture and introgression across more taxonomic groups.

We believe our estimates of the proportion of the genome introgressed were conservative for several reasons. First, introgression between *S. pallidus* and *S. droebachiensis* is likely historical, making it harder to detect because recombination breaks introgressed haplotypes into progressively shorter tracts over time, and new mutations obscure the history of introgression. Most of the strongylocentrotid genes

that showed introgression were not fully introgressed, especially those with histories of positive selection. Instead, many genes had one or more small regions with very strong support for introgression. Due to the old age of introgression and the high expected amount of recombination in strongylocentrotid urchins, the scale of introgression is likely at the exon level rather than the whole gene level. Detecting introgressed regions at this small scale is a major challenge for any statistical method, given the limited number of variants in an individual exon. Second, randomly resolving heterozygous sites to create multiple sequence alignments causes switching between maternal and paternal chromosomes, fragmenting introgressed haplotypes that are heterozygous in our samples and biasing our detection towards introgressed variation that has been fixed. When an introgression tract is heterozygous for ancestry, switching between introgressed and non-introgressed ancestry may lead to ambiguous posterior probabilities (Schumer et al. 2016). For this reason, the introgression tracts we detected are likely fixed and old.

Theory predicts a positive correlation between the extent of introgression and the local recombination rate (Veller et al. 2023). Unfortunately, information on recombination in the strongylocentrotid sea urchins is extremely limited, preventing us from testing this relationship. An outstanding question remains whether differences in recombination rates drove the differences in introgression among the different strongylocentrotid chromosomes. If the number of crossovers per meiosis is constant among chromosomes, shorter chromosomes should have higher per-base recombination rates and retain more introgressed variation. However, we did not find

a significant relationship between introgression and chromosome length, and the smallest chromosome had the least amount of introgression, which is inconsistent with expectations.

Without polymorphism data for *S. pallidus* and *S. droebachiensis*, we can only speculate about the proportion of introgression tracts driven to fixation by positive selection. However, selection is expected to be very efficient in these sea urchin species. For example, it was conservatively estimated that 15% of strongylocentrotid single-copy orthologs had experienced positive selection (Kober and Pogson 2017) and *S. purpuratus* shows selection on preferred usage of synonymous codons (Kober and Pogson 2013). Furthermore, the introgression tracts documented in this study are likely historical and fixed. *S. pallidus* and *S. droebachiensis* diverged 5.3 - 7.6 million years ago (Kober and Bernardi 2013b) and natural hybrids between the two are rarely observed (Vasseur 1952). Given the likely old age of introgression, the high expected efficiency of selection, and our bias toward detecting high-frequency variants, it is not unreasonable to assume that a small proportion of the introgression tracts spanning coding regions contained advantageous mutations. Future studies will test for recent selection at the genomic intervals inferred to have been introgressed and look for adaptive introgression in promoter regions upstream of genes.

In summary, our study documented strong evidence for genome-wide selection against introgressed variation, suggesting that slowly evolving, low-divergence genomic regions are more likely to move between species and avoid negative selection following hybridization and introgression. However, despite strong selection against

introgression, we also identified numerous candidate adaptively introgressed genes, suggesting that introgression has been an important source of adaptive genetic variation. The strongylocentrotid sea urchin family represents a valuable model system for further characterization of introgression given the high amount of gene flow and genetic diversity among the different species.

2.4 Materials and Methods

2.4.1 Study System

Four species forming a rooted triplet were used in the present study: (*Hemicentrotus pulcherrimus*, (*Strongylocentrotus pallidus*, (*S. droebachiensis*, and *S. fragilis*))). The metadata for the sample accessions are presented in Supplementary Table S11. The three *Strongylocentrotus* taxa were sampled from the East Pacific: *S. droebachiensis* and *S. pallidus* were dredged from Friday Harbor, WA, and *S. fragilis* was collected in Monterey Bay. *H. pulcherrimus* was chosen as the outgroup as it was sampled from the West Pacific (coastal Japan by Y. Agatsuma) and diverged from the *Strongylocentrotus* taxa 10 – 14 mya (Kober and Bernardi 2013b). *S. pallidus* and *S. droebachiensis* have broad, overlapping Holarctic distributions with ample opportunity for hybridization. They co-occur in the West Pacific, East Pacific, Arctic, West Atlantic, and East Atlantic Oceans. The geographic history of speciation is challenging to interpret, but fossil evidence confirms that both species speciated in the Pacific and crossed the Bering Sea in the late Miocene to colonize the Arctic and Atlantic Oceans (Durham and MacNeil 1967). Both species show little differentiation between the Pacific and Atlantic due to high trans-Arctic gene flow (Palumbi and Kessing 1991; Addison and Hart 2005b; Addison and Kim 2022).

The eggs of *S. droebachiensis* are highly susceptible to fertilization by heterospecific sperm (Strathmann 1981; Levitan 2002a), and hybrid matings readily occur when spawning *S. droebachiensis* females are closer to heterospecific males than conspecific males (Levitan 2002a). Hybrids between *S. pallidus* and *S. droebachiensis*

have also been successfully reared and backcrossed in the lab (Strathmann 1981). Although reproductive isolation between *S. pallidus* and *S. droebachiensis* appears incomplete, the two species remain distinct across their overlapping ranges (Vasseur 1952; Strathmann 1981). However, hybrids of *S. pallidus* and *S. droebachiensis* morphologically resemble *S. pallidus* as larvae and *S. droebachiensis* as adults, so the frequency of natural hybrids may be underestimated. Introgression between *S. pallidus* and *S. droebachiensis* has been previously detected (Addison and Hart 2005b; Harper et al. 2007b; Addison and Pogson 2009a; Pujolar and Pogson 2011; Glasenapp and Pogson 2023).

2.4.2 Data Pre-Processing

A single genome from each strongylocentrotid species had been previously sequenced on the Illumina HiSeq 2500 (Kober and Bernardi 2013b; Kober and Pogson 2017). The raw sequencing reads were pre-processed following GATK's Best Practices (Van der Auwera et al. 2013). Briefly, adapters were marked with Picard MarkIlluminaAdapters, and sequencing reads were mapped to the *S. purpuratus* reference genome (Spur_5.0) with bwa-mem2 v2.2.1 (Vasimuddin et al. 2019). Duplicates were marked with Picard MarkDuplicates, and reference mapping was evaluated using samtools flagstat (Danecek et al. 2021) and mosdepth v0.3.3 (Pedersen and Quinlan 2018). Variant calling and joint genotyping were performed with GAKT's HaplotypeCaller and GenotypeGVCFs. Variants were hard-filtered for skewed values across all samples following GATK recommendations (Caetano-Anolles 2023).

Further filtering was done for genotypes with low-quality scores (GQ<20) and low read depth (DP<3), and single nucleotide variants (SNVs) within three base pairs of an indel were excluded.

2.4.3 PhyloNet-HMM

We used an updated version of PhyloNet-HMM called PHiMM (Wuyun et al. 2019) to identify genomic regions supporting introgression between *S. pallidus* and *S. droebachiensis*. PhyloNet-HMM is a hidden Markov model that detects breakpoints between regions supporting different phylogenetic relationships, accounting for incomplete lineage sorting (ILS) by allowing for changes in gene trees within both the species and introgression trees (Liu et al. 2014; Schumer et al. 2016). PhyloNet-HMM walks across each chromosome, locates changes in the underlying genealogy, and outputs posterior probabilities for each SNV site, reflecting the likelihood that the site evolved along the species and introgression trees. PhyloNet-HMM has been used to detect introgressed regions in swordtails (Schumer et al. 2016; Powell et al. 2020), house mouse (Liu et al. 2015), North American admiral butterfly (Mullen et al. 2020), *Danaus* butterfly (Aardema and Andolfatto 2016), and snowshoe hare (Jones et al. 2020) genomes. Schumer et al. (2016) conducted performance tests of PhyloNet-HMM on simulated swordtail data and concluded that the approach could accurately distinguish between ILS sorting and hybridization.

Multiple sequence alignments of single nucleotide variant (SNV) sites were created for the 21 largest *S. purpuratus* scaffolds using vcf2phylip (Ortiz 2019). The

21 largest scaffolds correspond to the 21 *S. purpuratus* chromosomes (2n=42) and represent 90% of the 921,855,793 bases in the *S. purpuratus* reference genome assembly (Spur_5.0). Although it is known that *S. purpuratus* has a genetically based sex determination, sex chromosomes have yet to be identified (Pieplow et al. 2023). PhyloNet-HMM only allows for DNA base characters in the input sequence alignments, so all indels were excluded, and only variable sites where all four samples had genotypes passing filter were used. Heterozygous genotypes were randomly resolved because PhyloNet-HMM does not support IUPAC ambiguity codes, and accurate phasing could not be performed across entire chromosomes with a single diploid genome per species and no reference panel (Bukowicki et al. 2016). We ran PhyloNet-HMM on each chromosome 100 times using the default settings to avoid the effects of reaching local optima during hill climbing and averaged the posterior probabilities across independent runs. The average distance between variable sites in the alignments was 85 base pairs. The multiple sequence alignments including invariant sites had 2,361 gaps greater than 25 kb in length, with the largest gap being 585,224 base pairs.

We used two different posterior probability thresholds to identify introgression tracts: 90% and 80%. Introgression tracts were inferred by recording the genomic coordinates of consecutive sites with posterior probabilities at or above the threshold. The mean coverage depth for each introgression tract for each species was calculated with mosdepth (Pedersen and Quinlan 2018), and introgression tracts where any species had coverage depth less than 5x or greater than 100x were excluded.

Introgression tracts overlapping a gap of 25 kb or more between adjacent genotypes (including invariant sites) were identified using bedtools intersect (Quinlan and Hall 2010) and trimmed to remove the gap. Coverage depth metrics were estimated for all introgression tracts passing filter using mosdepth (Pedersen and Quinlan 2018). For each introgression tract, overlapping genes and coding bases were recorded using the gff file from the *S. purpuratus* assembly and bedtools intersect (Quinlan and Hall 2010). We also intersected the introgression tracts with the set of genes with a history of positive selection within the strongylocentrotid family identified by (Kober and Pogson 2017).

2.4.4 Properties of Introgressed Regions

To characterize the genomic landscape of introgression, we compared estimates of absolute nucleotide divergence (d_{XY}), gene density, coding base density, and rates of evolution (d_N, d_S, and d_N/d_S) for the set of introgressed intervals greater than 10 kb in length to estimates for the non-introgressed genome-wide background (i.e., species tree regions).

2.4.4.1 Divergence

We compared the mean absolute nucleotide divergence (d_{XY}) of introgression tracts to the mean d_{XY} of a random sample of non-introgressed genomic regions of the same number and size as the introgressed intervals. To avoid the confounding effect of introgression between *S. pallidus* and *S. droebachiensis* on d_{XY} estimates, we used *H.*

pulcherrimus and *S. fragilis*, who have experienced little-to-no introgression. The two species involved in introgression (*S. pallidus*, *S. droebachiensis*) were not included in the divergence measures because introgression from *S. droebachiensis* into *S. pallidus* would decrease *S. fragilis* - *S. pallidus* divergence, introgression from *S. pallidus* into *S. droebachiensis* would decrease *S. fragilis* - *S. droebachiensis* divergence, and introgression between *S. pallidus* and *S. droebachiensis* in either direction would reduce *S. pallidus* - *S. droebachiensis* divergence (see Forsythe et al. 2020).

To obtain distributions of d_{XY}, we first generated a new genotype (vcf) file for *S. fragilis* and *H. pulcherrimus,* including invariant sites. Vcf files typically only contain variant sites, and d_{XY} estimates can be downwardly biased by assuming missing sites are invariant (Korunes and Samuk 2021). We generated the new genotype file by combining the single sample vcf files for *H. pulcherrimus* and *S. fragilis* and performing joint genotyping using GATK's GenotypeGVCFs with the –include-non-variant-sites option. Variant and invariant sites were then split into separate files for filtering. Variant sites were hard-filtered for skewed values across all samples following GATK recommendations (Caetano-Anolles 2023). The variant and invariant site vcf files were then merged back together, and genotypes with low-quality scores (GQ<30), low read depth (DP<8), or low reference genotype confidence (RGQ<30) were set to missing.

The random sample of non-introgressed regions was created by randomly permuting intervals of the same number and size of the 10kb introgression tracts into regions confidently called for the species tree using bedtools shuffle (Quinlan and Hall

2010). We used pixy (Korunes and Samuk 2021) to calculate d_{XY} for each genomic interval in the sets of introgressed and non-introgressed intervals and implemented a bootstrap comparison of means to test for a significant difference between the mean d_{XY} of introgressed and non-introgressed regions. We pooled all d_{XY} values for both sets, bootstrap resampled the pool in pairs 100,000 times, and calculated the difference in mean d_{XY} between bootstrapped pairs to generate the distribution of differences in means expected if there were no difference between mean introgressed d_{XY} and mean non-introgressed d_{XY}. We then compared the true difference in mean d_{XY} between the introgressed and non-introgressed regions to the null distribution to calculate a p-value.

2.4.4.2 Rate of Evolution

We next compared the rate of evolution of introgressed genes to that of non-introgressed genes. For *H. pulcherrimus* and *S. fragilis*, we first identified genes with more than half of their bases declared introgressed at each posterior probability threshold, excluding those with mean coverage depth less than 10x or greater than 100x, with fewer than 75% of their coding bases covered by one read or fewer than 50% by ten reads, or with premature stop codons. We next created sequence alignments of *H. pulcherrimus* and *S. fragilis* and for each introgressed gene passing filter using vcf2fasta, and estimated d_N, d_S, and d_N/d_S using codeml model M0 of PAML (Yang 2007). We specified the cleandata=1 option to remove sites with ambiguity data. To obtain estimates of d_N, d_S, and d_N/d_S for non-introgressed genes, we identified all genes with more than half of their bases confidently called for the species tree, filtering by

the same metrics as the introgressed genes. We then randomly sampled the identified non-introgressed genes to get a sample the same size as the number of introgressed genes and estimated d_N, d_S, and d_N/d_S for each gene. We compared the means of each metric between introgressed and non-introgressed genes using the same bootstrap comparison of means procedure used in the *d*XY analysis. For each metric (d_N, d_S, and d_N/d_S), all values from both the introgressed and non-introgressed sets were pooled. The pool was bootstrap resampled in pairs 100,000 times, and the difference in means for each metric was calculated between bootstrapped pairs to generate the distribution of differences in means expected if there were no difference between mean introgressed d_N, d_S, and d_N/d_S and mean non-introgressed d_N, d_S, and d_N/d_S. We then compared the true difference in means between the introgressed and non-introgressed regions to the null distribution to calculate a p-value.

2.4.4.3 Gene Density

To determine whether introgressed regions were more or less likely to overlap protein-coding genes, genes with a history of positive selection, and coding bases than the genome-wide background, we bootstrap resampled the introgression tracts with replacement to create 1,000 pseudoreplicate datasets. For each, we counted the number of overlapping coding bases, the number of protein-coding genes with more than half of their bases declared introgressed, and the number of genes with a history of positive selection identified by Kober and Pogson (2017). We identified the overlapping genes and coding bases by intersecting the introgression tract interval files with the protein-coding gene and CDS coordinates for *S. purpuratus* using bedtools intersect. To

standardize the protein-coding gene counts, we divided the values by the total length of the interval files in megabases. To normalize the coding base counts, we divided the number of coding bases by the total number of bases in the interval file. To generate null distributions representative of the genome-wide background for protein-coding genes, genes with a history of positive selection, and coding base counts, we created 1,000 replicate interval sets by randomly permuting intervals of the same number and size of the 10kb introgression tracts into regions confidently called for the species tree using bedtools shuffle (Quinlan and Hall 2010). We then compared the mean and standard deviation of each metric for the introgressed set to the 95% confidence intervals of the null distribution representative of the genome-wide background.

2.5 Tables and Figures

Table 2.1. Summary of DNA sequencing and coverage

Species	% Bases Covered by 5 reads			Mean Coverage Depth		
	Whole Genome[a]	Coding[b]	Introgression Tracts[c]	Whole Genome[d]	Coding[e]	Introgression Tracts[f]
Sdro	62	90	66	24.7x	52.9x	25.4x
Sfra	69	91	73	32.1x	54.8x	32.7x
Spal	57	85	61	11.9x	17.3x	13.5x
Hpul	53	83	54	24.5x	54.2x	23.5x

[a]Percentage of bases in the *S. purpuratus* reference genome covered with at least five reads.

[b]Percentage of coding bases in the *S. purpuratus* reference genome covered with at least five reads.

[c]Percentage of bases in the introgression tracts identified at the 90% posterior probability threshold covered with at least five reads.

[d]Mean genome-wide coverage depth of the *S. purpuratus* reference genome

[e]Mean coverage depth for coding sequences in the *S. purpuratus* genome assembly

[f]Mean coverage depth for the introgression tracts identified at the 90% posterior probability threshold.

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Hpul - H. pulcherrimus*

Table 2.2. Genomic features of the distribution of 10 kb introgression tracts and the genome-wide background at the 90% posterior probability threshold. Means are shown for the introgression tracts. The 95% confidence intervals are shown for the genome-wide background.

	d_{XY}	d_N	d_S	d_N/d_S	Genes / Mb	Percent Coding	PSGs[a]/ Mb
Introgression Tracts/Genes	0.041	0.001	0.05	0.19	17.6	3.75	0.68
Genome-Wide Background	0.048 - 0.054	0.011 - 0.019	0.042 - 0.069	0.25 - 0.45	28.6 - 28.9	5.19 - 5.23	1.58 - 1.65

[a]Positively Selected Genes: Genes with significant sites tests of positive selection across the strongylocentrotid family.

Table 2.3. A selection of genes with a history of positive selection within the strongylocentrotid sea urchin family that overlapped introgression tracts. The list is organized by the percentage of coding bases introgressed at the 90% posterior probability threshold.

Gene Info		Percent Bases Introgressed		Percent Coding Bases Introgressed			Mean Coverage Depth			
NCBI LOC ID	Name	prob > 0.9	prob > 0.8	prob > 0.9	prob > 0.8	Spal - Sdro bootstrap[a]	Sdro	Sfra	Spal	Hpul
LOC591845	arachidonate 5-lipoxygenase	92.4	100	71.4	100	99	25.8	33.5	11.9	26.6
LOC764716	helicase domino	52.9	66.3	63.2	79.2	100	34.3	44.9	13.5	56.2
LOC587208	kinesin-II 95 kDa subunit	14.8	36.8	51	61.3	93	32.1	42.3	17.5	24.2
LOC105444929	5-hydroxytryptamine receptor 6	6.1	33	39.6	100	98	45.9	41.2	14.1	45.5
LOC100893326	kremen protein 1*	46.3	60.7	34.9	54.7	82	26	35.9	9.7	19.9
LOC100893626	FAT tumor suppressor homolog 3-like	52	60.9	29.8	40.8	95	28.7	34.9	13	45
LOC100891604	transcription termination factor 1	8.9	66.5	21.8	100	46	22.2	29	7.8	23.5
LOC586606	MAK16 homolog	29.1	100	4.2	100	94	21.7	19.6	8.8	27.7
LOC115921720	glutathione peroxidase-like	0.0	100	0.0	100	n/a[b]	15.2	17.7	6.6	12
LOC115917953	2',3'-cyclic-nucleotide 3'-phosphodiesterase-like	0.0	100	0.0	100	34	18.2	23	6.2	20.4

[a]Bootstrap support for the branch grouping *S. pallidus* and *S. droebachiensis* as sister taxa.

