Comparative Genomics Insights into Speciation and Evolution of Hawaiian *Drosophila*

Lin Kang

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics and Computational Biology

Pawel Michalak, Chair
Igor Sharakhov, co-Chair
Hehuang (David) Xie
Liqing Zhang

March 10th, 2017
Blacksburg, VA

Keywords: Genomics, Comparative genomics,
Drosophila, Evolution, Adaptation

Comparative Genomics Insights into Speciation and Evolution of Hawaiian *Drosophila*

Lin Kang

ABSTRACT

Speciation and adaptation have always been of great interest to biologists. The Hawaiian archipelago provides a natural arena for understanding adaptive radiation and speciation, and genomics and bioinformatics offer new approaches for studying these fundamental processes. The mode of speciation should have profound impacts on the genomic architecture and patterns of reproductive isolation of new species. The Hawaiian *Drosophila* are a spectacular example of sequential colonization, adaptive radiation, and speciation in the islands with nearly 1,000 estimated species, of which more than 500 have been described to date.

This dissertation gives an overview of the Hawaiian *Drosophila* system (Chapter 1), new insights into genomes of three recently diverged species of Hawaiian picture-winged *Drosophila* (Chapter 2), as well as estimated gene flow patterns (Chapter 3). Additionally, I present a new approach of mapping genomic scaffolds onto chromosomes, based on NextGen sequencing from chromosomal microdissections (Chapter 4), and gene expression profiles of backcross hybrids and their parental forms (Chapter 5). Overall, obtained results were used to address such fundamental questions as the role of adaptive changes, founder effects (small effective population size in isolation), and genetic admixture during speciation.

Comparative Genomics Insights into Speciation and Evolution of Hawaiian *Drosophila*

Lin Kang

GENERAL AUDIENCE ABSTRACT

Speciation, or the origin of new species, and adaptation have always been of great interest not only to biologists, but also to the general public. Understanding how the fascinating diversity of nature is formed, and the processes behind it, may benefit the development and improvement of human society as a whole. The Hawaiian archipelago provides a natural arena for apprehending topics related to colonization, adaptation, and speciation. This microcosm, with the new approaches in genomics and bioinformatics, offers us the study of these very fundamental processes. The Hawaiian *Drosophila*'s 1,000 estimated species are a spectacular example of adaptive radiation and speciation, and an excellent model of research, as exemplified by this study. Using Hawaiian *Drosophila*, it addresses such fundamental questions in speciation as the roles of adaptive changes, founder effects (small effective population size in isolation), and genetic admixture during speciation.

Acknowledgements

I would like to thank numerous people for their support. Fist I would like to thank my advisor, Dr. Pawel Michalak, for encouragement and mentoring during my Ph.D. study. I have benefited much from his guidance and generosity in both academic and personal life over the past four years.

Additionally, I wholeheartedly thank Dr. Donald Price for providing the great resource of Hawaiian flies that made this project possible. The kind help and suggestions from my committee members, Dr. Igor Sharakhov, Dr. David Xie, and Dr. Liqing Zhang, made this dissertation better. I would like to thank Katarzyna Michalak for her effort and time in helping wet-lab experiments, and all my co-authors of publications for their contributions.

Many friends in Blacksburg have supported me through the years and greatly enriched my life outside of scientific research.

Finally and foremost, I dedicate this dissertation to my family for all of their continued love that makes all I do possible and worthwhile.

Table of Contents

List of Tables

List of Figures

Chapter 1 : Introduction

Hawaiian Islands

The Hawaiian archipelago consists of the exposed peaks of a huge undersea mountain range, formed by volcanic activity. They are all geologically young. The northwestern most island, Kure Atoll, is the oldest at about 28 million years (Ma); while the southeastern most island, Hawai'i, is about 0.4 Ma (Clague, Dalrymple, & Moberly, 1975). The Hawaiian Islands are about 2,500 miles from the nearest continent, and have never been connected to any body of land. Due to this extreme isolation, the archipelago is characterized by exceptional uniqueness of flora and fauna, which remained intact until 1,200 and 1,600 years ago when the humans who began colonizing the islands brought there many other plant and animal species (Olson, 2004). The geo-temporal chain of Hawaiian archipelago and the extensive elevations of its "high islands" allow distinct climate regimes and lead to many interesting characteristics for the native Hawaiian species. The majority native Hawaiian species are endemic, that is, they can only be found in Hawaii. Those endemic species include around 850 species of flowering plants, 240 species of ferns and allies, 5200 insects species, 130 spider species, 60 species of birds, and about 500 species of terrestrial molluscs (Eldredge & Miller, 1997). The rich variety of native species includes no native species of reptiles or amphibians. The only native mammals are a bat species and a seal species (Olson, 2004). Several of the Hawaiian taxa provide striking examples of adaptive radiation, making Hawaii an excellent system and incredibly important resource to study topics and models related to the origin of species and biodiversity.