[b]No high-quality variant genotypes were called for this gene despite sufficient coverage.

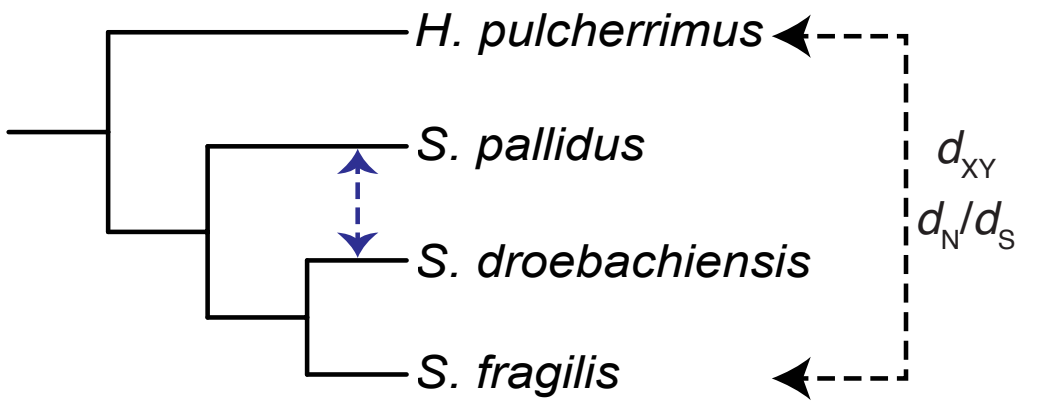
Species abbreviations: *Sdro*, *S. droebachiensis*; *Sfra*, *S. fragilis*; *Spal*, *S. pallidus*; *Hpul*, *H. pulcherrimus*.

Table 2.4. Summary of the top genes with introgressed bases that had significant branch-sites tests on either the *S. pallidus* or *S. droebachiensis* terminal branches.

Gene Info		Selective History		Percent Coding Bases Introgressed		Spal - Sdro boot[c]	Mean Coverage Depth			
NCBI LOC ID	**Name**	**PSG[a]**	**Branch[b]**	**prob > 0.9**	**prob > 0.8**		**Sdro**	**Sfra**	**Spal**	**Hpul**
LOC100893326	kremen protein 1	yes	*Spal*	34.9	54.7	82	26	35.9	9.7	19.9
LOC575079	arylsulfatase	yes	*Spal*	18.5	25.7	67	34.3	47.9	17.3	38.2
LOC580597	sodium- and chloride-dependent neutral and basic amino acid transporter B(0+)	yes	*Spal*	14.5	21.1	72	28	28.9	10.8	31.9
LOC590964	fibrosurfin-like	yes	*Spal*	7.1	18.3	100	44.5	51.3	16.3	110.7
LOC105445324	muscarinic acetylcholine receptor M5-like	no	*Spal*	0.0	39.1	87	33.1	33.8	6.2	53.8
eef1g	eukaryotic translation elongation factor 1 gamma	yes	*Spal*	0.0	28.6	89	34.3	35.4	16.8	40.2
LOC577068	prominin 1	yes	*Spal*	0.0	12.3	96	30.7	41.3	16.1	40.2
LOC584837	PHD finger protein 8	no	*Sdro*	20.3	42.8	100	23.2	38.3	9.6	41.7
LOC580943	structural maintenance of chromosomes 1A	yes	*Sdro*	4.3	39.7	59	25	35.7	11.7	37.5
LOC105442321	toll-like receptor 3	no	*Sdro*	0.0	13.3	94	58.1	54.8	18.6	73.7
LOC575208	death-inducer obliterator 1	no	*Sdro*	0.0	10.0	90	41.2	41.8	11.6	46.4

[a]PSG - Positively Selected Gene. A gene with a significant sites tests of positive selection across the strongylocentrotid family.
[b]Branch with significant branch-sites test.
[c]Bootstrap support for the branch grouping *S. pallidus* and *S. droebachiensis* as sister taxa

Species abbreviations: *Sdro*, *S. droebachiensis*; *Sfra*, *S. fragilis*; *Spal*, *S. pallidus*; *Hpul*, *H. pulcherrimus*.



alignment site	1	2	3	4	5	6	7	8	9	10
chromosome site	109	2,290	3,720	3,750	6,400	7,011	9,450	9,998	10,070	11,450
Hpul	A	C	G	T	T	A	C	G	G	T
Spal	G	**A**	**T**	T	**A**	T	**A**	**T**	**A**	T
Sdro	A	**A**	**T**	A	**A**	T	**A**	**T**	**A**	C
Sfra	A	C	G	T	T	T	G	G	G	C
introgression probability	.75	.9	.95	.95	.95	.95	.95	.95	.93	.5

introgression tract

Hypothetical Introgression Tract 1						Mean Coverage Depth			
Chromosome	Start	Stop	Length	SNV Sites	SNV/kb	Sdro	Sfra	Spal	Hpul
NW_022145609	109	11,450	11,341	10	0.89	w	x	y	z

Figure 2.1. Schematic of the study design. Introgressed genomic regions between *S. pallidus* and *S. droebachiensis* were identified using PhyloNet-HMM. PhyloNet-HMM outputs posterior probabilities of introgression for each site in a multiple sequence alignment. Introgression tracts were inferred by recording the genomic coordinates of consecutive sites with posterior probabilities at or above two different thresholds (90%, 80%). For each introgression tract, coverage depth metrics were estimated, and a gene tree was reconstructed for the region. For introgression tracts longer than 10kb, estimates of *H. pulcherrimus* - *S. fragilis* d_{XY} were made and compared to estimates obtained from genomic regions with high support for the species tree. All genes overlapping introgression tracts were identified. Estimates of d_N/d_S from sequence alignments of *H. pulcherrimus* and *S. fragilis* were made for genes with more than half of their bases declared introgressed and compared to estimates for non-introgressed genes. *S. pallidus* and *S. droebachiensis* were not used in the d_{XY} or d_N/d_S estimates because introgression in either direction could confound the estimates.

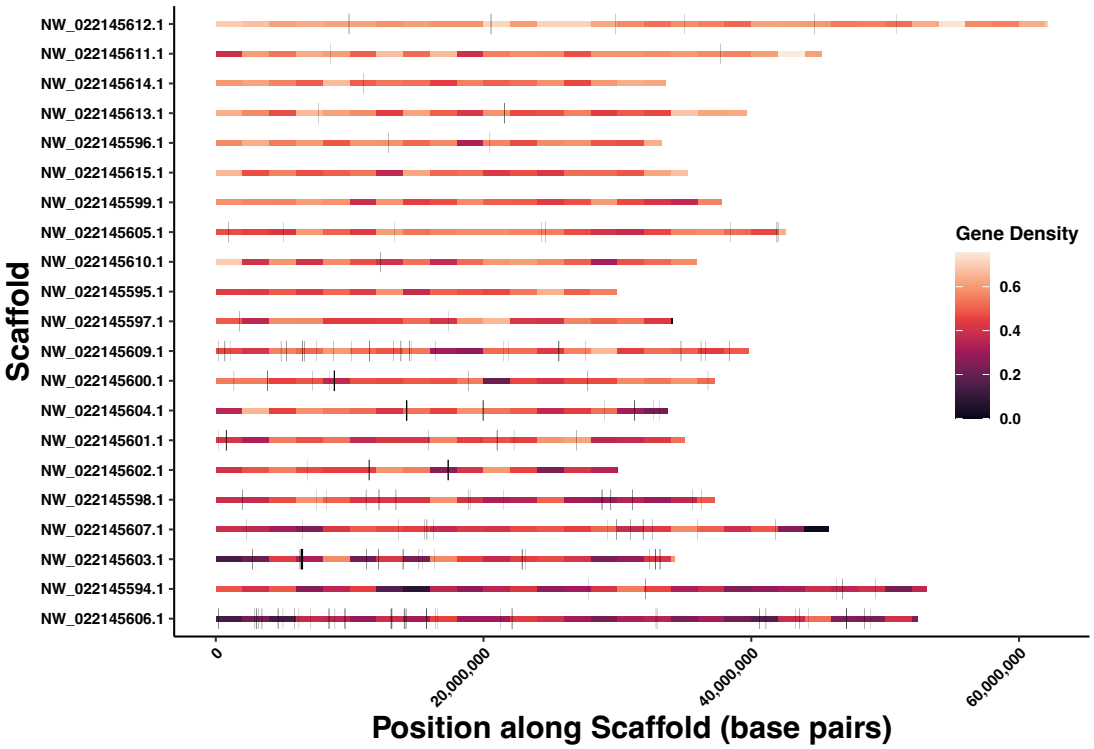


Figure 2.2. The 164 introgression tracts greater than 10 kb in length by chromosome (posterior probability > 90%). The introgression tracts are displayed as black rectangles along the chromosomes. The chromosomes are ordered by gene density (descending). The chromosomes are colored by gene density in windows of 2Mb.

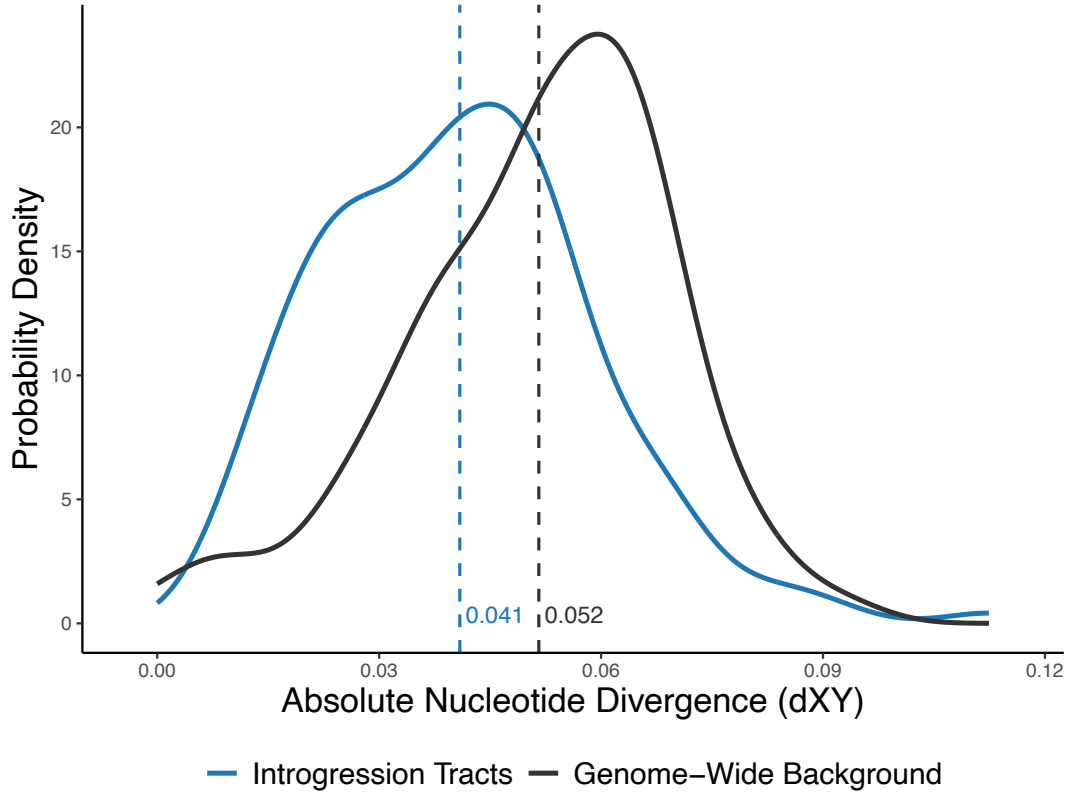


Figure 2.3. Absolute nucleotide divergence (*d*XY) between *H. pulcherrimus* and *S. fragilis* for the introgressed intervals vs. a random sample of non-Introgressed intervals of the same number and length confidently called for the species tree by PhyloNet-HMM.

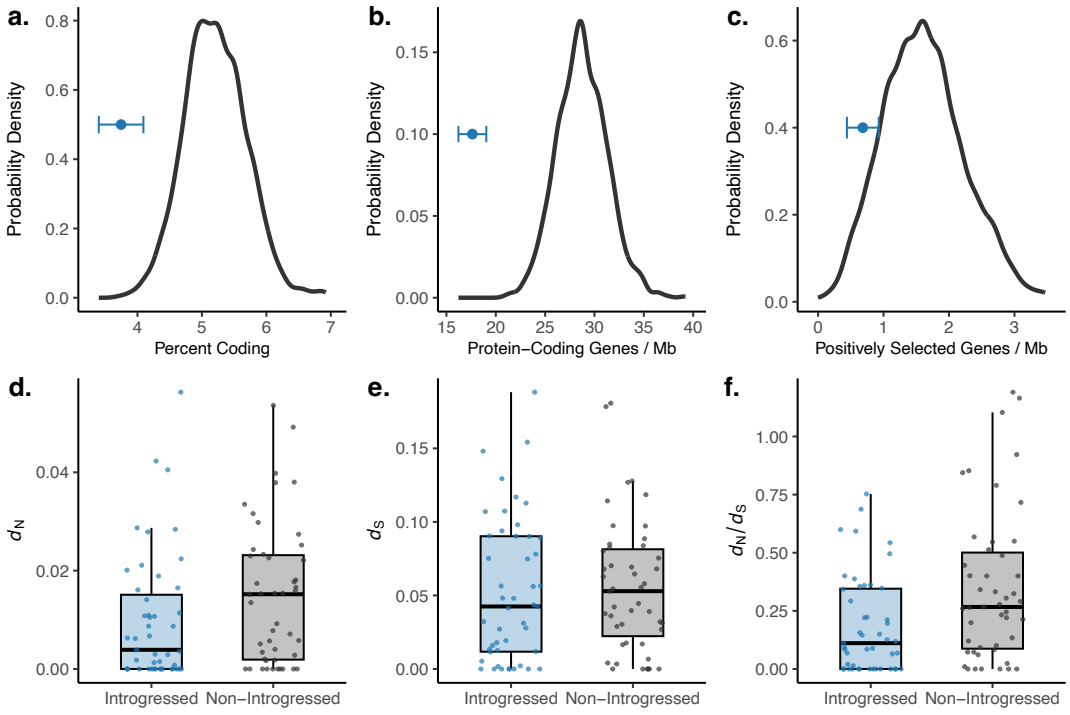


Figure 2.4 . Properties of introgressed regions and genes relative to random non-introgressed genes representative of the genome-wide background. (a) The percentage of bases that are coding for the introgression tracts (blue) is lower than the genome-wide background. (b) The number of overlapping protein coding genes, standardized by the combined number of introgressed bases in Mb, is lower than the genome-wide background. (c) The number of overlapping positively selected genes, standardized by the number of bases in the interval files, is lower for introgression tracts than the genome-wide background. Errors bars in (a-c) represent the standard deviation. (d-f) d_N, d_S, and d_N/d_S are lower for introgressed genes than non-introgressed genes. d_N, d_S, and d_N/d_S were estimated on protein-coding alignments of *H. pulcherrimus* and *S. fragilis*.

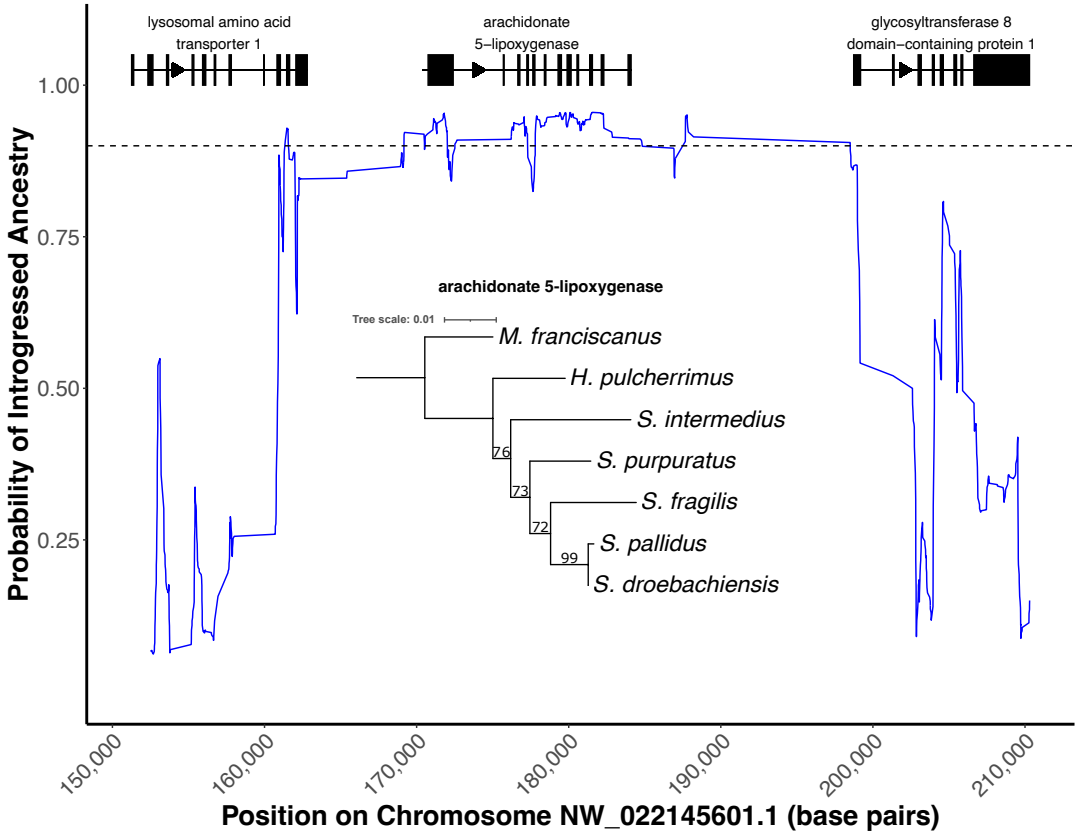


Figure 2.5. The introgression tract overlapping arachidonate 5-lipoxygenase, a gene with a history of positive selection within the strongylocentrotid sea urchin family. The maximum likelihood gene tree for the longest isoform of arachidonate 5-lipoxygenase shown inside the plot strongly supports the introgression tree topology ((*S. droebachiensis*, *S. pallidus*), *S. fragilis*).

Chapter 3 Positive selection and adaptive introgression at sea urchin gamete recognition proteins

Abstract

Broadcast spawning marine invertebrates often exhibit species-selective fertilization mediated by gamete recognition proteins (GRPs) located on the surfaces of sperm and egg cells. Although GRPs often evolve rapidly under positive selection and have been implicated in establishing reproductive isolation early in speciation, the selective pressures responsible for GRP divergence remain poorly understood. We characterized the molecular evolutionary histories of the sperm GRP bindin and its egg receptor EBR1 among nine species of the strongylocentrotid family of sea urchins. While the patterns of selection at bindin have been characterized across several echinoid genera, previous studies lack either complete taxa sampling or full bindin sequences. Additionally, few studies have characterized selection on EBR1 due to its large size and variable structure among species. To address these limitations, we tested for positive selection using multi-species multiple sequence alignments of complete bindin and EBR1 for all nine strongylocentrotid species. Our analysis revealed strong signatures of pervasive and episodic selection at both genes and identified the codon sites responsible. We also observed considerable gene tree discordance at both bindin and EBR1 and tested for introgression, finding strong support for historical introgression at both genes between divergent, non-sister taxa. It is highly likely that

the introgression of these proteins was adaptive, given the critical role of these proteins in fertilization and that the introgression involved amino acid substitutions. While our findings confirm rapid evolution at both bindin and EBR1, the introgression observed is inconsistent with bindin and EBR1 playing an important role in establishing reproductive isolation during strongylocentrotid sea urchin speciation.

3.1 Introduction

Sea urchins have played an outsized role in our understanding of fertilization, cell division, and embryology. They became model organisms in developmental biology during the 19th century due to their transparent embryos and the ease of obtaining large amounts of sperm and eggs while rearing them in the lab. Sea urchins have been instrumental in numerous pivotal discoveries, including the first evidence of sperm and egg pronuclear fusion (Hertwig 1876), the chromosomal theory of inheritance (Boveri 1902), and the discovery of the cyclin family of proteins that control the progression of the cell cycle (Evans et al. 1983). In evolutionary biology, sea urchins have been studied extensively for the role of fertilization in reproductive isolation. Early on, sea urchin fertilization was revealed to be species-selective, where selectivity is mediated by gamete recognition proteins (GRPs) on the surfaces of sperm and egg cells involved in gamete binding and fusion (Summers and Hylander 1975; Glabe and Vacquier 1977a; Metz et al. 1994). These GRPs were identified as the sperm protein bindin (Vacquier and Moy 1977) and its glycoprotein receptor EBR1 (Kamei and Glabe 2003).

The discovery that species selective fertilization was determined by recognition molecules on the sperm and egg surfaces fueled considerable investigation into their importance in reproductive isolation. During sea urchin fertilization, the sperm undergoes the acrosome reaction after contacting the egg jelly, where the acrosomal vesicle is exocytosed, creating a protrusion coated in mature bindin protein called the acrosomal process. Bindins on the acrosomal process must attach to EBR1 in the

vitelline envelope for the acrosomal process to penetrate the vitelline envelope and fuse with the egg plasma membrane. As congeneric sea urchin species show species-selectivity in gamete binding and often lack other common isolating barriers, it was suggested that gametic isolation may be one of the first reproductive barriers to evolve during echinoid speciation (Palumbi 1992; S R Palumbi 2009). Consistent with this theory, bindin has been shown to evolve rapidly under positive selection in several genera (Palumbi and Metz 1991; Metz and Palumbi 1996; Geyer and Palumbi 2003; McCartney and Lessios 2004; Calderón et al. 2009; Pujolar and Pogson 2011) and its sequence divergence is negatively correlated with gametic compatibility (Kirk S Zigler et al. 2005). Furthermore, bindin divergence sometimes accumulates faster among sympatric species than allopatric species (Palumbi and Lessios 2005; Lessios and Zigler 2012), and several theoretical models have demonstrated that speciation is a possible consequence of rapid bindin and EBR1 coevolution (Van Doorn et al. 2001; Gavrilets and Waxman 2002). Bindin has thus been categorized as a speciation gene involved in establishing reproductive isolation early in speciation (Noor and Feder 2006; Nei and Nozawa 2011; Blackman 2016). However, the forms of selection acting on bindin remain poorly understood and little is known about selection acting on EBR1 due to its large size and considerable structural variation among species. Empirical research on the role of bindin and EBR1 in speciation remains limited.

The major hypotheses for the rapid diversification of sea urchin gamete recognition proteins are sperm competition, sexual conflict, and reinforcement. Differentiating between these hypotheses is difficult because they are not mutually

exclusive and make overlapping predictions. Consistent with reinforcement, sea urchin genera with more sympatric species tend to have higher rates of bindin divergence (Lessios and Zigler 2012). However, this pattern could also be produced without reinforcement if significant bindin divergence is required for species to co-exist in sympatry (Lessios 2007). Studies characterizing the signal of selection at bindin across different genera indicate that reinforcement is possible (Geyer and Palumbi 2003; Zigler et al. 2003) but not common (McCartney and Lessios 2004; Geyer and Lessios 2009; Pujolar and Pogson 2011). The presence of selection not only between species but also between alleles within the same species has been repeatedly observed (Palumbi 1999; Levitan and Ferrell 2006; Levitan and Stapper 2010; Stapper et al. 2015), which cannot be explained by reinforcement. The leading hypothesis with the most support explaining the polymorphism observed at GRPs within species is sexual conflict mediated by polyspermy risk (Levitan and Stapper 2010; Pujolar and Pogson 2011). However, inference on the role of polyspermy is severely limited by the scarcity of information about selection acting on EBR1.

Here, we provide the most comprehensive characterization yet of the molecular evolution of bindin and EBR1 using multi-species multiple sequence alignments of both complete genes. We first tested for signals of historical positive selection and then tested for introgression after observing considerable gene tree discordance in both genes. To place the molecular evolution of bindin and EBR1 into context, we tested whether the other strongylocentrotid sea urchin sperm and egg proteins involved in reproduction showed similar signals of selection and introgression. Finally, we

attempted to characterize the interaction between bindin and EBR1 using molecular docking software.

3.2 Results

3.2.1 DNA Sequencing and Multiple Sequence Alignments

To characterize the molecular evolution of bindin, EBR1, and other sea urchin reproductive proteins, we created multi-species multiple sequence alignments for the protein-coding regions of each gene, composed of a single sequence from each of the nine strongylocentrotid species. Sequence alignments were constructed from whole-genome sequencing data from each species aligned to the *S. purpuratus* reference genome. To guide the multi-species multiple sequence alignments, we used the cDNA sequences of *S. purpuratus* bindin (Gao et al. 1986), *S. purpuratus* and *M. franciscanus* EBR1 (Kamei and Glabe 2003), and the gene models from the Spur_3.1 NCBI RefSeq assembly. We removed gaps and nucleotide sites with missing data for the selection and introgression tests. Our final multiple sequence alignment lengths were 1,260 base pairs (420 amino acids) for bindin and 7,575 base pairs (2,525 amino acids) for EBR1. Below, we will refer to individual nucleotide and codon sites using the coordinates from our multiple sequence alignments. A mapping of the coordinates in our multiple sequence alignments to the complete *S. purpuratus* NCBI proteins (BND, NCBI Gene ID: 373276; ebr1, NCBI Gene ID: 577775) is available in Supplementary Table S28.

3.2.2 Selection Tests

To test for positive selection at bindin and EBR1, we performed complementary analyses using the codeml program of PAML (Yang 2007) and the BUSTED (Murrell et al. 2015), FUBAR (Murrell et al. 2013), MEME (Murrell et al. 2012), and aBSREL (Smith et al. 2015) programs of the HyPhy package (Kosakovsky Pond et al. 2006). PAML and FUBAR test for pervasive selection across the phylogeny, while BUSTED, MEME, and aBSREL test for episodic selection. We also implemented the GARD (Kosakovsky Pond et al. 2006) program of HyPhy to screen the bindin and EBR1 multiple sequence alignments for recombination and assess whether the presence of recombination could cause spurious results. We reran the HyPhy selection tests using separate partitions corresponding to the recombination blocks identified by GARD. We also performed a similar exercise for the PAML analysis by running the codeml program on the individual exon alignments with their own inferred maximum-likelihood tree topologies.

3.2.2.1 Bindin

There was strong evidence of positive selection at bindin. Three bindin codon sites showed significant pervasive positive selection in both the PAML and HyPhy FUBAR analyses (Table 3.1, Figure 3.1). There were another two significant codon sites identified by PAML only and nine sites identified by FUBAR only. The BUSTED test of episodic diversifying selection was significant (p=0.0007), and the MEME test identified four sites under episodic diversifying selection, three of which were not

identified by PAML or FUBAR. A single branch was under episodic diversifying selection with the aBSREL test, which was the common ancestor of *M. franciscanus*, *M. nudus*, and *P. depressus*. The FUBAR test also revealed pervasive negative selection at ten sites.

HyPhy Gard identified four recombination breakpoints, and the analysis with the partitioned data returned similar results to the non-partitioned data. The BUSTED test of episodic selection remained significant (p=0.0082). Eight of the thirteen sites under pervasive positive selection in the non-partitioned FUBAR analysis remained significant, and the five sites that lost significance retained high posterior probabilities of positive selection (> 85%). The ten codon sites under pervasive negative selection in the non-partitioned FUBAR analysis remained significant. Finally, three of the five sites under positive selection identified by PAML remained significant in the exon-by-exon analysis at the 95% posterior probability threshold, and the two sites that lost significance had Bayes Empirical Bayes posterior probabilities > 90%. The results of the partitioned selection tests demonstrate that the presence of recombination did not lead to false inference of selection at bindin.

3.2.2.2 EBR1

EBR1 also showed strong positive selection, evidenced by eleven codon sites with significant posterior probabilities of pervasive positive selection in both the PAML and HyPhy Fubar analyses (Table 3.1, Figure 3.2). An additional eight significant codon sites were identified by PAML only, and twelve were identified by

FUBAR only. The BUSTED test of episodic diversifying selection was significant ($p<0.0001$) and the MEME test found evidence of episodic positive selection at 46 sites. The aBSREL test found nine branches to have experienced episodic diversifying selection, including the *P. depressus*, *S. purpuratus*, *S. intermedius*, and *S. pallidus* terminal branches. The FUBAR test also revealed pervasive negative selection at 264 sites.

HyPhy Gard inferred 30 recombination breakpoints in our EBR1 multiple sequence alignment. When the selection tests were rerun with the partitioned data, the BUSTED test remained highly significant (p=8.854e-7), but only six of the 23 positively selected codon sites in the non-partitioned FUBAR analysis remained significant. However, the posterior probabilities generally remained high for the 18 codon sites no longer significant in the partitioned analysis; fourteen had posterior probabilities of positive selection greater than 80%. The number of sites showing significant pervasive negative selection decreased from 264 to 145. In the exon-by-exon tests with PAML, fourteen of the original nineteen positively selected codon sites remained significant. As with bindin, recombination among our EBR1 sequences did not lead to false inferences of selection.

We examined whether positively selected codons were enriched in any of the different domains present within EBR1. The *S. purpuratus* EBR1 is composed of a propeptide, a reprolysin, an ADAM cys-rich domain, eight thrombospondin-type-1 (TSP-1) domains, eight and one-half alternating TSP-1 and CUB (C1s/C1r, uEGF, Bmp1) domain repeats, eleven hyalin-like (HYR) domains, and one final CUB domain

(Kamei and Glabe 2003). The alternating TSP-1 and CUB domains have been referred to as EBR repeats (Kamei and Glabe 2003; Vacquier and Swanson 2011). The second EBR repeat has a duplicated CUB domain for a gene-wide total of 10 CUB domains and 17 TSP1 domains. Much of EBR1's structure is conserved between *S. purpuratus* and *M. franciscanus*, who diverged 13-19 mya (Kober and Bernardi 2013c). However, the HYR repeats of *S. purpuratus* do not appear present in *M. franciscanus* EBR1, and *M. franciscanus* has an additional ten EBR repeats relative to *S. purpuratus* (Kamei and Glabe 2003). The propeptide, reprolysin, and ADAM cys-rich domains had no codon sites under pervasive positive selection (expected: 1.6, 2.4, 0.9, respectively). Twelve of the 31 codon sites under pervasive positive selection fell within TSP1 domains (expected: 10.0), and another ten fell within CUB domains (expected 11.0). For the branch-sites tests of episodic selection with HyPhy MEME, the propeptide and reprolysin each had one positively selected codon (expected: 2.5, 3.8, respectively), the TSP-1 domains had fourteen sites (expected: 15.9), and the CUB domains had nineteen sites (expected: 17.4). The positively selected codons do not appear to be notably enriched or under enriched in any domain type and do not appear to be nonrandomly clustered in any regions along the gene (Figure 3.2).

3.2.2.3 Auxiliary Reproductive Proteins

We also characterized the selective histories of 111 additional sea urchin reproductive genes (64 sperm proteins and 47 egg proteins) using PAML and HyPhy BUSTED. A total of 48 (43%) genes had significant BUSTED or PAML tests, 18 of

which were significant in both the PAML and BUSTED tests (Supplementary Table S29). PAML identified 35 genes (32%) under pervasive positive selection (15/64 sperm proteins; 20/47 egg proteins), while BUSTED identified 31 (30%) genes under episodic diversifying selection (13/64 sperm proteins; 18/47 egg proteins). We used Fisher's exact test to determine whether reproductive proteins were more likely to evolve under pervasive positive selection than a set of 6,441 single-copy orthologs not known to function in reproduction. The reproductive proteins were more likely to have evolved under pervasive positive selection than a random distribution of non-reproductive proteins ($p<0.0001$). The egg genes were also more likely to be under selection than a random draw ($p=0.003$), while the sperm genes were not significantly different ($p=0.35$).

3.2.3 Introgression Tests

The tree topologies of both bindin and EBR1 were discordant with the species tree relationships, suggestive of potential introgression between *S. pallidus* and *S. droebachiensis* at bindin and *H. pulcherrimus* and *S. droebachiensis* at EBR1 (Figure 3.3). Furthermore, the phylogenetic relationships varied across each gene, and HyPhy GARD identified recombination in both multiple sequence alignments. To determine whether the observed discordance was caused by introgression or incomplete lineage sorting, we applied PhyloNet-HMM (Liu et al. 2014), a comparative genomic model for detecting introgression, to our multiple sequence alignments of bindin, EBR1, and the set of auxiliary reproductive proteins. PhyloNet-HMM scans sequence alignments

for regions supporting different phylogenetic relationships and outputs posterior probabilities of introgression for each site in a multiple sequence alignment. The model is incomplete-lineage sorting aware and accounts for convergence using a finite-sites model (Liu et al. 2014; Liu et al. 2015; Schumer et al. 2016). We ran PhyloNet-HMM 100 times on each multiple sequence alignment and averaged the posterior probabilities across runs to avoid reaching local optima during hill climbing. We used a posterior probability threshold of 90% as evidence for introgression and recorded the sites at or above this threshold within each gene.

To complement the PhyloNet-HMM analysis, we also used Patterson's *D* (ABBA-BABA test) (Green et al. 2010; Durand et al. 2011) as a framework to characterize the patterns of allele sharing between *S. pallidus* and *S. droebachiensis* in bindin and *H. pulcherrimus* and *S. droebachiensis* in EBR1. Normally, Patterson's *D* should not be applied to individual genes or small genomic regions, as *D* outliers tend to cluster in low-divergence regions and are prone to false positives (Martin et al. 2015). However, the EBR1 gene model spans over 70kb, shows high pairwise genetic distance between the species in our alignment, and has experienced considerable recombination, making the block jackknife procedure for testing for significance reasonable. Furthermore, the two species pairs have significant genome-wide tests for introgression (Glasenapp and Pogson 2023), and we had reason to believe these genes had experienced introgression *a priori* (Pujolar and Pogson 2011).

3.2.3.1 Bindin

The gene tree for bindin places *S. droebachiensis* sister to *S. pallidus* rather than its true sister taxa, *S. fragilis* (Figure 3.3). PhyloNet-HMM identified one introgression tract spanning 40 bases towards the end of exon five of our multiple sequence alignment (Figure 3.1). Within this introgression tract are four codon sites where *S. pallidus* and *S. droebachiensis* share a derived amino acids not seen in the other species (Figure 3.4). Across the entire protein-coding gene alignment, *S. pallidus* and *S. droebachiensis* share six derived nucleotide sites and five amino acid sites (ABBA site pattern) relative to one shared, derived nucleotide site and amino acid site between *S. pallidus* and *S. fragilis* (BABA sites) (Supplementary Tables S30-S31). Five of the six shared, derived nucleotide sites between *S. pallidus* and *S. droebachiensis* fell within the introgression tract identified by PhyloNet-HMM. Two introgressed codons identified by PhyloNet-HMM were also significant in the tests for positive selection (codons 396, 398) (Table 3.1, Figure 3.4). None of the codon sites with evidence of pervasive negative selection showed introgression.

3.2.3.2 EBR1

There was also considerable phylogenetic discordance across EBR1. The overall tree topology for the gene has *S. droebachiensis* pulled up as the first branching member of *Strongylocentrotus*, following the branching of *H. pulcherrimus* (Figure 3.3). This unusual placement of *S. droebachiensis* was caused by the sharing of a large excess of unique mutations between *H. pulcherrimus* and *S. droebachiensis* relative to

the other *Strongylocentrotus* congeners (Table 3.2), which is qualitatively inconsistent with incomplete lineage sorting. To survey gene tree discordance across the gene, we reconstructed gene trees for each EBR1 exon in our multiple sequence alignment. Three exons (38, 40, and 41) group *H. pulcherrimus* and *S. droebachiensis* together as sister taxa, with 91%, 99%, and 26% bootstrap support, respectively. Another eleven exons place *S. droebachiensis* as the first branching member of *Strongylocentrotus*, consistent with the topology of the overall gene tree (Supplementary Table S32). The strongest signal of discordance occurs among exons 37-41 in our alignment (693 base pairs), where 88 parsimony-informative sites place *S. droebachiensis* sister to *H. pulcherrimus* with 97% bootstrap support (Supplementary Figure S7). These exons are also strong outliers in comparisons of genetic distance (K2P) between sister taxa *S. droebachiensis* and *S. fragilis* (Figure 3.5), consistent with the introgression of *H. pulcherrimus* alleles into *S. droebachiensis*. The net direction of introgression can be inferred as *H. pulcherrimus* into *S. droebachiensis* because the rest of the gene tree is concordant with the species tree, and only the introgression of *H. pulcherrimus* alleles into *S. droebachiensis* would cause it to be pulled up as sister to *H. pulcherrimus* (Figure 3.3). Introgression in the opposite direction would cause *H. pulcherrimus* to be pulled down near the position of *S. droebachiensis* within the *Strongylocentrotus* species tree.