Adaptive radiation and speciation in the Hawaiian Islands

Adaptive radiation is the development of many species derived from a single ancestral population. It is one of the most important outcomes of the evolution process (Schluter, 2000; Simpson, 1955). Islands are considered the best places to observe this phenomenon owing to their geological changes and dynamic environments. The multiple examples of adaptive radiation in the isolated Hawaiian Islands include silverswords, honeycreepers, crickets, and the most spectacular one, Hawaiian Drosophilidae (Seehausen, 2004). Many interesting topics are involved in the process of radiation, such as ecological release, adaptive plasticity, founder events, hybridization and gene flow, and sexual selection. Determining the processes by which species form is critical to our understanding of the diversity and sustainability of life. A central paradigm of biological speciation posits that reproductive isolation is important not only in defining species but also in determining the process by which speciation occurs. The biological speciation concept and related hypotheses indicate that species arise through a multi-stage process that begins with population divergence due to adaptive differences or drift. A critical difference is then proposed to occur between species that come to occur in sympatry compared with those that remain geographically separated and evolve in allopatry (Seehausen et al., 2014). The mode of speciation and the extent of gene flow between populations should have profound impacts on the genomic architecture and patterns of reproductive isolation of new species. Instances of speciation with gene flow can promote the rapid evolution of reproductive isolating barriers that are predicted to build up much more slowly in allopatry (Coyne & Orr, 2004; Seehausen et al., 2014). Speciation with gene flow should also promote relatively larger genomic changes, especially for gene sequences associated with reproductive isolating mechanisms and chromosomal inversions that restrict recombination and reduce the fitness of hybrids (Presgraves, 2010). Although divergence at both the phenotypic and genetic levels associated with speciation are documented, comprehensive analyses of genome changes within young radiations comprising both sympatric and allopatric speciation are lacking (Feder, Egan, & Nosil, 2012).

The Hawaiian Drosophila

The Hawaiian *Drosophila* are an iconic example of sequential colonization, adaptive radiation and speciation in the islands with 800-1000 estimated species. More than 550 have been described (Lapoint, Gidaya, & O'Grady, 2011; O'Grady et al., 2011). Due to the sensitivity to changes of climate regimes or even breeding-host density, the Hawaiian *Drosophila* are considered to be valuable “indicator species” to the ecosystem (D. K. Price & Muir, 2008; Sene & Carson, 1977). The sequential geological formation of islands in the Hawaiian archipelago have promoted the colonization and founder events by Hawaiian taxa to have occurred largely in sequential order from the oldest (northwest) to the youngest (southeast) islands (J. P. Price & Clague, 2002). In the Hawaiian picture-winged *Drosophila*, the charismatic subgroup of 120 species, host plant-specific adaptations, as well as sexual selection involving morphological ornaments and aggressive and mating behaviors have been proposed to underlie reproductive isolation and speciation (Hoikkala & Kaneshiro, 1993), especially in closely related species sharing the same island (H. L. Carson, 1997; H. L. Carson, Clague, D.A.,, 1995). For closely related species occurring on different islands, geographic isolation coupled with genetic drift or differential adaptation between island colonists and their source populations may be more important for speciation. This system is thus unique in that it provides natural allopatry-sympatry contrasts between pairs and clusters of closely related species, as well as distinct populations within certain species. One of the hottest debates about speciation concerns whether populations can diverge into new species when they geographically overlap and are connected by gene flow (i.e., speciation-with-gene-flow; (Berlocher & Feder, 2002; Coyne & Orr, 2004; Papadopulos, Baker, & Savolainen, 2013; Rashkovetsky, Frenkel, Michalak, & Korol, 2015)). Briefly, speciation-with-gene-flow includes situations in which populations overlap entirely in their ranges (sympatric) or only partially (parapatric). It also applies to cases in which populations have evolved in the face of continuous gene flow and have always been in contact (primary contact) or have experienced periods of geographic isolation (allopatry) and episodes of subsequent gene flow (secondary contact). The key question is whether additional reproductive isolation can evolve that culminates in speciation when gene flow is occurring between populations, either *de novo* in primary contact or by reinforcement in secondary contact.

Another significant aspect of this system concerns sexual selection and its role in speciation-with-gene-flow. Unlike males of other *Drosophila* that typically court females at feeding sites or their immediate vicinity, picture-winged *Drosophila* males establish territories within a mating arena (lek) where they exhibit male-male aggressive behaviors and engage in elaborate courtship in front of visiting females (H. I. Spieth, 1964; H. T. Spieth, 1967, 1981, 1982). Some of the most striking morphological differences between picture-winged *Drosophila* species are secondary sexual characters of males, such as hammer-shaped heads of *D. heteroneura* males. Male head width in this species positively correlates with aggressiveness and mating success through female choice

(Boake, DeAngelis, & Andreadis, 1997). The contribution of sexual selection to speciation and sexual selection signatures at the genome-wide level are difficult to disentangle from adaptive responses to environmental challenges, such as host-plant-specific interactions, especially due to the fact that sensory pathways used by *Drosophila* for food detection and mate recognition largely overlap (Grosjean et al., 2011; Nozawa & Nei, 2007; Ziegler, Berthelot-Grosjean, & Grosjean, 2013). Indeed, the Hawaiian picture-winged *Drosophila* are specialist saprophages uniquely adapted to native trees (Magnacca & Price, 2015). However, this study takes advantage of the biogeographical contrasts (allopatry versus sympatry), as well as the wealth of ecological data amassed for the group, including host plant information (Magnacca, Foote, & O'Grady, 2008; Magnacca & Price, 2015), such that at least in some comparisons effects of sexual selection and host-specific adaptations can be separated.

The Dissertation Chapters

Despite significance of Hawaiian *Drosophila*, genome information from this spectacular natural system has so far been limited to a single assembly from one species only. Here I endeavor to establish a new resource for speciation genomics using Hawaiian *Drosophila* as a model system. In the following chapters, I discuss studies based on this system.

In Chapter 2, three new genomes of Hawaiian picture-winged *Drosophila* are analyzed, represented by both allopatric and sympatric species in which host-plant associations and sexual selection involving elaborate mating, aggressive behaviors and their associated morphological ornaments have been proposed as important drivers of reproductive isolation and speciation. Their phylogenetic relationship, genomic changes, and signatures of positive selection are discussed. In Chapter 3, a modified four-taxon test is used to examine gene flow during speciation among the three Hawaiian *Drosophila* species. In Chapter 4, a method of mapping scaffolds from draft genome assembly to chromosomes using laser capture microdissection is described. Lastly, in Chapter 5, gene expression profiles of backcrosses with different sperm phenotypes between F1 hybrids of *D. planitibia*-*D. silvestris* and their parental species are surveyed to understand further the reproductive isolation during speciation.