PhyloNet-HMM identified 653 (8.6%) nucleotide sites introgressed between *H. pulcherrimus* and *S. droebachiensis* (Figure 3.2). The introgression tracts were relatively short, which was expected given the high amount of recombination detected by HyPhy GARD and the old age of the introgression. Seven of the 31 codon sites

under pervasive positive selection in Table 3.1 fell completely within *H. pulcherrimus* - *S. droebachiensis* introgression tracts (codons 512, 1067, 1329, 1360, 2039, 2124). Another two positively selected codons (codons 1283, 2083) had one base declared introgressed, and two had one introgressed base (codons 646, 1332). Twelve of the 49 codons identified by the MEME test of episodic selection showed introgression. At codon site 1220, the MEME test of episodic diversifying selection was significant, *H. pulcherrimus* and *S. droebachiensis* share a derived amino acid (Serine) not seen in the other species, and all three bases in this codon had posterior probabilities of introgression > 99%. Nineteen of the 264 codon sites showing pervasive negative selection showed introgression.

The PhyloNet-HMM introgression signal was highest in EBR1 exons 37-41 (Figure 3.2, sites 2072 - 2341), the same exons showing the strongest gene tree discordance (Supplementary Figure S7). Several of these exons also showed strong selection in the individual exon tests. Within this region, the clearest introgression tract occurs in exon 40 of our multiple sequence alignment, where numerous sites had posterior probabilities of introgression >99% (Figure 3.6). This exon had 36 parsimony informative sites, 27 of which support a sister relationship between *H. pulcherrimus* and *S. droebachiensis*. Furthermore, at six of these sites, *H. pulcherrimus* and *S. droebachiensis* share a unique allele, while all seven other strongylocentrotid urchin species share a different allele. The tree topology for exon 40 places *S.* droebachiensis sister to *H. pulcherrimus* with 99% bootstrap support (Figure 3.6).

We further looked at patterns of allele sharing between *H.* pulcherrimus and *S. droebachiensis* to complement the PhyloNet-HMM analysis. In the coding sequence alignment of EBR1 for the rooted triplet (((*S. fragilis*, *S. droebachiensis*), *H. pulcherrimus*), *M. franciscanus*), there are 28 ABBA site patterns and 14 BABA patterns (D = 0.33; p=5.25e-11) (Supplementary Table S33). In the complete gene sequence alignment, including introns, there are 127 ABBA patterns and 43 BABA patterns (D = 0.49; p=0.0016) (Supplementary Table S34), and in the translated protein sequence, there are 2-3 times the number of ABBA amino acid sites than BABA amino acid sites, depending on the outgroup (Supplementary Table S35). Regardless of the choice of outgroup and P1 taxa, *H. pulcherrimus* and *S. droebachiensis* have an excess of shared, derived alleles, which cannot be explained by incomplete lineage sorting alone.

3.2.3.3 Auxiliary Reproductive Proteins

We tested the additional 111 reproductive proteins for both *S. pallidus* - *S. droebachiensis* and *H. pulcherrimus* - *S. droebachiensis* and introgression using the same PhyloNet-HMM protocol as used for bindin and EBR1 (Supplementary Table S29). Surprisingly, EBR1 had the 8th most introgressed nucleotide sits between *H. pulcherrimus* and *S. droebachiensis*, indicating that introgression between these two species is not unique to EBR1. Similar to introgression at EBR1, six of the seven other genes showing *H. pulcherrimus – S. droebachiensis* introgression were only partially introgressed (9.1% – 37%), except for spermatogenesis-associated protein 6

(SPU_025715, Sp-Spata6L, LOC757232), which had all of its bases declared introgressed. For introgression between *S. pallidus* and *S. droebachiensis*, 25 genes showed more introgressed sites than bindin. This was unsurprising given the small number of sites introgressed at bindin (40 bp, 3.2%) and the finding of significant genome-wide introgression between the two taxa (Glasenapp and Pogson 2023; Glasenapp and Pogson 2024). It was previously estimated that 1-5% of the genome had experienced introgression between *S. pallidus* and *S. droebachiensis* (Glasenapp and Pogson 2024).

3.2.4 Bindin and EBR1 Protein Structure and Interaction

Bindin exon five contains a repeat region with a seven amino acid motif repeated in variable numbers across species. It is nearly impossible to align short, paired-end reads to this region and create multi-species multiple sequence alignments. Therefore, we used Sanger sequencing to recover as much exon five as possible. Unfortunately, most of the region had to be excluded from the multi-species alignment because the *Mesocentrotus* and *Pseudocentrotus* species only have one or two of the repeat motifs, while the other species have seven to thirteen. To investigate the diversification of the repeats, we repeated the exercise performed in Figure 3 of Biermann (1998), comparing the structure and length of the repeat section across species. Similar to Biermann (1998), we found an expansion of nearly 2x in the number of repeats in *S. intermedius* and *S. pallidus* relative to *H. pulcherrimus*, *S. purpuratus*, *S. droebachiensis* and *S. fragilis* (Supplementary Figure S8). It is difficult to say

whether this expansion (a) occurred in the common ancestor of *S. intermedius*, *S. pallidus*, *S. droebachiensis*, and *S. fragilis*, (b) evolved in *S. intermedius* and *S. pallidus* independently, or (c) evolved in either *S. pallidus* or *S. intermedius* and was homogenized via introgression, as *S. intermedius* and *S. pallidus* do show limited genome-wide introgression (Glasenapp and Pogson 2023).

We attempted to predict the interface resides between bindin and EBR1 using AlphaFold (Jumper et al. 2021) and HDOCK (Yan et al. 2017; Yan et al. 2020). We downloaded the bindin structure prediction (Accession P06651) from AlphaFold Protein Structure Database (Varadi et al. 2022; Varadi et al. 2024). EBR1 was not in the AlphaFold database due to its long length (3,712 amino acids), so we ran the AlphaFold algorithm to obtain a structure prediction. We then ran HDOCK (Yan et al. 2017; Yan et al. 2020), a web server-based application for analyzing protein-protein docking, to predict interface residues in the bindin-EBR1 interaction. The best HDOCK model had a confidence score of 0.98, indicating a high likelihood that the two proteins would bind. The model identified 53 interface residues in EBR1 and 31 in bindin. The bindin interface residues clustered near the boundary of exons three and four and the beginning of exon five (Figure 3.1). Forty-one of the 53 EBR1 interface residues fell within the 3' most CUB domain following the *S. purpuratus* HYR repeat region, while another eight fell in the HYR-like repeats (Figure 3.2). Caution should be taken in interpreting the HDOCK results (interface residues) as the confidence in the EBR1 structure was low for most of the protein (Supplementary Figure S27). Most of the EBR1 sequence does not have homology to other known proteins with predicted

structures. The regions with the highest confidence were the reprolysin, several CUB domains, and the hyaline-like repeat domains. The TSP-1 domains and the last CUB domain had very low confidence scores, where the structure should not be interpreted.

We intersected the bindin and EBR1 interface residues with the positively selected, negatively selected, and introgressed codons. For bindin, two of the 31 interface residues showed positive selection (codons 148 and 155 in our sequence alignment), while none showed negative selection or introgression. To determine whether the number of overlaps observed differed from the number expected due to chance, we randomly sampled 31 codon sites without replacement 100,000 times and counted the number of times a positively selected, negatively selected, or introgressed codon was selected. Observing two interface residues showing positive selection was slightly more than expected due to chance (mean: 1.25, 99% confidence interval: 1.24 - 1.26), while observing zero interface residues showing negative selection or introgression was slightly less than expected due to chance (expected: 0.73, 0.96, respectively). For EBR1, one interface residue showed positive selection (codon 2476, expected: 1.4), two showed negative selection (expected: 5.5), and none showed introgression (expected: 3.4).

3.3 Discussion

Our study has provided strong evidence of historical positive selection and introgression at both bindin and EBR1. It is also the first to characterize the molecular evolution of the complete EBR1 gene and the most comprehensive characterization of

selection at bindin to date. Although our findings demonstrate that GRPs have evolved under strong positive selection in the strongylocentrotid family, the observed historical adaptive introgression between well-diverged, non-sister taxa is inconsistent with bindin and EBR1 playing a major role in establishing reproductive isolation early in speciation, as introgression should be suppressed near loci involved in reproductive isolation (Nosil 2012; Ravinet et al. 2017; Elmer 2019). Our results indicate that the rapid diversification of strongylocentrotid GRPs likely did not cause reproductive isolation, emphasizing the importance of additional reproductive isolating barriers during speciation.

Both PAML and HyPhy many codons evolving under positive selection at both bindin and EBR1. We were able to analyze 420 of the 481 (87%) amino acids in *S. purpuratus* bindin, and 2,525 of the 3,712 (68%) amino acids in *S. purpuratus* EBR1. The likelihood ratio tests for selection were considerably higher for EBR1 than bindin, although bindin had a higher proportion of codon sites showing evidence of selection. Neither the positively nor the negatively selected codon sites at either gene appeared localized to specific exons, regions, or domains. In EBR1, the number of positively selected sites by domain was similar to expectations due to chance based on the proportion of EBR1 covered by each domain type. The HyPhy aBSREL branch models of lineage-specific diversification were qualitatively consistent with sexual conflict as a driver of selection. Sexual conflict is expected to be strongest under high population density, where multiple males compete to fertilize individual eggs. The species more likely to exist at high population densities (e.g., *S. purpuratus*) had elevated

nonsynonymous substitution rates, while those more likely to occur at lower population densities did not (e.g., *S. droebachiensis* and *S. fragilis*) did not. However, further population-level sampling is needed to confidently rule out a history of reinforcement selection.

Bindin and EBR1 both contain tandem repeat regions of variable length between and within species, where reference genome alignment, variant calling, and multiple sequence alignment are not possible. Near the end of bindin exon five, there is a repeated motif of 21 nucleotides (7 amino acids) that varies widely in copy number among the *Strongylocentrotus* species, prohibiting multi-species alignments. Additionally, the *S. purpuratus* EBR1 contains eleven hyalin-like (HYR) repeats of around 81 amino acids in length that are thought to play a role in the species-selective binding of gametes (Kamei and Glabe 2003). These repeats are not present in *M. franciscanus* (Kamei and Glabe 2003) and were not included in our multi-species multiple sequence alignments. Unfortunately, we could not test whether selection was responsible for the considerable diversification in the repeat region at the end of bindin exon five and the hyalin-like repeat region of EBR1. The repeats in both sections have too few phylogenetically informative sites to trace their evolutionary origin and diversification. A more complete analysis of protein structure, polymorphism, and divergence for these genes will require *de novo* assembly and RNA-Seq.

It is highly likely that the introgression at bindin and EBR1 was adaptive. Introgressed regions contained amino acid substitutions in the species involved in introgression, and there was limited overlap between positively selected and

introgressed codons. Based on estimates of silent nucleotide polymorphism, the introgression at EBR1 preceded the introgression at bindin. Given that both introgression events involved *S. droebachiensis*, we speculate that the adaptive introgression at bindin may have evolved in response to EBR1 introgression, which may be explained by sexual conflict mediated by polyspermy. While the finding of introgression at bindin was unexpected, it was not surprising that the introgression occurred between *S. pallidus* and *S. droebachiensis*. The two species co-occur across their ranges in both the Pacific and Atlantic, and there is well-documented evidence of introgression between the pair (Addison and Hart 2005b; Harper and Hart 2007; Addison and Pogson 2009b; Pujolar and Pogson 2011; Glasenapp and Pogson 2023; Glasenapp and Pogson 2024). More surprising was the finding of introgression between *H. pulcherrimus* and *S. droebachiensis*, who diverged 10-14 mya (Kober and Bernardi 2013c) and are thought to be currently allopatric. However, there may be limited geographic overlap in the West Pacific, where their ranges are not well-characterized, and there certainly could have been overlap in the past. *S. droebachiensis* occurs in the Sea of Okhotsk and has been sampled as far south as Onekotan Island (Vasileva et al. 2017). *H pulcherrimus* has been sampled 2,000 kilometers south of Onekotan Island at Onagawa Bay, Honshu, Japan (Agatsuma et al. 2006). Given that the *S. droebachiensis* sample was taken from the San Juan Islands and the *H. pulcherrimus* sample was taken from Japan, the introgression detected between the two did not occur recently and is most likely fixed. Previous studies have documented the heightened susceptibility of *S. droebachiensis* eggs to fertilization by heterospecific sperm relative to other

strongylocentrotid species (Strathmann 1981; Levitan 2002c). It would be worth testing whether this increased susceptibility is caused by the introgressed *H. pulcherrimus* alleles.

The main signal of introgression in EBR1 occurred in exons 37-41 of our multiple sequence alignment, which spans the 5th -7th EBR repeat motif of alternating TSP-1 and CUB domains. Both pervasive and episodic selection have occurred in this region, and many of the positively selected codon sites remained significant in the partitioned analysis, indicating that the selection signal was not caused by gene tree misspecification. This region is thought to play an important role in the interaction with bindin (Kamei and Glabe 2003). It is unclear why introgressed alleles were favored so heavily in this particular region, but we speculate that it must have been adaptive, given the presence of introgressed amino acid substitutions (Supplementary Table S35) and the high efficiency of selection expected in sea urchins (Kober and Pogson 2013; Kober and Pogson 2017). As EBR1 is under strong selection, if the introgressed alleles from *H. pulcherrimus* resulted in reduced fitness in *S. droebachiensis*, they almost certainly would have been removed by selection. Since EBR1 is only expressed in females, introgression could have proceeded through males, where introgressed haplotypes could have avoided selection and been fragmented by recombination, weakening the strength of negative selection when expressed in females. One adaptive hypothesis for the introgression of EBR1 is that sexual conflict mediated by polyspermy within *S. droebachiensis* favored the incorporation of EBR1 mutations from a different species. The risk of polyspermy is thought to favor egg protein mutations that reduce

compatibility with common sperm protein genotypes, leading to fewer polyspermic zygotes (Tomaiuolo and Levitan 2010; Levitan et al. 2019). Although *S. droebachiensis* currently tends to exist at low population densities in the Northeast Pacific, sexual conflict is still likely to occur under sperm limitation (Franke et al. 2002; Levitan et al. 2007), and *S. droebachiensis* exists at high population densities in other locations in the Atlantic. If the introgressed *H. pulcherrimus* alleles from *H. pulcherrimus* slowed the rate of binding of sperm for the most common bindin genotypes and reduced polyspermy, they may have been favored by selection. In response to the introgression from *H. pulcherrimus*, we hypothesized that we would see more adaptive substitutions on the *S. droebachiensis* branch to compensate for the new alleles. However, branch-sites tests did not reveal elevated adaptive diversification on this branch of the phylogeny.

Our molecular docking analysis revealed candidate amino acid residues involved in the binding interaction between bindin and EBR1. The EBR1 interface residues identified mainly fell within the hyalin-like domain repeats and the final CUB domain that follows the HYR repeats. For bindin, the interface residues clustered near the boundary of exons three and four and the beginning of exon five. None of the interface residues for either gene showed introgression. The interface residues should be interpreted cautiously due to the low confidence in the protein structure prediction for EBR1. Furthermore, identifying recognition sites in EBR1 may be difficult because the interaction with bindin also involves EBR1's carbohydrate components (Foltz 1994; Biermann et al. 2004) .

Consistent with previous studies, we found reproductive proteins to have higher rates of adaptive diversification (Swanson and Vacquier 2002a; Turner and Hoekstra 2008). Caution should be taken when interpreting these results, as our list of 111 reproductive proteins likely includes many housekeeping genes that are expressed in sperm and egg cells but do not have a direct role in reproduction. Contrary to expectations and empirical work in *Drosophila* (Swanson et al. 2001; Meiklejohn et al. 2003; Clark et al. 2007), we found elevated evolution of female reproductive proteins relative to male proteins, which may highlight the importance of sexual conflict as a driver of adaptive diversification in sea urchins. On the male side, seven of the fifteen positively selected genes encode sea urchin receptor for egg jelly (suREJ) proteins. Sea urchin REJ proteins are expressed in the sea urchin plasma membrane, likely play a role in the acrosomal reaction (Mengerink et al. 2002), and have been shown to be under selection in *Strongylocentrotidae* (Mah et al. 2005; Pujolar and Pogson 2011). However, suREJ protein expression is not restricted to the sperm. Excluding these proteins from the analysis would dampen the signal of positive selection among male reproductive proteins and bolster our finding of faster female protein evolution. The introgression tests revealed that many of these genes have small introgressed regions, with introgression more pronounced between *S. pallidus* and *S. droebachiensis*.

Among the sea urchin genera studied to date, there is considerable variation in the rate of bindin diversification. Furthermore, when positive selection does occur, the selective pressure responsible for GRP divergence may vary across genera and locale. Variation in sperm availability across genera and species may explain why GRPs in

some genera show positive selection but not in others (Levitan 2002c; Levitan 2002c; Levitan et al. 2007). Under conditions of sperm limitation, purifying selection is expected to favor common GRP alleles with high binding affinity to increase zygote production (Tomaiuolo and Levitan 2010; Levitan et al. 2019). Conditions of high spawning density and strong sperm competition should lead to positive selection, where the risk of polyspermy favors egg protein mutations that reduce compatibility with common sperm protein genotypes, leading to fewer polyspermic zygotes.

Although GRPs sometimes experience rapid diversification, it is unlikely that their divergence commonly causes reproductive isolation. Bindin is not one of the fastest-evolving proteins in sea urchins (Lessios and Zigler 2012; Geyer et al. 2020b), appears to be under positive selection in fewer than half the genera studied thus far, and sometimes evolves under purifying selection (Geyer et al. 2020b). While bindin divergence predicts gamete compatibility, there appears to be a threshold level of bindin divergence necessary for it to act as a true barrier locus, as high gamete compatibility is usually still achieved following several amino acid replacements (Kirk S Zigler et al. 2005). Additionally, gametic incompatibilities are rarely bidirectional (Strathmann 1981; McCartney and Lessios 2004; Kirk S Zigler et al. 2005), and asymmetric incompatibilities are incapable of preventing interspecific gene flow (Addison and Pogson 2009b). Within *Strongylocentrotidae*, gametic incompatibilities have not been an effective barrier to introgression, and the signal of introgression does not appear to be related to bindin or EBR1 divergence (Glasenapp and Pogson 2023). Furthermore, the observed signal of introgression between non-sister taxa at both

bindin and EBR1 is inconsistent with them having a role in reproductive isolation because genes affecting reproductive isolation should not flow readily between species and should reflect species boundaries (Wu 2001; Nosil and Schluter 2011). Finally, studies within *Arbacia* (Metz et al. 1998), *Heliocidaris* (Binks et al. 2012), and *Pseudoboletia* (Zigler et al. 2012) have demonstrated that substantial reproductive isolation can evolve between differentiated groups without bindin divergence. The only system studied thus far where GRPs may have played a role in the evolution of reproductive isolation appears to be the Indo-Pacific *Echinometra* species, where strong reciprocal incompatibilities have formed between recently diverged sympatric species (Palumbi and Metz 1991; Metz et al. 1994; Kirk S Zigler et al. 2005). More work is needed to identify other potential reproductive isolating barriers across diverse groups and quantify their relative strength.

The term "species-specific" is used widely in the literature to describe sea urchin fertilization and the interaction between bindin and EBR1. This idea seems to have come from early studies that compared highly divergent taxa and found that heterospecific mixtures of gametes tend to result in fewer fertilizations than mixtures of homospecific gametes (Summers and Hylander 1975; Glabe and Vacquier 1977). "Species-specific" is misleading because it implies that hybrid crosses either aren't possible or extremely rare. In reality, hybrid crosses in laboratory no-choice experiments are generally successful between congeneric species pairs in at least one direction, and even inter-ordinal crosses can be made between sea urchins and sand dollars (Flickinger 1957; Moore 1957b; Brookbank 1970; Fujisawa 1993). The use of

strongylocentrotid hybrid individuals in aquaculture has been pursued recently as hybrids often have commercially desirable traits such as heat resistance and faster growth rate (Ding et al. 2007; Liu et al. 2020; Zhao et al. 2021). Natural hybridization has also been detected in at least five genera of sea urchins (*Diadema*: Lessios and Pearse (1996), *Lytechinus*: Zigler and Lessios (2004) *Strongylocentrotus*: Levitan (2002), *Pseudoboletia*: Zigler et al. (2012), *Arbacia*: Lessios et al. (2012)), and Levitan (2002) showed that hybrid larvae are readily formed between *S. droebachiensis* and *M. franciscanus* when *S. droebachiensis* females are closer to heterospecific males. For these reasons, sea urchin fertilization should be described as species-selective rather than species-specific (Vacquier et al. 1995).

In summary, we found that bindin and EBR1 have evolved rapidly under positive selection within the strongylocentrotid family of sea urchins. We also unexpectedly found that both genes have experienced historical adaptive introgression, a pattern inconsistent with the hypothesis of speciation via the rapid evolution of gamete recognition proteins. In the future, it will be important to further quantify the amount of bindin and EBR1 polymorphism present within species and further characterize the structural variation observed between species.

3.4 Materials and Methods

3.4.1 Study System

The strongylocentrotid family of sea urchins contains nine broadly distributed species throughout the North Pacific, including the purple sea urchin *S. purpuratus*,

which has a well-annotated reference genome (Spur_5.0) (Sodergren et al. 2006). The species tree relationships are consistent with a West Pacific common ancestor and subsequent colonization of the East Pacific (Kober and Bernardi 2013c). The family contains four genera, two of which are monotypic – *Pseudocentrotus* and *Hemicentrotus*. *Mesocentrotus* contains *M. nudus* and *M. franciscanus* (formerly *S. franciscanus*). *Strongylocentrotus* contains five species – *S. purpuratus*, *S. intermedius*, *S. pallidus*, *S. droebachiensis*, and *S. fragilis* (formerly *Allocentrotus fragilis*). The ranges of four of the strongylocentrotid species are limited to the West Pacific – *P. depressus*, *M. nudus* (formerly *S. nudus*), *H. pulcherrimus,* and *S. intermedius*. The other five taxa have overlapping ranges in the East Pacific – *M. franciscanus*, *S. purpuratus*, *S. pallidus*, *S. droebachiensis*, and *S. fragilis*. Two of the younger *Strongylocentrotus* species, *S. pallidus* and *S. droebachiensis*, are Holarctic, co-occurring in the West Pacific and East Pacific and Atlantic. Sympatric species generally have overlapping depth preferences and spawning times, and previous phylogenomic studies have uncovered introgression between several extant species pairs (Glasenapp and Pogson 2023; Glasenapp and Pogson 2024).

3.4.2 DNA Sequencing and Multiple Sequence Alignments

DNA samples for each of the strongylocentrotid species were collected for sequencing. The *S. pallidus* and *S. droebachiensis* samples were collected near Friday Harbor, WA, as described in Addison and Pogson (2009). *Mesocentrotus franciscanus* was collected near Santa Cruz, California (Addison and Pogson 2009b), and *S. fragilis*

was collected from whale falls in Monterey Bay (Kober and Pogson 2017). The *M. nudus* and *S. intermedius* samples came from the eastern coast of South Korea and were provided by Y-H Lee. The *Hemicentrotus pulcherrimus* and *Pseudocentrotus depressus* samples were supplied by Y. Agatsuma from Shimoda, Izu Peninsula, Shizuoka Prefecture, Japan.

The samples were sequenced to high coverage depth on the Illumina HiSeq 2000 (Kober and Bernardi 2013c; Kober and Pogson 2017). Raw sequence reads are available in the NCBI SRA (BioProject PRJNA391452). Reference genome alignment, variant calling, and multiple sequence alignments were performed by Kober and Pogson (2017). Briefly, paired-end reads were aligned to the *S. purpuratus* reference genome (Spur_3.1) using SSAHA2 v.2.5.5 (Ning et al. 2001). Reads with mapping qualities below 30 and nucleotides with quality scores below 25 were filtered. Variable sites were called heterozygous if the minor allele had a frequency greater than 0.125 and was covered by at least eight filtered reads. A multiple sequence alignment for the sperm protein bindin was constructed from the full-length *S. purpuratus* cDNA sequence of Gao et al. (1986) using the MARSS tool (https://github.com/kordk/marss). Paired-end reads from the last portion of exon four in *bindin* from *M. franciscanus*, *M. nudus*, and *P. depressus* failed to align with the *S. purpuratus* reference sequence. Additional gaps were present in most species near the end of exon five due to the presence of highly divergent protein repeat motifs of variable size and number. Sequences from these regions for each species were obtained by Sanger sequencing using the PCR primers listed in Supplementary Table S36. The multiple sequence

alignment for bindin's egg receptor, EBR1, was constructed from the cDNA sequences of *S. purpuratus* and *M. franciscanus* EBR1 (Kamei and Glabe 2003). We also constructed multiple sequence alignments for other *S. purpuratus* genes involved in reproduction (Song et al. 2006) or showing expression in sperm or egg cells (Tu et al. 2012; Tu et al. 2014). These gene models were obtained from the Spur_3.1 NCBI RefSeq assembly to guide multi-species sequence alignments.

3.4.3 Selection Tests

We tested the multi-species multiple sequence alignments of bindin, EBR1, and the other known sperm/egg genes for selection using both PAML v4.10.7 (Yang 2007) and HyPhy v2.5.32 (Pond et al. 2020). HyPhy was used to complement the PAML analyses as it models synonymous rate variation, allowing d_S to vary across sites and/or branches. To test for pervasive positive selection across the whole phylogeny, we used the codeml program of PAML to compare the codon sites models M7 (Beta) vs. M8 (Beta plus ω) (Yang 2005) and the FUBAR program (Murrell et al. 2013) of HyPhy. To test for episodic selection, we implemented the HyPhy models BUSTED (Murrell et al. 2015), MEME (Murrell et al. 2012), and aBSREL (Smith et al. 2015). BUSTED tests for gene-wide episodic selection, while MEME identifies specific sites under selection on a subset of branches, and aBSREL identifies individual branches with evidence of episodic diversification. For the PAML analyses, gene-wide significance was assessed by calculating the M8 vs. M7 likelihood ratio score ($2\Delta\ell$) and assuming a chi-square distribution with two degrees of freedom (Álvarez-Carretero et al. 2023).

We used a Bayes Empirical Bayes posterior probability threshold of 0.95 as evidence of selection for individual codon sites in genes with significant tests for selection. For the HyPhy analyses, we used a posterior probability threshold of 0.9 for FUBAR and $p < 0.05$ for BUSTED, MEME, and aBSREL, following the recommendations in the HyPhy documentation.

In our multiple sequence alignments of bindin and EBR1, both genes have sizable regions that support a topology other than the overall gene tree topology with strong support. Because specifying the wrong topology can lead to false inferences of selection, we ran the GARD (Kosakovsky Pond et al. 2006) program of HyPhy to screen the multiple sequence alignments for recombination. To determine whether the presence of recombination led to false selection signals, we reran the selection tests using the partitions defined by GARD and compared the results to the selection tests ran on the non-partitioned data. We also split the multiple sequence alignments by exon, inferred individual exon topologies using IQ-TREE2 (Nguyen et al. 2015), and reran the codeml program of PAML. Finally, to characterize reproductive proteins more generally and put our bindin and EBR1 results in context, we tested for selection at 64 additional sperm genes and 47 additional egg genes using the PAML codeml and HyPhy BUSTED programs.

3.4.4 Introgression Tests

The bindin and EBR1 gene tree topologies are discordant with the species tree and show recombination and variation in topologies across the gene. To determine

whether the gene tree discordance present was caused by introgression, we applied the phylogenetic hidden Markov model PhyloNet-HMM (Liu et al. 2014) to our multiple sequence alignments. PhyloNet-HMM detects breakpoints between regions with different underlying phylogenetic relationships and outputs posterior probabilities of introgression for each site in the input multiple sequence alignment. The model employs a finite-sites model to account for convergence and is incomplete lineage sorting-aware. PhyloNet-HMM was previously used to detect introgression between *S. pallidus* and *S. droebachiensis* (Glasenapp and Pogson 2024) and has been applied in other systems, including swordtails (Schumer et al. 2016), house mice (Liu et al. 2015), North American admiral butterflies (Mullen et al. 2020) and snowshoe hares (Jones et al. 2020). Using the default settings, we ran PhyloNet-HMM on the multiple sequence alignments 100 times to avoid local optima and averaged the posterior probabilities across runs, using a mean posterior probability threshold of 90% as evidence of introgression. Introgression tracts were inferred by recording the coordinates of consecutive sites with posterior probabilities above the threshold.

We also used Patterson's *D* (Green et al. 2010; Durand et al. 2011) as a framework to characterize patterns of allele sharing between non-sister taxa at bindin and EBR1. Although Patterson's *D* is not supposed to be applied to individual genes (Martin et al. 2015), we found it to be a useful complementary approach to the formal tests for introgression with PhyloNet-HMM. Patterson's *D* is usually not appropriate for single genes because adjacent site patterns are not independent due to physical linkage, and when applied to whole genomes, *D* outlier loci tend to cluster in low-

divergence regions and are prone to false positives (Martin et al. 2015). However, the EBR1 gene model spans over 70 kb, contains several dozen exons, and shows considerable recombination and divergence in our multiple sequence alignment, making the block-jackknife procedure for assessing significance reasonable. We applied Patterson's *D* to our bindin and EBR1 sequence alignments, using a block size of 1 kb for the EBR1 concatenated exon alignment, 10 kb for the EBR1 alignment containing introns, and 129 base pairs for the bindin concatenated exon alignment using the EvobiR package (Jonika et al. 2023). We further identified ABBA and BABA amino acid sites in our translated sequence alignments.

3.4.5 Bindin and EBR1 Protein Structure and Interaction

For successful fertilization, bindin and EBR1 must adhere to one another following the sperm's contact with the egg jelly. To identify candidate protein regions involved in binding, we modeled the bindin-EBR1 protein-protein interaction using the molecular docking program HDOCK (Yan et al. 2017; Yan et al. 2020). The HDOCK program uses the protein structures for the receptor and ligand and a scoring function to rank all possible binding modes and outputs confidence scores and predicted interface residues for each mode. We ran HDOCK using the bindin and EBR1 protein data bank (pdb) files as input. The *S. purpuratus* bindin pdb file was downloaded from the AlphaFold Protein Structure Database (Accession P06651) (Varadi et al. 2022; Varadi et al. 2024). EBR1 is outside the primary structure length range for proteins in

the AlphaFold database, so we ran AlphaFold using the *S. purpuratus* ebr1 primary structure (NCBI gene ID: 577775).

3.5 Tables and Figures

Table 3.1. Summary of tests for positive selection at bindin and EBR1.

					PAML	HyPhy BUSTED		
Gene	Codons[a]	d_N	d_S	d_N/d_S	2Δℓ M8 vs. M7[b]	LRT[c]	p[d]	PSCs[e]
bindin	420	0.23	0.35	0.67	42.2***	12.0	0.0007	**28**,**42**,**127**,**137**,148, **219**,246,247,**295**,**297**,**299**,309,***396***,*398*
EBR1	2525	0.22	0.53	0.42	123.0***	54.0	9.22E-13	11,**489**,***512***,574,**630**, 642,646,691,**733**,769, **947**,1007,***1067***,**1151**, 1283,***1329***,**1332**,***1360***, 1373,1403,**1469**,1704, 1874,***2039***,**2083**,*2099*, *2124*,**2215**,**2317**,2466, **2476**

[a]Number of codons tested by PAML and HyPhy.

[b]Likelihood ratio test score comparing PAML models M7 and M8.

[c]HyPhy BUSTED likelihood ratio test.

[d]HyPhy BUSTED p value.

[e]The positively selected codon sites (PSCs) from our multiple sequence alignments under *pervasive* positive selection, identified by PAML and HyPhy. A mapping of the codon coordinates in our multiple sequence alignments to the complete *S. purpuratus* NCBI proteins (BND, NCBI Gene ID: 373276; ebr1, NCBI Gene ID: 577775) is available in Supplementary Table S28. Sites colored black were significant in both the HyPhy FUBAR and PAML tests. Sites colored blue were significant in FUBAR only. Sites colored red were significant in PAML only. Sites in **bold** were significant in the HyPhy FUBAR partitioned analysis or PAML exon-by-exon analysis. Italicized sites were declared introgressed between *S. pallidus* and *S. droebachiensis* for bindin, and between *H. pulcherrimus* and *S. droebachiensis* for EBR1.

Table 3.2. Counts of mutations shared between *Hemicentrotus* pulcherrimus and the *Strongylocentrotus* taxa at variable sites. Invariable and singleton sites were removed.

taxon 1	taxon 2	Total Shared Mutations	Private Shared Mutations
Hpul	Spur	160	34
Hpul	Sint	119	13
Hpul	Spal	87	10
Hpul	Sdro	225	81
Hpul	Sfra	91	8

Species abbreviations: *Sdro - S. droebachiensis*; *Sfra - S. fragilis*; *Spal - S. pallidus*; *Sint - S. intermedius*; *Spur - S. purpuratus*; *Hpul - H. pulcherrimus*

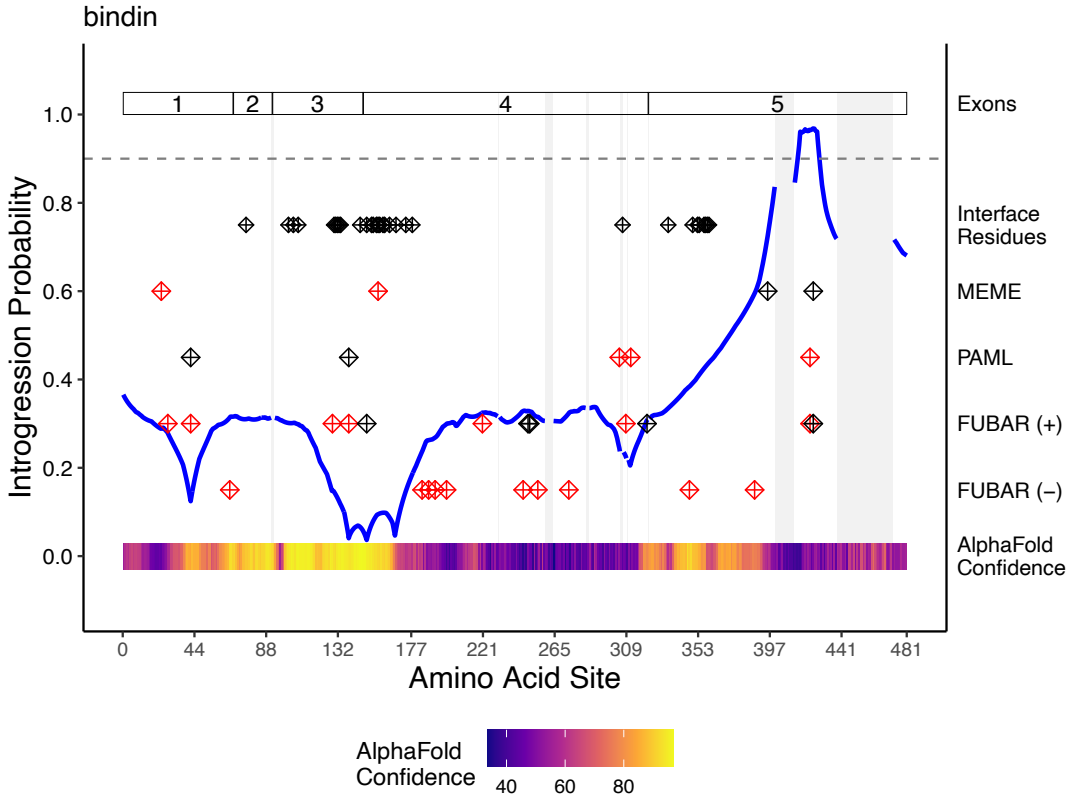


Figure 3.1. Probability of introgression across the protein-coding regions of bindin (blue line). Grey dashed line: 95% posterior probability. The bindin exons are labeled at the top of the plot. The positively selected codon sites identified by MEME, PAML and FUBAR (+) are shown. The location of codon sites under pervasive negative selection identified by FUBAR is shown (FUBAR -). Codon sites that remained significant in the partitioned analysis accounting for recombination are colored in red. The AlphaFold confidence score (plDDT) values for each amino acid residue are plotted at the bottom. Grey bars indicate missing data gaps in the bindin multi-species multiple sequence alignment.