References

Berlocher, S. H., & Feder, J. L. (2002). Sympatric speciation in phytophagous insects: moving beyond controversy? *Annu Rev Entomol, 47*, 773-815. doi:10.1146/annurev.ento.47.091201.145312

Boake, C. R. B., DeAngelis, M. P., & Andreadis, D. K. (1997). Is sexual selection and species recognition a continuum? Mating behavior of the stalk-eyed fly Drosophila heteroneura. *Proceedings of the National Academy of Sciences of the United States of America, 94*(23), 12442-12445.

Carson, H. L. (1997). The Wilhelmine E. Key 1996 Invitational Lecture. Sexual selection: a driver of genetic change in Hawaiian Drosophila. *J Hered, 88*(5), 343-352.

Carson, H. L., Clague, D.A.,. (1995). Geology and biogeography of the Hawaiian Islands. In W. Wagner, Funk, V. (Ed.), *Hawaiian Biogeography: Evolution in a Hotspot Archipelago.* (pp. 14-29). Washington, DC: Smithsonian Institution Press.

Clague, D. A., Dalrymple, G. B., & Moberly, R. (1975). Petrography and K-Ar Ages of Dredged Volcanic Rocks from the Western Hawaiian Ridge and the Southern Emperor Seamount Chain. *Geology Society of America Bulletin, 129*(86(7)), 8.

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA ;: Sinauer Associates.

Eldredge, L. G., & Miller, S. E. (1997). Numbers of Hawaiian species: supplement 2, including a review of freshwater invertebrates.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends Genet, 28*(7), 342-350.

Grosjean, Y., Rytz, R., Farine, J. P., Abuin, L., Cortot, J., Jefferis, G. S., & Benton, R. (2011). An olfactory receptor for food-derived odours promotes male courtship in Drosophila. *Nature, 478*(7368), 236-240.

Hoikkala, A., & Kaneshiro, K. (1993). Change in the signal-response sequence responsible for asymmetric isolation between Drosophila planitibia and Drosophila silvestris. *Proc Natl Acad Sci U S A, 90*(12), 5813-5817.

Lapoint, R. T., Gidaya, A., & O'Grady, P. M. (2011). Phylogenetic relationships in the spoon tarsus subgroup of Hawaiian Drosophila: conflict and concordance between gene trees. *Mol Phylogenet Evol, 58*(3), 492-501. doi:S1055-7903(10)00496-3 [pii]

Magnacca, K. N., Foote, D., & O'Grady, P. M. (2008). A review of the endemic Hawaiian Drosophilidae and their host plants. *Zootaxa, 1728*, 1-58.

Magnacca, K. N., & Price, D. K. (2015). Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing Drosophila (Diptera: Drosophilidae). *Mol Phylogenet Evol, 92*, 226-242.

Nozawa, M., & Nei, M. (2007). Evolutionary dynamics of olfactory receptor genes in Drosophila species. *Proc Natl Acad Sci U S A, 104*(17), 7122-7127.

O'Grady, P. M., Lapoint, R. T., Bonacum, J., Lasola, J., Owen, E., Wu, Y., & DeSalle, R. (2011). Phylogenetic and ecological relationships of the Hawaiian Drosophila inferred by mitochondrial DNA analysis. *Mol Phylogenet Evol, 58*(2), 244-256.

Olson, S. (2004) *Evolution in Hawaii: A Supplement to Teaching about Evolution and the Nature of Science*. Washington (DC).

Papadopulos, A. S. T., Baker, W. J., & Savolainen, V. (2013). Sympatric speciation of island plants. In P. Michalak (Ed.), *Speciation. Natural Processes and Biodiversity.* (pp. 59-81). New York: Nova Science Publishers.

Presgraves, D. C. (2010). The molecular evolutionary basis of species formation. *Nat Rev Genet, 11*(3), 175-180. doi:10.1038/nrg2718

Price, D. K., & Muir, C. (2008). Conservation implications of hybridization in Hawaiian picture-winged Drosophila. *Mol Phylogenet Evol, 47*(3), 1217-1226. doi:10.1016/j.ympev.2007.12.003

Price, J. P., & Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc Biol Sci, 269*(1508), 2429-2435. doi:10.1098/rspb.2002.2175

Rashkovetsky, E., Frenkel, Z., Michalak, P., & Korol, A. (2015). Sympatric differentiation and speciation: Insights from *Drosophila* studies. In P. Pontarotti (Ed.), *Evolutionary Biology: Biodiversification from Genotype to Phenotype.* (pp. 107-140). Switzerland: Springer Int. Publ.

Schluter, D. (2000). *The ecology of adaptive radiation*: OUP Oxford.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends Ecol Evol, 19*(4), 198-207. doi:10.1016/j.tree.2004.01.003

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., . . . Widmer, A. (2014). Genomics and the origin of species. *Nat Rev Genet, 15*(3), 176-192. doi:10.1038/nrg3644

Sene, F. M., & Carson, H. L. (1977). Genetic variation in Hawaiian Drosophila. IV. Allozymic similarity between D. silvestris and D. heteroneura from the island of Hawaii. *Genetics, 86*(1), 187-198.

Simpson, G. G. (1955). *Major features of evolution*: Columbia University Press: New York.

Spieth, H. I. (1964). Studies of the Mating Behavior of Endemic Hawaiian Drosophila. *American Zoologist, 4*(4), 406-406.