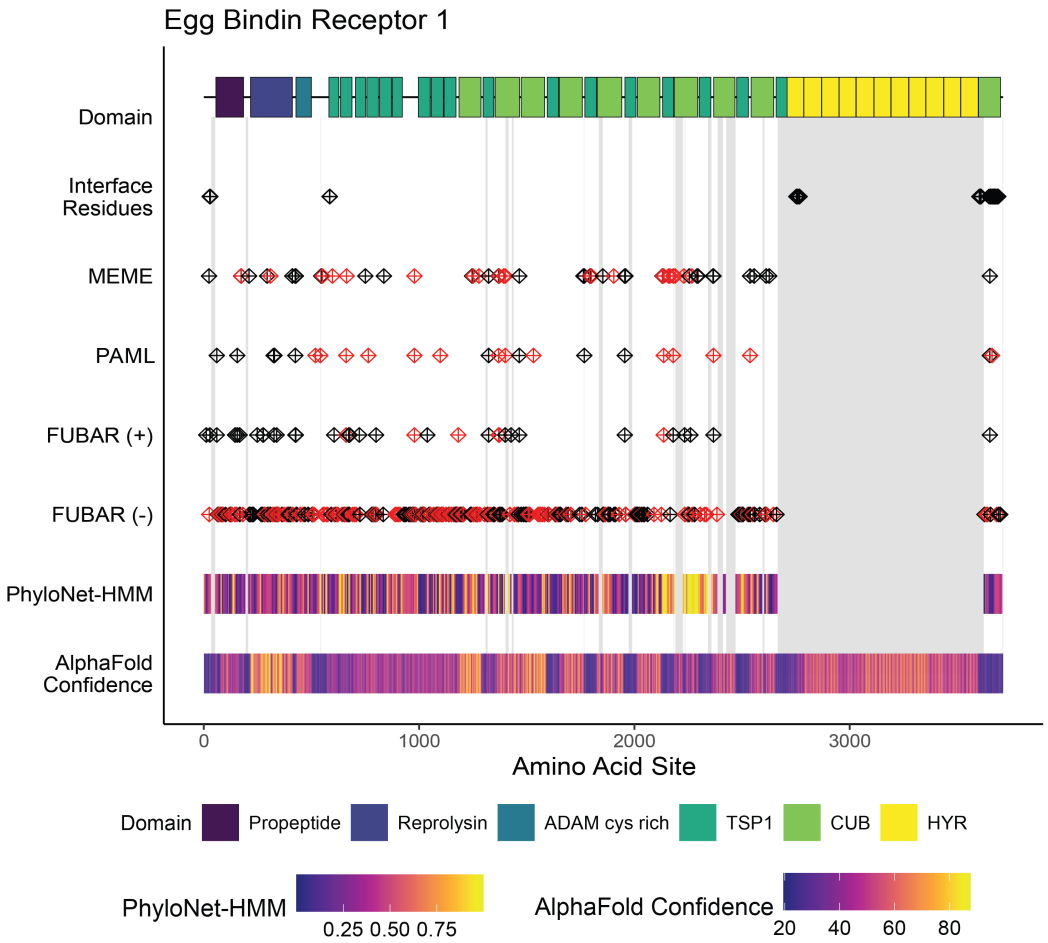


Figure 3.2. Location of the positively and negatively selected codons across the EBR1 protein-coding alignment (MEME, PAML, FUBAR (+), FUBAR (-)). Codon sites that remained significant in the partitioned analysis accounting for recombination are colored in red. The known EBR1 domains are labeled at the top. The PhyloNet-HMM posterior probabilities of introgression between H. pulcherrimus and S. droebachiensis and the AlphaFold confidence scores (plDDT) are shown at the bottom. Grey bars indicate missing data gaps in the EBR1 multi-species multiple sequence alignment.

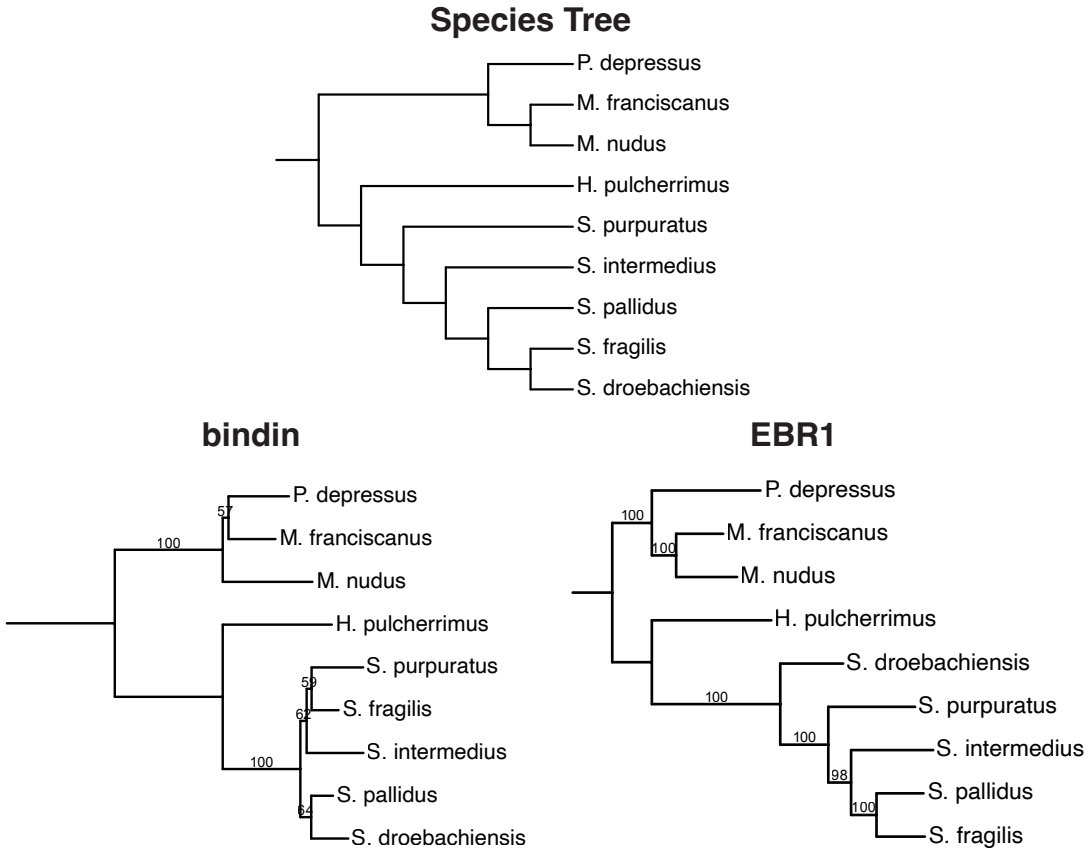


Figure 3.3. Maximum-likelihood gene trees of bindin and EBR1 for the strongylocentrotid sea urchin family. Only the protein-coding regions of bindin and EBR1 were used in tree reconstruction. The species tree relationships are shown at the top.

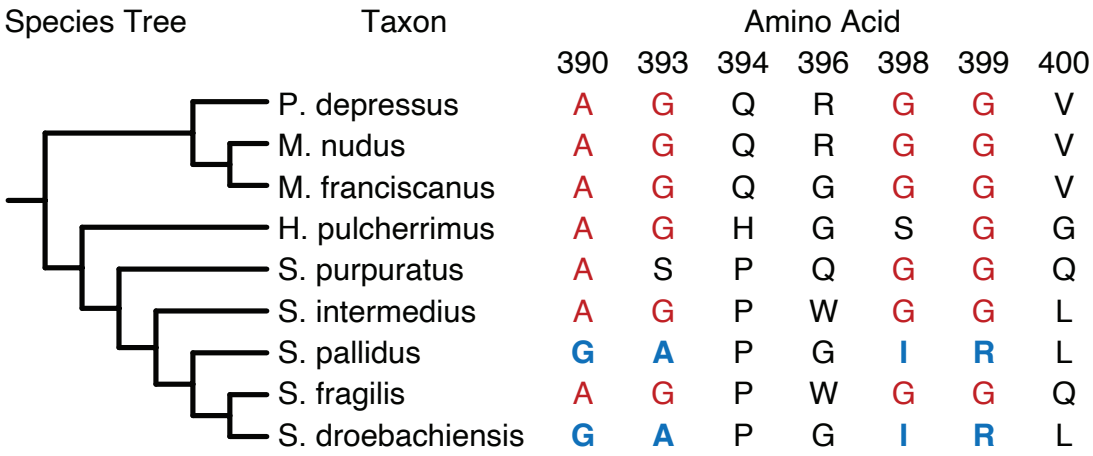


Figure 3.4. The amino acid sequence from the bindin introgression tract, spanning amino acids 389-401 in our multiple sequence alignment. Invariable and singleton sites were removed from the table. Shared, derived amino acids between *S. pallidus* and *S. droebachiensis* are shown in bold blue. Amino acids 389 – 401 in our multiple sequence alignment correspond to amino acids 415 – 427 in the complete gene record for *S. purpuratus* (BND, NCBI Gene ID: 373276). A mapping of the codon coordinates in our multiple sequence alignment to the complete NCBI protein is available in Supplementary Table S28.

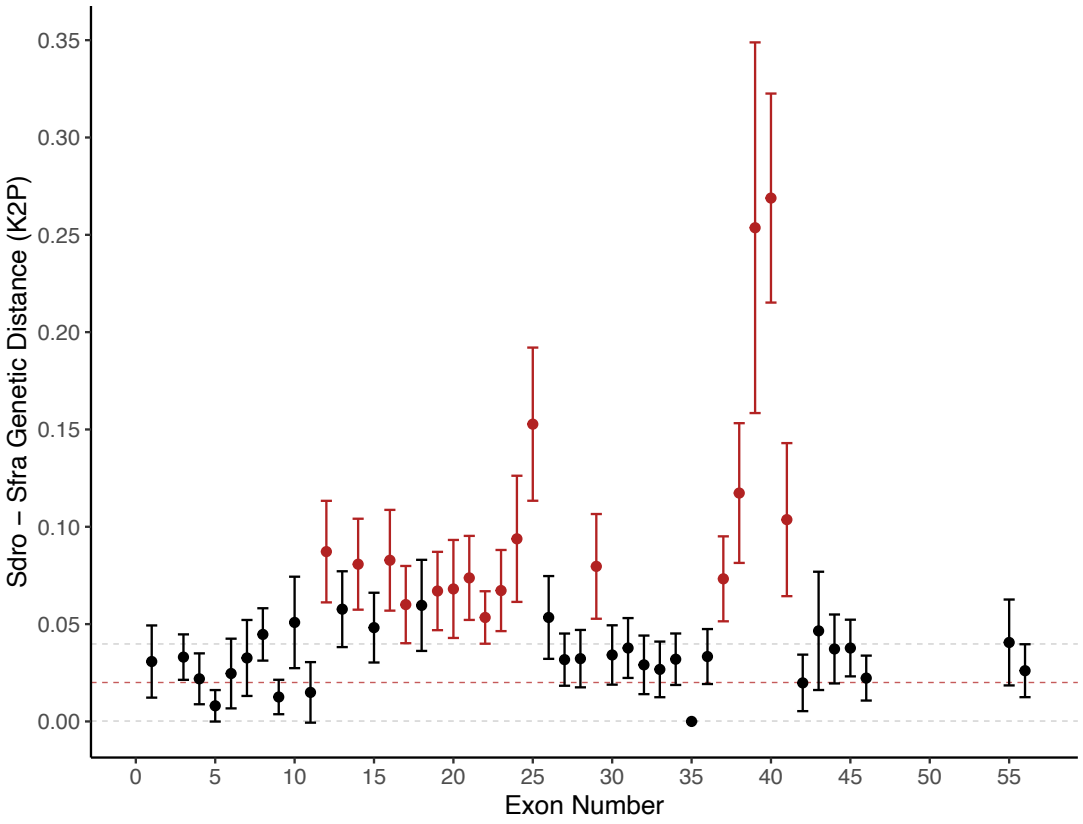


Figure 3.5. Pairwise genetic distance (K2P) between sister taxa *S. droebachiensis* and *S. fragilis* by EBR1 exon. The dashed red line represents the mean pairwise distance between *S. droebachiensis* and *S. fragilis* for 6,520 single-copy orthologs from Kober and Pogson (2017). The dashed grey lines represent the 95% confidence intervals. Exons with K2P distances exceeding the genome-wide 95% confidence intervals are shown in red.

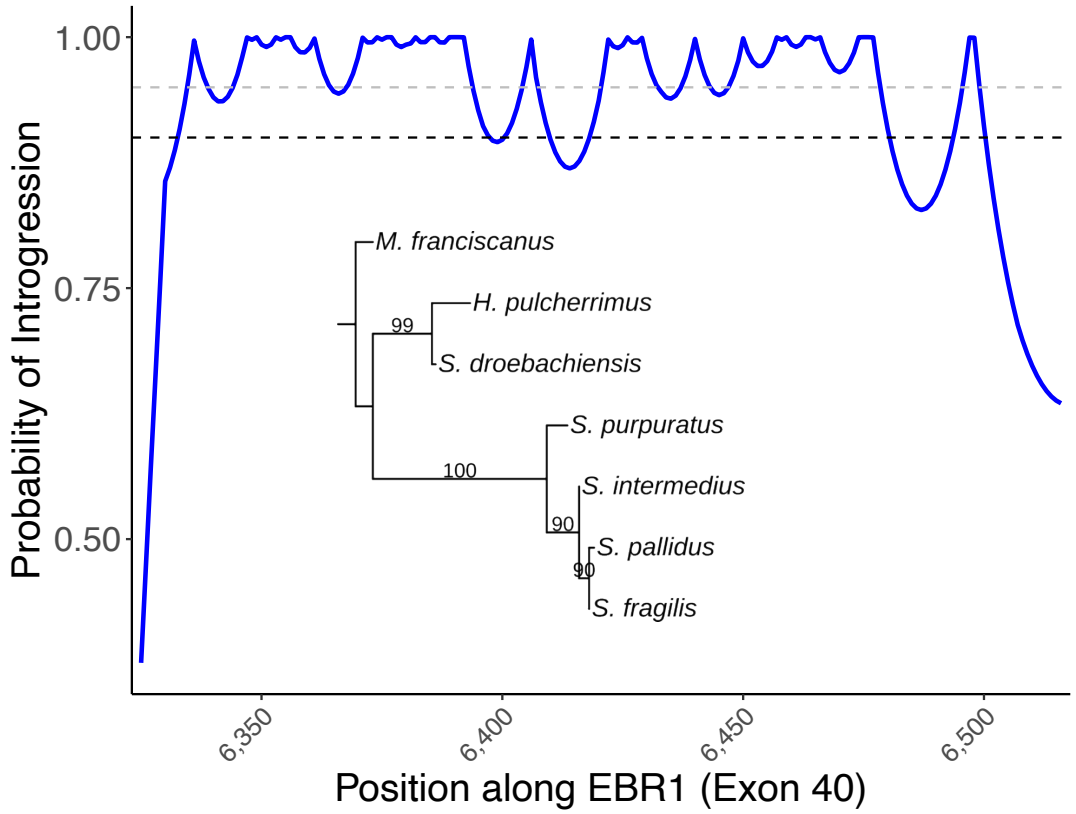


Figure 3.6. Probably of introgression across exon 40 in our EBR1 protein-coding multiple sequence alignment. Grey dashed line: 95% posterior probability; black dashed line: 90% posterior probability. The discordant gene tree for exon 40 (192 base pairs) strongly supports introgression between non-sister taxa *H. pulcherrimus* and *S. droebachiensis*. Exon 40 represents bases 6,325 - 6,516 in our multiple sequence alignment and bases 6,733 – 6,924 in the complete gene record for *S. purpuratus* (ebr1, NCBI Gene ID: 577775). A mapping of the nucleotide coordinates in our multiple sequence alignment to the complete NCBI protein is available in Supplementary S28.

Synthesis

Introgression has long been predicted to be an important source of new adaptive variation for selection to act upon, but this was only recently confirmed by the rapidly accumulating body of empirical studies formally testing for introgression using Next Generation Sequencing data. The recent expansion of genomics resources for non-model systems has uncovered a surprising amount of hybridization and introgression, prompting investigation into their importance in adaptation and speciation. As many of these studies are centered on the genetic basis of adaptive phenotypes already suspected to have introgressed between species, the functional consequence of most introgressed variation remains unknown, and there are few estimates of the proportion of introgressed variation fixed by positive selection. Further studies correlating genome-wide signals of selection and introgression are needed in diverse taxonomic groups to evaluate the relative contributions of mutation and introgression to adaptive evolution, especially in those whose diversification hasn't left clear morphological or ecological signatures. For example, the nine strongylocentrotid sea urchin species diverged within the last twenty million years with only slight changes in pigmentation, depth preference, and spawning time, and although positive selection is widespread, it remains unclear whether adaptive or non-adaptive processes have been more important in their divergence. In this dissertation, I used whole-genome sequencing data and new phylogenomic approaches to evaluate the importance of introgression during the adaptive diversification of the strongylocentrotid sea urchins.

I was surprised to find that introgression was widespread across the strongylocentrotid family, as these species were thought to be strongly reproductively isolated by gametic incompatibilities. Although hybridization might be expected to be more common among external than internal fertilizers, genetic testing for hybrids or for historical or contemporary introgression has only been performed in a few broadcast spawner groups. It's unclear how much hybridization is ongoing among the strongylocentrotid species because the methods employed here test for ancient introgression. However, characterizing introgression remains important even when reproductive isolation is complete because a deeper history of hybridization and introgression can have significant contemporary evolutionary consequences, such as reinforced prezygotic barriers or adaptively introgressed phenotypes. Although there was strong selection against introgression throughout much of the genome, a surprising number of introgressed protein-coding genes showed historical positive selection, indicating that introgression has likely been an important source of adaptive genetic variation.

The evolution of gametic incompatibilities is thought to play an important role in the evolution of reproductive isolation between broadcast-spawning marine invertebrates. However, it remains unclear whether gametic isolation typically develops early in the speciation process or accumulates only after other reproductive barriers are already in place. The high amount of introgression observed here is inconsistent with gametic isolation playing a major role in strongylocentrotid speciation. If gametic incompatibilities were insufficient to limit introgression

following speciation, it is unlikely that they were the primary barrier causing speciation. Furthermore, the introgression signal observed was not correlated with phylogenetic or gamete recognition protein genetic distances, and the proteins responsible for gamete compatibility show adaptive introgression between divergent non-sister taxa, inconsistent with predictions for barrier loci. The continued divergence of the strongylocentrotid urchins in the face of widespread introgression underscores the importance of postzygotic isolation in maintaining species boundaries.

The introgression inference here could only have been made with whole-genome sequencing data and benefitted tremendously from complete taxonomic representation. Observing hybridization and introgression among the strongylocentrotids in the wild is infeasible because they have broad, poorly characterized distributions, and several species are only found at considerable depth. Additionally, hybrid individuals are not easily morphologically identified and may drift far from their parents during their planktonic larval stages. Given the pervasive phylogenetic discordance across the genomes of these taxa, sampling many nuclear loci was necessary to accurately distinguish between introgression and other sources of discordance, such as incomplete lineage sorting and ancestral population structure. Additionally, having data from each species in the family was critical because incomplete taxa sampling would have led to the incorrect attribution of introgression to extant taxa in cases when it actually involved an ancestral lineage (i.e., an internal branch in the phylogeny).

Identifying the specific introgressed genomic regions was difficult due to the old age of introgression. Recombination has fragmented introgressed haplotypes considerably, and subsequent substitutions have eroded the introgression signal, presenting a serious challenge for current local ancestry inference methods. Window-based approaches typically require the user to define a window size in units of variable sites, leading to a tradeoff between resolution and confidence. In species with high amounts of recombination, small ancestry blocks may not contain enough informative sites to provide high confidence in the phylogenetic relationship inferred, while larger windows may average over multiple recombination blocks with distinct phylogenetic histories. Hidden Markov model approaches avoid this problem by attempting to define the recombination blocks but may return false positives or fail to reach statistical significance when recombination blocks have few informative sites. In Chapter 2, the hidden Markov model I applied identified many introgression tracts on the scale of tens to hundreds of bases, but I only included introgression tracts greater than ten kilobases to mitigate the effect of false positives in downstream analyses. As the introgression tracts identified at bindin and EBR1 in Chapter 3 were much shorter than this threshold, I have likely missed a considerable amount of biologically relevant introgression.

Moving forward, there is still much to be done to continue developing the strongylocentrotid urchin family as a model system in evolutionary genomics. Population-level sampling is needed for each species to determine the variation in introgressed ancestry between individuals and populations and to assess the frequency distribution of introgressed alleles. It will also be important to compare the

introgression signal at known regulatory regions to that of protein-coding sequences and to test for reduced introgression near sex-determining loci. A common pattern among introgression studies in other systems is the enhancement barriers to introgression on the sex chromosomes where recombination is limited, and Dobzshansky-Muller incompatibilities are thought to be widespread. Unfortunately, although strongylocentrotid sea urchins have a genetically based sex-determination system, the details and loci involved in sex determination have not been resolved. Finally, a more comprehensive characterization of polymorphism at bindin and EBR1 is needed to better understand the behavior of selection acting within species and whether this selection has cascading effects on gametic compatibility between species.

List of Supplemental Files

1. Appendix A (appendix_a.pdf)
2. Supplementary Tables (supplementary_tables.xlsx)
3. Supplementary Figures (supplementary_figures.docx)

References

Aardema ML, Andolfatto P. 2016. Phylogenetic incongruence and the evolutionary origins of cardenolide-resistant forms of Na+,K+-ATPase in *Danaus* butterflies. *Evolution* 70:1913–1921.

Addison JA, Hart MW. 2005a. Colonization, dispersal, and hybridization influence phylogeography of North Atlantic sea urchins (Strongylocentrotus droebachiensis). *Evol. Int. J. Org. Evol.* 59:532–543.

Addison JA, Hart MW. 2005b. Colonization, dispersal, and hybridization influence phylogeography o North Atlantic sea urchins (*Strongylocentrotus droebachiensis*). *Evolution* 59:532–543.

Addison JA, Kim J. 2022. Trans-Arctic vicariance in *Strongylocentrotus* sea urchins. *PeerJ* 10:e13930.

Addison JA, Pogson GH. 2009a. Multiple gene genealogies reveal asymmetrical hybridization and introgression among strongylocentrotid sea urchins. *Mol. Ecol.* 18:1239–1251.

Addison JA, Pogson GH. 2009b. Multiple gene genealogies reveal asymmetrical hybridization and introgression among strongylocentrotid sea urchins. *Mol. Ecol.* 18:1239–1251.

Adelson DL, Humphreys T. 1988. Sea urchin morphogenesis and cell–hyalin adhesion are perturbed by a monoclonal antibody specific for hyalin. *Development* 104:391–402.

Agatsuma Y, Yamada H, Taniguchi K. 2006. Distribution of the sea urchin *Hemicentrotus pulcherrimus* along a shallow bathymetric gradient in Onagawa Bay in northern Honshu, Japan. *J. Shellfish Res.* 25:1027–1036.

Álvarez-Carretero S, Kapli P, Yang Z. 2023. Beginner's guide on the use of PAML to detect positive selection. *Mol. Biol. Evol.* 40:msad041.

Alves A-P, Mulloy B, Diniz JA, Mourão PAS. 1997. Sulfated polysaccharides from the egg jelly layer are species-specific inducers of acrosomal reaction in sperms of sea urchins. *J. Biol. Chem.* 272:6965–6971.

Anderson E, Hubricht L. 1938. Hybridization in *Tradescantia*. iii. the evidence for introgressive hybridization. *Am. J. Bot.* 25:396–402.

Anderson E, Stebbins GL. 1954. Hybridization as an evolutionary stimulus. *Evolution* 8:378–388.

Arnegard ME, McGee MD, Matthews B, Marchinko KB, Conte GL, Kabir S, Bedford N, Bergek S, Chan YF, Jones FC, et al. 2014. Genetics of ecological divergence during speciation. *Nature* 511:307–311.

Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, Bomblies K, Yant L. 2016. Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci.* 113:8320–8325.

Arnold M, Fogarty N. 2009. Reticulate evolution and marine organisms: the final frontier? *Int. J. Mol. Sci.* 10:3836–3860.

Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *eLife* 6:e24133.

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

Barton NH. 1983. Multilocus clines. *Evolution* 37:454–471.

Biermann CH. 1998a. The molecular evolution of sperm bindin in six species of sea urchins (Echinoida: Strongylocentrotidae). *Mol. Biol. Evol.* 15:1761–1771.

Biermann CH. 1998b. The molecular evolution of sperm bindin in six species of sea urchins (Echinoida: *Strongylocentrotidae*). *Mol. Biol. Evol.* 15:1761–1771.

Biermann CH, Marks JA, Vilela-Silva A-CES, Castro MO, Mourao PAS. 2004. Carbohydrate-based species recognition in sea urchin fertilization: another avenue for speciation? *Evol. Dev.* 6:353–361.

Bierne N, Bonhomme F, David P. 2003. Habitat preference and the marine-speciation paradox. *Proc. R. Soc. Lond. B Biol. Sci.* 270:1399–1406.

Binks RM, Prince J, Evans JP, Kennington WJ. 2012. More than bindin divergence: reproductive isolation between sympatric subspecies of a sea urchin by asynchronous spawning: reproductive barriers between sea urchin subspecies. *Evolution* 66:3545–3557.

Blackman BK. 2016. Speciation genes. In: Encyclopedia of evolutionary biology. Elsevier. p. 166–175. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128000496000664

Boveri T. 1902. Über mehrpolige mitosen als mittel zur analyse des zellkerns. *Verhandlungen Phys.-Med. Ges. Zu Würzbg.* 35:67–90.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. 2014. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLOS Genet.* 10:e1004410.

Brennan RS, Garrett AD, Huber KE, Hargarten H, Pespeni MH. 2019. Rare genetic variation and balanced polymorphisms are important for survival in global change conditions. *Proc. R. Soc. B Biol. Sci.* 286:20190943.

Broad Institute. 2018. Picard tools - by broad institute. Available from: http://broadinstitute.github.io/picard/

Brookbank JW. 1970. DNA synthesis and development in reciprocal interordinal hybrids of a sea urchin and a sand dollar. *Dev. Biol.* 21:29–47.

Bukowicki M, Franssen SU, Schlötterer C. 2016. High rates of phasing errors in highly polymorphic species with low levels of linkage disequilibrium. *Mol. Ecol. Resour.* 16:874–882.

Caetano-Anolles D. 2023. (How to) Filter variants either with VQSR or by hard-filtering. *GATK* [Internet]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering

Calderón I, Turon X, Lessios HA. 2009. Characterization of the sperm molecule bindin in the sea urchin genus *Paracentrotus*. *J. Mol. Evol.* 68:366–376.

Calfee E, Gates D, Lorant A, Perkins MT, Coop G, Ross-Ibarra J. 2021. Selective sorting of ancestral introgression in maize and teosinte along an elevational cline. *PLOS Genet.* 17:e1009810.

Campbell CR, Poelstra JW, Yoder AD. 2018. What is speciation genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biol. J. Linn. Soc.* 124:561–583.

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65:997–1008.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319:1395–1398.

Cooper BS, Sedghifar A, Nash WT, Comeault AA, Matute DR. 2018. A maladaptive combination of traits contributes to the maintenance of a *Drosophila* hybrid zone. *Curr. Biol.* 28:2940-2947.e6.

Corbett-Detig R, Nielsen R. 2017. A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLOS Genet.* 13:e1006529.

Coyne JA, Orr HA. 2004. Speciation. Oxford, New York: Oxford University Press

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of samtools and bcftools. *GigaScience* 10:giab008.

Ding J, Chang Y, Wang C, Cao X. 2007. Evaluation of the growth and heterosis of hybrids among three commercially important sea urchins in China: *Strongylocentrotus nudus*, *S. intermedius* and *Anthocidaris crassispina*. *Aquaculture* 272:273–280.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28:2239–2252.

Durham JW, MacNeil FS. 1967. Cenozoic migrations of marine invertebrates through the Bering Strait region. In: The Bering Land Bridge. Stanford University Press. p. 326–349. Available from: https://www.vliz.be/en/imis

Elmer KR. 2019. Barrier loci and evolution. In: Encyclopedia of life sciences. John Wiley & Sons, Ltd. p. 1–7. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0028138

Evans JP, Sherman CDH. 2013. Sexual selection and the evolution of egg-sperm interactions in broadcast-spawning invertebrates. *Biol. Bull.* 224:166–183.

Evans T, Rosenthal ET, Youngblom J, Distel D, Hunt T. 1983. Cyclin: A protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* 33:389–396.

Feng C, Wang J, Liston A, Kang M. 2023. Recombination variation shapes phylogeny and introgression in wild diploid strawberries. *Mol. Biol. Evol.* 40:msad049.

Flickinger RA. 1957. Evidence from sea urchin-sand dollar hybrid embryos for a nuclear control of alkaline phosphatase activity. *Biol. Bull.* 112:21–27.

Foltz KR. 1994. The sea urchin egg receptor for sperm. *Semin. Dev. Biol.* 5:243–253.

Forsythe ES, Sloan DB, Beilstein MA. 2020. Divergence-based introgression polarization. *Genome Biol. Evol.* 12:463–478.

Fraïsse C, Belkhir K, Welch JJ, Bierne N. 2016. Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. *Mol. Ecol.* 25:269–286.

Franke E, C. Babcock R, A. Styan C. 2002. Sexual conflict and polyspermy under sperm-limited conditions: in situ evidence from field simulations with the free-spawning marine echinoid *Evechinus chloroticus*. *Am. Nat.* [Internet]. Available from: https://www.journals.uchicago.edu/doi/10.1086/342075

Fujisawa H. 1993. Temperature sensitivity of a hybrid between two species of sea urchin differing in thermotolerance. *Dev. Growth Differ.* 35:395–401.

Gao B, Klein LE, Britten RJ, Davidson EH. 1986. Sequence of mRNA coding for bindin, a species-specific sea urchin sperm protein required for fertilization. *Proc. Natl. Acad. Sci.* 83:8634–8638.

Gardner JPA. 1997. Hybridization in the sea. In: Advances in Marine Biology. Vol. 31. Elsevier. p. 1–78. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065288108602217

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. *Nature* 403:886–889.

Gavrilets S, Hayashi TI. 2005. Speciation and sexual conflict. *Evol. Ecol.* 19:167–198.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. *Proc. Natl. Acad. Sci.* 99:10533–10538.

Geyer LB, Lessios HA. 2009. Lack of character displacement in the male recognition molecule, bindin, in Altantic sea urchins of the genus *Echinometra*. *Mol. Biol. Evol.* 26:2135–2146.

Geyer LB, Palumbi SR. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. *Evolution* 57:1049–1060.

Geyer LB, Zigler KS, Tiozzo S, Lessios HA. 2020a. Slow evolution under purifying selection in the gamete recognition protein bindin of the sea urchin *Diadema*. *Sci. Rep.* 10:9834.

Geyer LB, Zigler KS, Tiozzo S, Lessios HA. 2020b. Slow evolution under purifying selection in the gamete recognition protein bindin of the sea urchin *Diadema*. *Sci. Rep.* 10:9834.

Glabe CG, Vacquier VD. 1977a. Species specific agglutination of eggs by bindin isolated from sea urchin sperm. *Nature* 267:836–838.

Glabe CG, Vacquier VD. 1977b. Species specific agglutination of eggs by bindin isolated from sea urchin sperm. *Nature* 267:836–838.

Glasenapp MR, Pogson GH. 2023. Extensive introgression among strongylocentrotid sea urchins revealed by phylogenomics. *Ecol. Evol.* 13:e10446.

Glasenapp MR, Pogson GH. 2024. Selection shapes the genomic landscape of introgressed ancestry in a pair of sympatric sea urchin species. :2023.12.01.566927. Available from: https://www.biorxiv.org/content/10.1101/2023.12.01.566927v2

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328:710–722.

Guerrero A, Nishigaki T, Carneiro J, Yoshiro Tatsu, Wood CD, Darszon A. 2010. Tuning sperm chemotaxis by calcium burst timing. *Dev. Biol.* 344:52–65.

Hagström BE, Lönning S. 1967. Experimental studies of Strongylocentrotus droebachiensis and S. pallidus. *Sarsia* 29:165–176.

Hahn MW. 2018. Molecular population genetics. Oxford, New York: Oxford University Press

Hamlin JAP, Hibbins MS, Moyle LC. 2020. Assessing biological factors affecting postspeciation introgression. *Evol. Lett.* 4:137–154.

Harper FM, Addison JA, Hart MW. 2007a. Introgression Versus Immigration in Hybridizing High-Dispersal Echinoderms. *Evolution* 61:2410–2418.

Harper FM, Addison JA, Hart MW. 2007b. Introgression versus immigration in hybridizing high-dispersal echinoderms. *Evolution* 61:2410–2418.

Harper FM, Hart MW. 2007. Morphological and phylogenetic evidence for hybridization and introgression in a sea star secondary contact zone: hybridization between *Asterias* sea stars. *Invertebr. Biol.* 126:373–384.

Harris K, Nielsen R. 2016. The genetic cost of Neanderthal introgression. *Genetics* 203:881–891.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22:4606–4618.

Hertwig O. 1876. Beiträge zur kenntniss der bildung, befruchtung und theilung des thierischen eies ... W. Engelmann

Hibbins MS, Hahn MW. 2022. Phylogenomic approaches to detecting and characterizing introgression.Turelli M, editor. *Genetics* 220:iyab173.

Hirase S, Yamasaki YY, Sekino M, Nishisako M, Ikeda M, Hara M, Merilä J, Kikuchi K. 2021. Genomic evidence for speciation with gene flow in broadcast spawning marine invertebrates.Crandall K, editor. *Mol. Biol. Evol.* 38:4683–4699.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. Ufboot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35:518–522.

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512:194–197.

Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner PA, Waterman M, editors. Research in Computational Molecular Biology. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. p. 233–249.

Janoušek V, Munclinger P, Wang L, Teeter KC, Tucker PK. 2015. Functional organization of the genome may shape the species boundary in the house mouse. *Mol. Biol. Evol.* 32:1208–1220.

Jones MR, Mills LS, Jensen JD, Good JM. 2020. The origin and spread of locally adaptive seasonal camouflage in snowshoe hares. 196:271–389.

Jonika MM, Chin M, Anderson NW, Adams RH, Demuth JP, Blackmon H. 2023. coleoguy/evobir: EvobiR version 2.1. Available from: https://zenodo.org/records/8033504

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

Juric I, Aeschbacher S, Coop G. 2016. The strength of selection against Neanderthal introgression. *PLOS Genet.* 12:e1006340.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14:587–589.

Kamei N, Glabe CG. 2003. The species-specific egg receptor for sea urchin sperm adhesion is EBR1, a novel ADAMTS protein. *Genes Dev.* 17:2502–2507.

Kim BY, Huber CD, Lohmueller KE. 2018. Deleterious variation shapes the genomic landscape of introgression. *PLOS Genet.* 14:e1007741.

Knowlton N, Maté JL, Guzmán HM, Rowan R, Jara J. 1997. Direct evidence for reproductive isolation among the three species of the Montastraea annularis complex in Central America (Panamá and Honduras). *Mar. Biol.* 127:705–711.

Kober KM, Bernardi G. 2013a. Phylogenomics of strongylocentrotid sea urchins. *BMC Evol. Biol.* 13:88.

Kober KM, Bernardi G. 2013b. Phylogenomics of strongylocentrotid sea urchins. *BMC Evol. Biol.* 13:88.

Kober KM, Bernardi G. 2013c. Phylogenomics of strongylocentrotid sea urchins. *BMC Evol. Biol.* 13:88.

Kober KM, Pogson GH. 2013. Genome-wide patterns of codon bias are shaped by natural selection in the purple sea urchin, *Strongylocentrotus purpuratus*. *G3 GenesGenomesGenetics* 3:1069–1083.

Kober KM, Pogson GH. 2017. Genome-wide signals of positive selection in strongylocentrotid sea urchins. *BMC Genomics* 18:555.

Kong S, Kubatko LS. 2021. Comparative performance of popular methods for hybrid detection using genomic data.Hahn M, editor. *Syst. Biol.* 70:891–907.

Korunes KL, Samuk K. 2021. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* 21:1359–1368.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* 23:1891–1901.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518:371–375.

Lee YH, Ota T, Vacquier V. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12:231–238.

Lee YH, Vacquier VD. 1992. The divergence of species-specific abalone sperm lysins is promoted by positive darwinian selection. *Biol. Bull.* 182:97–104.

Lessios HA. 2007. Reproductive isolation between species of sea urchins. *Bull. Mar. Sci.* 81:191–208.

Lessios HA. 2011. Speciation Genes in Free-Spawning Marine Invertebrates. *Integr. Comp. Biol.* 51:456–465.

Lessios HA, Lockhart S, Collin R, Sotil G, Sanchez-Jerez P, Zigler KS, Perez AF, Garrido MJ, Geyer LB, Bernardi G, et al. 2012. Phylogeography and bindin evolution in *Arbacia*, a sea urchin genus with an unusual distribution. *Mol. Ecol.* 21:130–144.

Lessios HA, Pearse JS. 1996. Hybridization and introgression between Indo-Pacific species of *Diadema*. *Mar. Biol.* 126:715–723.

Lessios HA, Zigler KS. 2012. Rates of sea urchin bindin evolution. In: Rapidly evolving genes and genetic systems. OUP Oxford.

Levitan DR. 1993. The importance of sperm limitation to the evolution of egg size in marine invertebrates. *Am. Nat.* 141:517–536.

Levitan DR. 2002a. The relationship between conspecific fertilization success and reproductive isolation among three congeneric sea urchins. *Evolution* 56:1599–1689.

Levitan DR. 2002b. Density-dependent selection on gamete traits in three congeneric sea urchins. *Ecology* 83:464–479.

Levitan DR. 2002c. The relationship between conspecific fertilization success and reproductive isolation among three congeneric sea urchins. *Evolution* 56:1599–1689.

Levitan DR, Buchwalter R, Hao Y. 2019. The evolution of gametic compatibility and compatibility groups in the sea urchin *Mesocentrotus franciscanus*: an avenue for speciation in the sea. *Evolution* 73:1428–1442.

Levitan DR, Ferrell DL. 2006. Selection on gamete recognition proteins depends on sex, density, and genotype frequency. *Science* 312:267–269.

Levitan DR, Fukami H, Jara J, Kline D, McGovern TM, McGhee KE, Swanson CA, Knowlton N. 2004. Mechanisms of reproductive isolation among sympatric

broadcast-spawning corals of the Montastraea annularis species complex. *Evolution* 58:308–323.