Spieth, H. T. (1967). Lek Behavior of Endemic Hawaiian Drosophila. *American Zoologist, 7*(2), 205-&.

Spieth, H. T. (1981). Courtship Behavior and Evolutionary Status of the Hawaiian Drosophila-Primaeva Hardy and Kaneshiro. *Evolution, 35*(4), 815-817.

Spieth, H. T. (1982). Behavioral Biology and Evolution of the Hawaiian Picture-Winged Species Group of Drosophila. *Evolutionary Biology, 14*, 351-437.

Ziegler, A. B., Berthelot-Grosjean, M., & Grosjean, Y. (2013). The smell of love in Drosophila. *Front Physiol, 4*, 72. doi:10.3389/fphys.2013.00072

Chapter 2 : Genomic Signatures of Speciation in Sympatric and Allopatric Hawaiian Picture-winged *Drosophila*

Lin Kang[1], Robert Settlage[1], Wyatt McMahon[2], Katarzyna Michalak[1], Hongseok Tae[1], Harold R. Garner[1], Elizabeth Stacy[3], Donald K. Price[3] & Pawel Michalak[1]

[1]Biocomplexity Institution, Virginia Tech, Blacksburg, VA 24061, USA

[2]Howard Hughes Medical Institute, Johns Hopkins Medical Institutes, Baltimore, MD 21287, USA

[3] Tropical Conservation Biology and Environmental Science Graduate Program, University of Hawai`i at Hilo, Hilo, HI 96720, USA

Genome Biol Evol. 2016 May; 8(5): 1482–1488.

Abstract

The Hawaiian archipelago provides a natural arena for understanding adaptive radiation and speciation. The Hawaiian *Drosophila* are one of the most diverse endemic groups in Hawai`i with up to 1000 species. We sequenced and analyzed entire genomes of recently diverged species of Hawaiian picture-winged *Drosophila*, *D. silvestris* and *D. heteroneura* from Hawai`i Island, in comparison with *D. planitibia*, their sister species from Maui, a neighboring island where a common ancestor of all three had likely occurred. Genome-wide SNP patterns suggest the more recent origin of *D. silvestris* and *D. heteroneura*, as well as a pervasive influence of positive selection on divergence of the three species, with the signatures of positive selection more prominent in sympatry than allopatry. Positively selected genes were significantly enriched for functional terms related to sensory detection and mating, suggesting that sexual selection played an important role in speciation of these species. In particular, sequence variation in *Olfactory receptor* and *Gustatory receptor* genes seems to play a major role in adaptive radiation in Hawaiian pictured-winged *Drosophila*.

Introduction

The advancement of genomics and bioinformatics provides new approaches to elucidate the relationships between evolutionary processes and genomic divergence patterns, as well as between genomic properties and speciation processes (Seehausen et al., 2014). The mode of speciation should have profound impacts on the genomic architecture and patterns of reproductive isolation of new species. Instances of speciation in sympatry with gene flow can promote the rapid evolution of reproductive isolating barriers that generally build up much more slowly in allopatry (Jerry A. Coyne & Orr, 2004; Noor, 1995; Seehausen et al., 2014). Although divergence at both the phenotypic and genomic levels associated with speciation are documented, combined analyses of genomic changes and empirical-based measures of reproductive isolation within young radiations comprising both sympatric and allopatric speciation are lacking (Feder, Egan, & Nosil, 2012). To investigate the genomic changes associated with both sympatric and allopatric settings, we have sequenced and assembled three new genomes from a recently diverged and well-studied clade of Hawaiian picture-winged *Drosophila*.

The Hawaiian *Drosophila* are a spectacular example of sequential colonization, adaptive radiation and speciation in the islands with nearly 1000 estimated species, of which more than 500 have been described to date (O'Grady et al., 2011). Within the system, the Hawaiian picture-winged *Drosophila* are a charismatic subgroup of ca. 120 species (Magnacca & Price, 2015). Sexual selection involving mating and aggressive behaviors and their associated morphological ornaments has been proposed as an important driver of reproductive isolation and speciation in the Hawaiian picture-winged *Drosophila* (Hoikkala & Kaneshiro, 1993), especially for closely related species sharing the same island (H. L. Carson, 1997; H. L. Carson, Clague, D.A.,, 1995). For closely related species occurring on different islands, geographic isolation coupled with genetic drift or differential adaptation between island colonists and their source populations may be more important for speciation.

Drosophila heteroneura and *D. silvestris* are two iconic Hawaiian picture-winged *Drosophila* species endemic to mid-altitude rainforests of the Big Island (Hawai`i), the youngest (< 0.5 million years old) of the Hawaiian Islands (H. L. Carson, 1982). Their closest relative, *D. planitibia*, presumably directly derived from the ancestral lineage that colonized the Big Island, is endemic to a similar habitat on Maui (Bonacum, O'Grady, Kambysellis, & Desalle, 2005; H. L. Carson & Kaneshiro, 1976; DeSalle & Giddings, 1986). All three species are morphologically distinct, particularly *D. heteroneura*, which possesses a novel stalk-eyed head shape (Fig 2.1A) and distinct male-male aggressive behaviors (K. Y. Kaneshiro, 1976; Price 1995). Over the past several decades, the species have experienced population declines, with *D. heteroneura* currently classified as federally endangered.

Interspecies mate discrimination is greater in experimental crosses between sympatric species relative to those between allopatric species, with *D. silvestris* and *D. heteroneura* females being less discriminatory against allopatric *D. planitibia* males (Ahearn, Carson, Dobzhansky, & Kaneshiro, 1974). Despite their greater mate discrimination, *D. silvestris* and *D. heteroneura* hybridize in nature (H. L. Carson, K. Y. Kaneshiro, F.C. Val, 1989) with asymmetrical mating between the species determined at an early stage of courtship, indicating the importance of species-recognition factors in the behavioral reproductive isolation and maintenance of species boundaries (Boake, 2005; K. Y. Kaneshiro, 1976; Price 1995; D. K. Price, Souder, & Varys, 2014). Patterns of fertility of F_1 hybrids also differ between the allopatric and sympatric species pairs. F_1 hybrid males are sterile in allopatric crosses, i.e., *D. planitibia* bred with either *D. silvestris* or *D. heteroneura* (Ahearn et al., 1974; Brill, Kang, Michalak, Michalak, & Price, 2016), but F_1 hybrid females are fertile, as is common in crosses between closely related *Drosophila* species (J. A. Coyne & Orr, 1997). Both F_1 hybrid males and females produced through crosses between sympatric *D. silvestris* and *D. heteroneura* are fully fertile and possess distinct combinations of morphological and behavioral traits consistent with both dominant and additive genetic factors that differ between the species (Boake, Price, & Andreadis, 1998). These contrasts suggest very different patterns of genome divergence in sympatry versus allopatry.

Due to the sequential geological formation of islands in the Hawaiian archipelago, founder events within the *Drosophila* appear to have occurred in sequential order from the oldest (northwest) to the youngest (southeast) islands (J. P. Price & Clague, 2002). Therefore, *D. silvestris* and *D. heteroneura* endemic to the slopes of the volcanoes on Hawai`i, the newest island in the chain (~0.5 MY), are not only the youngest species in the group, but their ancestry can be traced back to lineages from neighboring islands. *Drosophila planitibia* living on the geologically older island of Maui is the closest sister species to both Hawai`i Island species on the basis of morphological, behavioral, chromosomal, and genetic characteristics (DeSalle & Giddings, 1986; Hunt, Bishop, & Carson, 1984). An alternative hypothesis, however, based on morphological and behavioral data, places *D. planitibia* closer to *D. silvestris* and a separate species from Molokai, *D. differens*, situates closer to *D. heteroneura*, implying two independent ancestral lineages from different islands. Here, we report the sequencing and analysis of entire genomes of *D. planitibia*, *D. heteroneura* and *D. silvestris*, three recently diverged picture-winged species for which pre- and postzygotic reproductive isolating barriers are well characterized. We contrast the two leading phylogenetic hypotheses using the three new genomes and analyze the signatures of adaptive evolution that recent speciation may have left on genomes of these species.

Results and Discussion

Pooled genomes from 10 non-inbred individuals per species were used for construction of paired-end and mate pair libraries that were Illumina-sequenced at a total >80 x coverage. First, to determine if *D. silvestris* (SIL) originated from *D. planitibia* (PLA) and *D. heteroneura* (HET) diverged from another species, such as *D. differens*, we predicted that genetic distances should be lower between PLA and SIL than between PLA and HET, even under extensive gene flow between SIL and HET in sympatry. However, our genome-wide analysis of the average number of pairwise differences between sequences d_{XY} showed that SIL and HET were the closest relatives (Table 2.1, Fig 2.1B). Also, mean fixation index (F_{ST}) values based on 4,558,111 SNPs were lowest between SIL and HET (0.141), thus consistent with shorter divergence time and/or interbreeding between these two, but higher and very similar between PLA and SIL (0.306) and between PLA and HET (0.277). The average number of synonymous substitutions per synonymous site (Ks) showed a similar pattern, with the lowest value (0.041) between SIL and HET, and almost identical values for the other two pairwise comparisons (0.049). These estimates support an evolutionary scenario with the most recent phylogenetic split between SIL and HET.

A remarkable difference among the genomes of these species is that HET and SIL, the two sympatric species on Hawai`i Island and most closely related species of the three examined, have on average ~50% more genes with Ka/Ks > 1 than observed in the allopatric species pairs (PLA-SIL and PLA-HET; Fig 2.2, Appendix Table 2.1), many of them underlying sensory perception (GO term enrichment FDR < 0.002, Appendix Table 2.2). For comparison, most genes exhibiting signatures of purifying selection (Ka/Ks < 1) were conserved genes shared by all three species with patterns showing no relationship to sympatry or allopatry. The relative increase of positive selection in sympatry is possibly due to sexual selection and reinforcement, if hybrids are maladapted or when male secondary sex characters evolve through runaway processes (Higashi, Takimoto, & Yamamura, 1999; Noor, 1995). In fact, HET and SIL have very divergent morphologies that have been proposed to be associated with divergent mating or male aggressive behaviors (Boake 2005). The most conspicuous morphological trait of HET is the wide stalk-eyed-shaped head, while SIL and PLA have a round head more typical of other *Drosophila* (Fig 2.1A). Among the most abundant groups of over-represented genes driven by positive selection were many odorant (*Or*) and gustatory (*Gr*) receptor genes (Fig 2.3 and Appendix Table 2.3). In addition to *Or* and *Gr* genes, pheromone-binding *antennal protein 10* and neurological system- and courtship-related *spinster* were also driven by positive selection.