Levitan DR, Stapper AP. 2010. Simultaneous positive and negative frequency-dependent selection on sperm bindin, a gamete recognition protein in the sea urchin *Strongylocentrotus purpuratus*. *Evolution* 64:785–797.

Levitan DR, TerHorst CP, Fogarty ND. 2007. The risk of polyspermy in three congeneric sea urchins and its implications for gametic incompatibility and reproductive isolation. *Evolution* 61:2007–2014.

Liu KJ, Dai J, Truong K, Song Y, Kohn MH, Nakhleh L. 2014. An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Comput. Biol.* 10:e1003649.

Liu KJ, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *Proc. Natl. Acad. Sci.* 112:196–201.

Liu L, Sun J, Zhan Y, Zhao T, Zou Y, Yan H, Zhang W, Chang Y. 2020. Gonadal traits and nutrient compositions of novel sea urchin hybrids of *Hemicentrotus pulcherrimus* (♀) and *Strongylocentrotus intermedius* (♂). *Aquac. Rep.* 18:100439.

Mah SA, Swanson WJ, Vacquier VD. 2005. Positive selection in the carbohydrate recognition domains of sea urchin sperm receptor for egg jelly (suREJ) proteins. *Mol. Biol. Evol.* 22:533–541.

Maheshwari S, Barbash DA. 2011. The Genetics of hybrid incompatibilities. *Annu. Rev. Genet.* 45:331–355.

Malinsky M, Matschiner M, Svardal H. 2021. Dsuite - fast d-statistics and related admixture evidence from vcf files. *Mol. Ecol. Resour.* 21:584–595.

Mallet J. 2005. Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20:229–237.

Mao Y, Economo EP, Satoh N. 2018. The roles of introgression and climate change in the rise to dominance of Acropora corals. *Curr. Biol.* 28:3373-3382.e5.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32:244–257.

Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLOS Biol.* 17:e2006288.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* 47:69–74.

Maxwell CS, Sepulveda VE, Turissini DA, Goldman WE, Matute DR. 2018. Recent admixture between species of the fungal pathogen *Histoplasma*. *Evol. Lett.* 2:210–220.

McBride CS, Singer MC. 2010. Field studies reveal strong postmating isolation between ecologically divergent butterfly populations. *PLOS Biol.* 8:e1000529.

McCartney MA, Lessios HA. 2004. Adaptive evolution of sperm bindin tracks egg incompatibility in neotropical sea urchins of the genus *Echinometra*. *Mol. Biol. Evol.* 21:732–745.

McClay DR, Fink RD. 1982. Sea urchin hyalin: appearance and function in development. *Dev. Biol.* 92:285–293.

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci.* 100:9894–9899.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.* 75:35–45.

Mengerink KJ, Moy GW, Vacquier VD. 2002. suREJ3, a polycystin-1 protein, Is cleaved at the GPS domain and localizes to the acrosomal region of sea urchin sperm. *J. Biol. Chem.* 277:943–948.

Metz EC, Gomez-Gutierrez G, Vacquier VD. 1998. Mitochondrial DNA and bindin gene sequence evolution among allopatric species of the sea urchin genus *Arbacia*. *Mol. Biol. Evol.* 15:185–195.

Metz EC, Kane RE, Yanagimachi H, Palumbi SR. 1994. Fertilization between closely related sea urchins is blocked by incompatibilities during sperm-egg attachment and early stages of fusion. *Biol. Bull.* 187:23–34.

Metz EC, Palumbi SR. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13:397–406.

Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. 2019. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* 14:703–721.

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets.Rosenberg M, editor. *Mol. Biol. Evol.* 37:2727–2733.

Minor JE, Fromson DR, Britten RJ, Davidson EH. 1991. Comparison of the bindin proteins of Strongylocentrotus franciscanus, S. purpuratus, and Lytechinus variegatus: sequences involved in the species specificity of fertilization. *Mol. Biol. Evol.* 8:781–795.

Mo YK, Lanfear R, Hahn MW, Minh BQ. 2023. Updated site concordance factors minimize effects of homoplasy and taxon sampling.Schwartz R, editor. *Bioinformatics* 39:btac741.

Moore AR. 1957a. Biparental inheritance in an interordinal cross of sea urchin and sand dollar. *J. Exp. Zool.* 135:75–83.

Moore AR. 1957b. Biparental inheritance in an interordinal cross of sea urchin and sand dollar. *J. Exp. Zool.* 135:75–83.

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. 2021. The genomic consequences of hybridization. *eLife* 10:e69016.

Mullen SP, VanKuren NW, Zhang W, Nallu S, Kristiansen EB, Wuyun Q, Liu K, Hill RI, Briscoe AD, Kronforst MR. 2020. Disentangling population history and character evolution among hybridizing lineages. *Mol. Biol. Evol.* 37:1295–1305.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: A fast, unconstrained Bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30:1196–1205.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, et al. 2015. Gene-wide identification of episodic selection. *Mol. Biol. Evol.* 32:1365–1371.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. *PLOS Genet.* 8:e1002764.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310:321–324.

Nei M, Nozawa M. 2011. Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biol. Evol.* 3:812–829.

Newman HH. 1923. Hybrid vigor, hybrid weakness, and the chromosome theory of heredity. an experimental analysis of the physiology of heredity in the reciprocal crosses between two closely associated species of sea-urchins, Strongylocentrotus purpuratus and S. franciscanus. *J. Exp. Zool.* 37:169–205.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32:268–274.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res.* 11:1725–1729.

Noor MAF, Feder JL. 2006. Speciation genetics: evolving approaches. *Nat. Rev. Genet.* 7:851–861.

Nosil P. 2012. Ecological speciation. OUP Oxford

Nosil P, Schluter D. 2011. The genes underlying the process of speciation. *Trends Ecol. Evol.* 26:160–167.

Nydam ML, Harrison RG. 2011. Introgression despite substantial divergence in a broadcast spawning marine invertebrate: introgression and divergence in Ciona intestinalis. *Evolution* 65:429–442.

Nydam ML, Yanckello LM, Bialik SB, Giesbrecht KB, Nation GK, Peak JL. 2017. Introgression in two species of broadcast spawning marine invertebrate. *Biol. J. Linn. Soc.* 120:879–890.

Orr HA. 1995. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* 139:1805–1813.

Orr HA. 2005. The genetic basis of reproductive isolation: insights from Drosophila. *Proc. Natl. Acad. Sci.* 102:6522–6526.

Ortiz EM. 2019. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. Available from: https://zenodo.org/record/2540861

Palumbi SR. 1992. Marine speciation on a small planet. *Trends Ecol. Evol.* 7:114–118.

Palumbi SR. 1994. Genetic Divergence, Reproductive Isolation, and Marine Speciation. *Annu. Rev. Ecol. Syst.* 25:547–572.

Palumbi SR. 1999. All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins. *Proc. Natl. Acad. Sci.* 96:12632–12637.

Palumbi S. R. 2009. Speciation and the evolution of gamete recognition genes: pattern and process. *Heredity* 102:66–76.

Palumbi S R. 2009. Speciation and the evolution of gamete recognition genes: pattern and process. *Heredity* 102:66–76.

Palumbi SR, Kessing BD. 1991. Population biology of the trans-arctic exchange: mtDNA sequence similarity between Pacific and Atlantic sea urchins. *Evolution* 45:1790–1805.

Palumbi SR, Lessios HA. 2005. Evolutionary animation: How do molecular phylogenies compare to Mayr's reconstruction of speciation patterns in the sea? *Proc. Natl. Acad. Sci.* 102:6566–6572.

Palumbi SR, Metz EC. 1991. Strong reproductive isolation between closely related tropical sea urchins (genus *Echinometra*). *Mol. Biol. Evol.* 8:227–239.

Palumbi SR, Wilson AC. 1990. Mitochondrial Dna Diversity in the Sea Urchins Strongylocentrotus Purpuratus and S. Droebachiensis. *Evolution* 44:403–415.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 34:867–868.

Pennington JT. 1985. The ecology of fertilization of echinoid eggs: the consequences of sperm dilution, adult aggregation, and synchronous spawning. *Biol. Bull.* 169:417–430.

Petr M, Pääbo S, Kelso J, Vernot B. 2019. Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci.* 116:1639–1644.

Pieplow CA, Furze AR, Wessel GM. 2023. A case of hermaphroditism in the gonochoristic sea urchin, *Strongylocentrotus purpuratus*, reveals key mechanisms of sex determination. *Biol. Reprod.* 108:960–973.

Pond SLK, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A, et al. 2020. HyPhy 2.5 - a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.* 37:295–299.

Pool JE. 2015. The Mosaic ancestry of the *Drosophila* genetic reference panel and the *D. melanogaster* reference genome reveals a network of epistatic fitness interactions. *Mol. Biol. Evol.* 32:3236–3251.

Popovic I, Bierne N, Gaiti F, Tanurdžić M, Riginos C. 2021. Pre-introduction introgression contributes to parallel differentiation and contrasting hybridization outcomes between invasive and native marine mussels. *J. Evol. Biol.* 34:175–192.

Powell DL, García-Olazábal M, Keegan M, Reilly P, Du K, Díaz-Loyo AP, Banerjee S, Blakkan D, Reich D, Andolfatto P, et al. 2020. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science* 368:731–736.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* 11:175–180.

Pujolar JM, Pogson GH. 2011. Positive darwinian selection in gamete recognition proteins of *Strongylocentrotus* sea urchins. *Mol. Ecol.* 20:4968–4982.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Ramírez-Gómez HV, Jimenez Sabinina V, Velázquez Pérez M, Beltran C, Carneiro J, Wood CD, Tuval I, Darszon A, Guerrero A. 2020. Sperm chemotaxis is driven by the slope of the chemoattractant concentration field. *eLife* 9:e50532.

Ravinet M, Faria R, Butlin RK, Galindo J, Bierne N, Rafajlović M, Noor MAF, Mehlig B, Westram AM. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* 30:1450–1477.

Ravinet M, Kume M, Ishikawa A, Kitano J. 2021. Patterns of genomic divergence and introgression between Japanese stickleback species with overlapping breeding habitats. *J. Evol. Biol.* 34:114–127.

Ravinet M, Yoshida K, Shigenobu S, Toyoda A, Fujiyama A, Kitano J. 2018. The genomic landscape at a late stage of stickleback speciation: high genomic divergence interspersed by small localized regions of introgression. *PLOS Genet.* 14:e1007358.

Riffell JA, Krug PJ, Zimmer RK. 2004. The ecological and evolutionary consequences of sperm chemoattraction. *Proc. Natl. Acad. Sci.* 101:4501–4506.

Saarman NP, Pogson GH. 2015. Introgression between invasive and native blue mussels (genus Mytilus) in the central California hybrid zone. *Mol. Ecol.* 24:4723–4738.

Sanchez-Ramirez S. 2017. vcf2fasta. Available from: https://github.com/santiagosnchez/vcf2fasta

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507:354–357.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* 26:1241–1247.

Schluter D, Rieseberg LH. 2022. Three problems in the genetics of speciation by selection. *Proc. Natl. Acad. Sci.* 119:e2122153119.

Schumer M, Cui R, Powell DL, Rosenthal GG, Andolfatto P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. *Mol. Ecol.* 25:2661–2679.

Simon A, Fraïsse C, El Ayari T, Liautard-Haag C, Strelkov P, Welch JJ, Bierne N. 2021. How do species barriers decay? Concordance and local introgression in mosaic hybrid zones of mussels. *J. Evol. Biol.* 34:208–223.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32:1342–1353.

Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, Angerer RC, Angerer LM, Arnone MI, Burgess DR, Burke RD, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* 314:941–952.

Song JL, Wong JL, Wessel GM. 2006. Oogenesis: single cell development and differentiation. *Dev. Biol.* 300:385–405.

Song Y, Endepols S, Klemann N, Richter D, Matuschka F-R, Shih C-H, Nachman MW, Kohn MH. 2011. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between Old World mice. *Curr. Biol.* 21:1296–1301.

Stapper AP, Beerli P, Levitan DR. 2015. Assortative mating drives linkage disequilibrium between sperm and egg recognition protein loci in the sea urchin *Strongylocentrotus purpuratus*. *Mol. Biol. Evol.* 32:859–870.

Strathmann RR. 1981. On barriers to hybridization between *Strongylocentrotus droebachiensis* (O.F. Müller) and *S. pallidus* (G.O. Sars). *J. Exp. Mar. Biol. Ecol.* 55:39–47.

Summers RG, Hylander BL. 1975. Species-specificity of acrosome reaction and primary gamete binding in echinoids. *Exp. Cell Res.* 96:63–68.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell PJ, Lang M, Courtier-Orgogozo V, et al. 2022. Widespread introgression across a phylogeny of 155 Drosophila genomes. *Curr. Biol.* 32:111-123.e5.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc. Natl. Acad. Sci.* 98:7375–7379.

Swanson WJ, Vacquier VD. 2002a. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* 3:137–144.

Swanson WJ, Vacquier VD. 2002b. Reproductive protein evolution. *Annu. Rev. Ecol. Syst.* 33:161–179.

Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, Krenz JG, Sans-Fuentes MA, Nachman MW, Tucker PK. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res.* 18:67–76.

Than C, Ruths D, Nakhleh L. 2008. Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322.

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487:94–98.

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* 31:8–22.

Tomaiuolo M, Levitan DR. 2010. Modeling how reproductive ecology can drive protein diversification and result in linkage disequilibrium between sperm and egg proteins. *Am. Nat.* 176:14–25.

Tu Q, Cameron RA, Davidson EH. 2014. Quantitative developmental transcriptomes of the sea urchin *Strongylocentrotus purpuratus*. *Dev. Biol.* 385:160–167.

Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson EH. 2012. Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res.* 22:2079–2087.

Turner LM, Hoekstra HE. 2008. Causes and consequences of the evolution of reproductive proteins. *Int. J. Dev. Biol.* 52:769–780.

Vacquier VD, Moy GW. 1977. Isolation of bindin: the protein responsible for adhesion of sperm to sea urchin eggs. *Proc. Natl. Acad. Sci.* 74:2456–2460.

Vacquier VD, Swanson WJ. 2011. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. *Cold Spring Harb. Perspect. Biol.* 3:a002931.

Vacquier VD, Swanson WJ, Hellberg ME. 1995. What have we learned about sea urchin sperm bindin? *Dev. Growth Differ.* 37:1–10.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43:11.10.1-11.10.33.

Van Doorn GS, Luttikhuizen PC, Weissing FJ. 2001. Sexual selection at the protein level drives the extraordinary divergence of sex–related genes during sympatric speciation. *Proc. R. Soc. Lond. B Biol. Sci.* 268:2155–2161.

Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ, et al. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression.Jiggins CD, editor. *PLOS Biol.* 18:e3000954.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. 2022. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50:D439–D444.

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. 2024. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* 52:D368–D375.

Vasileva EA, Mishchenko NP, Fedoreyev SA. 2017. Diversity of polyhydroxynaphthoquinone pigments in North Pacific sea urchins. *Chem. Biodivers.* 14:e1700182.

Vasimuddin Md, Misra S, Li H, Aluru S. 2019. Efficient architecture-aware acceleration of bwa-mem for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Rio de Janeiro, Brazil: IEEE. p. 314–324. Available from: https://ieeexplore.ieee.org/document/8820962/

Vasseur E. 1952. Geographic variation in the Norwegian sea urchins, *Strongylocentrotus droebachiensis* and *S. pallidus*. *Evolution* 6:87–100.

Veller C, Edelman NB, Muralidhar P, Nowak MA. 2023. Recombination and selection against introgressed DNA. *Evolution* 77:1131–1144.

Vendrami DLJ, De Noia M, Telesca L, Brodte E, Hoffman JI. 2020. Genome-wide insights into introgression and its consequences for genome-wide heterozygosity in the Mytilus species complex across Europe. *Evol. Appl.* 13:2130–2142.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343:1017–1021.

Ward GE, Brokaw CJ, Garbers DL, Vacquier VD. 1985. Chemotaxis of Arbacia punctulata spermatozoa to resact, a peptide from the egg jelly layer. *J. Cell Biol.* 101:2324–2329.

Weber AA-T, Stöhr S, Chenuil A. 2019. Species delimitation in the presence of strong incomplete lineage sorting and hybridization: lessons from Ophioderma (Ophiuroidea: Echinodermata). *Mol. Phylogenet. Evol.* 131:138–148.

Wen D, Yu Y, Zhu J, Nakhleh L. 2018. Inferring phylogenetic networks using phylonet.Posada D, editor. *Syst. Biol.* 67:735–740.

Wessel GM, Berg L, Adelson DL, Cannon G, McClay DR. 1998. A molecular analysis of hyalin—a substrate for cell adhesion in the hyaline layer of the sea urchin embryo. *Dev. Biol.* 193:115–126.

Wu C-I. 2001. The genic view of the process of speciation. *J. Evol. Biol.* 14:851–865.

Wuyun Q, VanKuren NW, Kronforst M, Mullen SP, Liu KJ. 2019. Scalable statistical introgression mapping using approximate coalescent-based inference. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Niagara Falls NY USA: ACM. p. 504–513. Available from: https://dl.acm.org/doi/10.1145/3307339.3342165

Yan Y, Tao H, He J, Huang S-Y. 2020. The HDOCK server for integrated protein–protein docking. *Nat. Protoc.* 15:1829–1852.

Yan Y, Zhang D, Zhou P, Li B, Huang S-Y. 2017. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* 45:W365–W373.

Yang Z. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22:1107–1118.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.

Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17:1446–1455.

Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci.* 111:16448–16453.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22:2472–2479.

Zhao T, Sun J, Zhan Y, Liu L, Song J, Zhang W, Chang Y. 2021. Comparative metabolic analysis between distant sea urchin hybrids (*Heliocidaris crassispina* ♀ × *Strongylocentrotus intermedius* ♂) and their parental purebred offspring. *Aquaculture* 541:736796.

Zheng Y, Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics* 19:10.

Zigler KS, Byrne M, Raff EC, Lessios HA, Raff RA. 2012. Natural hybridization in the sea urchin genus *Pseudoboletia* between species without apparent barriers to gamete recognition. *Evolution* 66:1695–1708.

Zigler KS, Lessios HA. 2004. Speciation on the coasts of the new world: phylogeography and the evolution of bindin in the sea urchin genus *Lytechinus*. *Evolution* 58:1225–1241.

Zigler Kirk S, Mccartney MA, Levitan DR, Lessios HA. 2005. Sea urchin bindin divergence predicts gamete compatibility. *Evolution* 59:2399–2404.

Zigler Kirk S., McCartney MA, Levitan DR, Lessios HA. 2005. Sea Urchin Bindin Divergence Predicts Gamete Compatibility. *Evolution* 59:2399–2404.

Zigler KS, Raff EC, Popodi E, Raff RA, Lessios HA. 2003. Adaptive evolution of bindin in the genus *Heliocidaris* is correlated with the shift to direct development. *Evolution* 57:2293–2302.