We then used the McDonald-Kreitman (MK) test that contrasts levels of polymorphism and divergence at neutral and functional sites (John H. McDonald & Martin Kreitman, 1991) to further examine signatures of adaptive evolution at the genomic level. Similar to the Ka/Ks test results, genes related to sensory perception and odorant binding formed one of the largest functionally enriched groups (GO term enrichment FDR < 0.05) of all genes driven by positive selection according to the MK test (P < 0.05, Appendix Table 2.4 & 2.5). Overall, out of the 62 *Or* and *Gr* genes, the MK test showed 23 (37%) genes to be under significant positive selection (direction of selection DoS >0, P < 0.05, Appendix Table 2.3), which is a threefold enrichment relative to all other genes (372 (11%) out of 3,470 genes, P=6.67×10^{-8}, Fisher exact test). Out of the 11 *Or* and *Gr* genes under positive selection indicated by Ka/Ks tests, eight showed significant signatures of positive selection in the MK tests as well.

Chemosensation in *Drosophila* is critical for detecting food and avoiding toxicants, as well as for courtship and mating. Since all three species share the same primary host plant (lobeliad trees of the genus *Clermontia* (K. Y. Kaneshiro, Val, F.C., 1977)), sexual selection operating on traits related to mate discrimination may have taken precedence over food-related adaptations in sensory divergence. HET and SIL males also differ in their concentrations of epicuticular hydrocarbons (Alves et al., 2010), molecules that are important for courtship communication in *Drosophila*. Lastly, courtship displays of SIL and HET also differ in the timing for stage advancement, the degree of female responsiveness, and the speed of body and/or wing movement (Hoikkala & Kaneshiro, 1993; Watson, 1979). The substantial differences in Ka/Ks values between SIL and HET compared to both SIL-PLA and HET-PLA suggest that strong behavioral reproductive isolation may be driving genome divergences in sympatry. In sum, this study emphasizes the significance of genomic sequences of Hawaiian picture-winged *Drosophila* in the inference of processes involved in speciation and adaptive radiation.

Materials and Methods

Flies

Genomic DNA was extracted (Gentra Puregene Tissue Kit, Qiagen) from 10 *D. heteroneura,* 10 *D. silvestris,* and 10 *D. planitibia* non-inbred males and pooled within species. DNA pooling enabled us to compare allele frequencies per SNP and estimate differentiation between populations (Fst), while keeping sequencing costs relatively low (Kofler, Pandey, & Schlotterer, 2011). *D. heteroneura* and *D. silvestris* individuals were from populations initiated with wild-caught individuals collected at the same location in the rainforest at 1,400 m elevation in the Kukuiopa‘e section of South Kona Forest Reserve from 2009-2011. *D. planitibia* originated from Waikamoi Preserve, east Maui, and were collected in December 2012. All flies were raised at the University of Hawai`i at Hilo.

Genome sequencing

Illumina paired-end HiSeq (2x100 bp, 500 bp inserts), paired-end Miseq (2x300 bp, 500 and 800 bp inserts), and Nextera Mate Pair libraries were sequenced at a total sequence coverage >80 x.

Genome assembly

Adapters were removed from the raw sequencing reads, and low quality and duplicated reads were discarded using FastqMcf (Aronesty, 2013), error corrections were performed by SOAPec from SOAPdenovo2 package (Luo et al., 2012). The 300 bp pair-end reads from Miseq were merged into long reads using mergepairs from ABYSS package (Simpson et al., 2009). Mate-pair sequences were processed with nextclip (Leggett, Clavijo, Clissold, Clark, & Caccamo, 2014) with default parameters, and reads from categories A, B and C were used to make the assembly. To exclude possible contamination, all reads were aligned to bacterial database downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), and unmapped reads were used for the assembly. Processed reads were assembled with Spades (Bankevich et al., 2012) and duplicated reads were removed with picard (https://github.com/broadinstitute/picard) according to the alignments which mapped the processed reads against the first assembly. De-duplicated reads were then assembled with Spades combined with scaffolding step using SSPACE (Boetzer, Henkel, Jansen, Butler, & Pirovano, 2011) (default parameters) and contigs with length less than 500 bp were discarded from further analyses (Table 2.2).

Genome completeness

The completeness of assembly was estimated using CEGMA by examining 248 core eukaryotic genes (Parra, Bradnam, & Korf, 2007). Completeness estimates were in the range of 93-98% ('complete') and 98-99% ('partial').

Gene Prediction and Annotation

Protein-coding genes were predicted using MAKER2 (Holt & Yandell, 2011), which used *D. melanogaster* protein sequences from FlyBase (r6.02, http://flybase.org) as protein homology evidence and integrated with prediction methods including BLASTX and SNAP. Predicted genes were subsequently used as query sequences in a blastx database search of NR database (non-redundant database, http://www.ncbi.nlm.nih.gov/). Blastx alignments with e-value greater than 1e-10 were discarded, and the top hit (or top hit from *Drosophila* species if existed) was used to annotate the query genes.

Functional enrichment

All functional enrichment analyses were performed by importing the appropriate gene list into DAVID (Huang da, Sherman, & Lempicki, 2009) and using annotated genes (HET, SIL, PLA) or *D. melanogaster* as background. GO terms with a Benjamini-Hoschberg-adjusted p-value of <0.05 were considered significant.

Ka/Ks ratio

To reduce the possible impact of Ka/Ks ratio by wrong annotation, we used only annotations against Swissprot (http://www.ebi.ac.uk/uniprot). Blastx alignments with e-value greater than 1e-40 or identity less than 40% were discarded. Sequences with same annotation were grouped together, and Clustal-omega (Sievers et al., 2011) was used to conduct the multiple sequence alignments. Nucleotide sequences were parsed to amino acid sequences before carrying multiple-sequence alignments to avoid possible frame-shift, and the amino acid sequences of alignment were changed back to nucleotide sequences for Ka/Ks calculations. PAML (Yang, 2007) (version 4.7) was used to calculate the Ka/Ks ratio values, setting the model = 0 in the control file of codeml. To minimize further the possible effect by the wrong annotation and grouping, Ks values greater than 2 were excluded from further analyses, and the maximal Ka/Ks value was set to be 3. Models M7/M8 along with likelihood ratio tests were applied to test for the significance of positive selection, with p-values generated from chi-square distribution (Nielsen & Yang, 1998). Pairwise Ks and Ka/Ks values are presented in a pairwise fashion, except for Appendix Table 2.4 that contains results of Ks and Ka/Ks from both pairwise comparisons and a single multispecies alignment (Ks(all) and Ka/Ks(all)).

McDonald-Kreitman test

To test for signatures of selection, we also used the McDonald-Kreitman (MK) test that compares the number of synonymous (Ds) and non-synonymous (Dn) substitutions between species with the number of synonymous (Ps) and non-synonymous (Pn) polymorphisms within species (J. H. McDonald & M. Kreitman, 1991). *Drosophila planitibia* was used as reference for mapping and detection of polymorphisms and substitutions. We used GATK (DePristo et al., 2011) with default parameters for genotyping. Only sites with the minimum depth of 10 and minimum genotyping quality of 30 were used. Sites with at least two reads supporting an alternative allele were considered polymorphic. Sites showing polymorphism in at least one of the three species were counted as polymorphic sites, and those with fixed differences between species were counted as substitutions. P-values were computed using Fisher exact test. Statistic DoS was used to determine the direction of selection (Stoletzki & Eyre-Walker, 2011), as given by: $DoS = \frac{D_n}{D_n+D_s} - \frac{P_n}{P_n+P_s}$. Positive and negative DoS values suggest positive and purifying selection, respectively.

F_{ST} and d_{XY} estimates

Sequences were mapped using BWA (Li & Durbin, 2010) with default parameters and *D. grimshawi* assembly as reference. Samtools (Li et al., 2009) was used to generate the pileup result. SNPs within 10 bp of an indel were discarded and Poopolation2 (Kofler, Pandey, et al., 2011) was used to estimate the F_{ST} value for each SNP. All pairwise analyses used the maximum number of sites. That is, F_{ST} estimates are based on sites that are polymorphic in at least one of the three species or divergent (if monomorphic) between at least two species. progressiveMauve (Darling, Mau, & Perna, 2010) was used for multiple sequence alignments of *D. silvestris*, *D. heteroneura*, *D. planitibia*, and *D. grimshawi*. PoPoolation (Kofler, Orozco-terWengel, et al., 2011) was used to estimate pairwise divergence (d_{XY}) with the window size set to 10 Kb.

Phylogeny

A total of 100 orthologs with the highest confidence from BLASTX alignments of *D. heteroneura, D. silvestris, D. planitibia, D. gramshawi* and *D. melanogaster* were used to construct a phylogenetic tree (Fig 2.1B). Mcmctree based on a Bayesian Markov Chain Monte Carlo (MCMC) algorithm from PAML (Yang, 2007) package was used to estimate the divergence time, and the calibration time was set using the divergence time between *D. melanogaster* and *D. grimshawi* from TimeTree (www.timetree.org), i.e. 39 (Thomas & Hunt, 1991) to 62.9 Mya (Tamura, Subramanian, & Kumar, 2004).

Competing interests

Authors declare no financial competing interests.

Acknowledgements

This work was funded in part by NSF CREST award HRD 0833211 to DKP and VBI's Medical Informatics and Systems funds to PM. We also thank Chris Parypa for his help with graphics.

Tables:

Table 2.1. Average pair-wise divergence (d_{XY}) values (below diagonal) and F_{ST} values (above diagonal) for *D. heteroneur*a (HET), *D. silvestris* (SIL), and *D. planitibia* (PLA) pairwise comparisons.

	HET	**SIL**	**PLA**
HET	-	0.141	0.277
SIL	0.0077	-	0.306
PLA	0.0130	0.0121	-

Table 2.2. Genome assembly attributes of the three Hawaiian picture-winged *Drosophila*.

	Assembly attributes					
	Total size	**No. of scaffold**	**No. of scaffold ≥ 1k**	**Contig N50**	**scaffold N50**	**GC content (%)**
D. silvestris	146,901,421	8,486	6,624	16,915	92,229	38.92
D. heteroneura	144,943,455	10,998	7,322	17,229	92,746	39.01
D. planitibia	188,994,020	15.471	11,830	154,334	399,542	40.55

Figure legends:

Fig 2.1. (A) Recent speciation in Hawaiian *Drosophila silvestris*, *D. heteroneura* and *D. planitibia*. (**B)** A phylogenetic tree based on 100 homologs from mcmctree (Yang, 2007). Divergence time in million years ago (Mya), with intervals in parentheses.

Fig 2.2. A Venn diagram illustrating overlap between *Drosophila silvestris*, *D. heteroneura* and *D. planitibia* in the number of genes driven by positive selection.

Fig 2.3. Gene Ontology terms and their statistical significance, showing an overrepresentation of genes related to sensory detection and cognition.

References

Ahearn, J. N., Carson, H. L., Dobzhansky, T., & Kaneshiro, K. Y. (1974). Ethological isolation among three species of the planitibia subgroup of Hawaiian Drosophila. *Proc Natl Acad Sci U S A, 71*(3), 901-903.

Alves, H., Rouault, J.-D., Kondoh, Y., Nakano, Y., Yamamoto, D., Kim, Y.-K., & Jallon, J.-M. (2010). Evolution of Cuticular Hydrocarbons of Hawaiian Drosophilidae. *Behavioral Genetics, 40*, 694–705.

Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal, 7*, 1-8. doi:10.2174/1875036201307010001

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol, 19*(5), 455-477.

Boake, C. R. B. (2005). Sexual selection and speciation in Hawaiian Drosophila. *Behavior Genetics, 35*(3), 297-303. doi:10.1007/s10519-005-3221-4

Boake, C. R. B., Price, D. K., & Andreadis, D. K. (1998). Inheritance of behavioural differences between two interfertile, sympatric species, Drosophila silvestris and D-heteroneura. *Heredity (Edinb), 80*, 642-650.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics, 27*(4), 578-579.

Bonacum, J., O'Grady, P. M., Kambysellis, M., & Desalle, R. (2005). Phylogeny and age of diversification of the planitibia species group of the Hawaiian Drosophila. *Mol Phylogenet Evol, 37*(1), 73-82.

Brill, E., Kang, L., Michalak, K., Michalak, P., & Price, D. K. (2016). Hybrid sterility and evolution in Hawaiian *Drosophila*: differential gene and allele-specific expression analysis of backcross males. *Heredity, , in press.*

Carson, H. L. (1982). Evolution of Drosophila on the newer Hawaiian volcanoes. *Heredity (Edinb), 48*(Pt 1), 3-25.

Carson, H. L. (1997). The Wilhelmine E. Key 1996 Invitational Lecture. Sexual selection: a driver of genetic change in Hawaiian Drosophila. *J Hered, 88*(5), 343-352.

Carson, H. L., Clague, D.A.,. (1995). Geology and biogeography of the Hawaiian Islands. In W. Wagner, Funk, V. (Ed.), *Hawaiian Biogeography: Evolution in a Hotspot Archipelago.* (pp. 14-29). Washington, DC: Smithsonian Institution Press.

Carson, H. L., K. Y. Kaneshiro, F.C. Val. (1989). Natural hybridization between the sympatric Hawaiian species Drosophila silvestris and Drosophila heteroneura. *Evolution, 43*(1), 190-203.

Carson, H. L., & Kaneshiro, K. Y. (1976). Drosophila of Hawaii - Systematics and Ecological Genetics. *Annual Review of Ecology and Systematics, 7*, 311-345. doi:DOI 10.1146/annurev.es.07.110176.001523

Coyne, J. A., & Orr, H. A. (1997). "Patterns of speciation in Drosophila" revisited. *Evolution, 51*(1), 295-303. doi:Doi 10.2307/2410984

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA ;: Sinauer Associates.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE, 5*(6), e11147. doi:10.1371/journal.pone.0011147

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498.

DeSalle, R., & Giddings, L. V. (1986). Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian Drosophila. *Proc Natl Acad Sci U S A, 83*(18), 6902-6906.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends Genet, 28*(7), 342-350.

Higashi, M., Takimoto, G., & Yamamura, N. (1999). Sympatric speciation by sexual selection. *Nature, 402*(6761), 523-526. doi:10.1038/990087

Hoikkala, A., & Kaneshiro, K. (1993). Change in the signal-response sequence responsible for asymmetric isolation between Drosophila planitibia and Drosophila silvestris. *Proc Natl Acad Sci U S A, 90*(12), 5813-5817.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics, 12*, 491. doi:10.1186/1471-2105-12-491

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57.

Hunt, J. A., Bishop, J. G., 3rd, & Carson, H. L. (1984). Chromosomal mapping of a middle-repetitive DNA sequence in a cluster of five species of Hawaiian Drosophila. *Proc Natl Acad Sci U S A, 81*(22), 7146-7150.

Kaneshiro, K. Y. (1976). Ethological isolation and phylogeny in the planitibia subgroup of Hawaiian *Drosophila*. *Evolution, 30*, 740-745.

Kaneshiro, K. Y., Val, F.C. (1977). Natural hybridization between a sympatric pair of Hawaiian Drosophila. *American Naturalist, 111*, 897-902.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., . . . Schlotterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE, 6*(1), e15925. doi:10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics, 27*(24), 3435-3436.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., & Caccamo, M. (2014). NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics, 30*(4), 566-568. doi:10.1093/bioinformatics/btt702

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., . . . Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience, 1*(1), 18. doi:10.1186/2047-217X-1-18

Magnacca, K. N., & Price, D. K. (2015). Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing Drosophila (Diptera: Drosophilidae). *Mol Phylogenet Evol, 92*, 226-242. doi:10.1016/j.ympev.2015.06.014

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature, 351*(6328), 652-654.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature, 351*(6328), 652-654. doi:10.1038/351652a0

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics, 148*(3), 929-936.

Noor, M. A. (1995). Speciation driven by natural selection in Drosophila. *Nature, 375*(6533), 674-675. doi:10.1038/375674a0

O'Grady, P. M., Lapoint, R. T., Bonacum, J., Lasola, J., Owen, E., Wu, Y., & DeSalle, R. (2011). Phylogenetic and ecological relationships of the Hawaiian Drosophila inferred by mitochondrial DNA analysis. *Mol Phylogenet Evol, 58*(2), 244-256.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067. doi:10.1093/bioinformatics/btm071

Price , D. K., Boake , C.R.B. . (1995). Behavioral reproductive isolation in Drosophila silvestris, D. heteroneura and their F1 hybrids (Diptera: Drosophilidae). *Journal of Insect Behavior, 8*, 595-616.

Price, D. K., Souder, S., & Varys, T. (2014). Sexual Selection, epistasis and species boundaries in sympatric Hawaiian picture-winged Drosophila. *Journal of Insect Behavior, 27*(1), 27-40.

Price, J. P., & Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc Biol Sci, 269*(1508), 2429-2435. doi:10.1098/rspb.2002.2175

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., . . . Widmer, A. (2014). Genomics and the origin of species. *Nat Rev Genet, 15*(3), 176-192. doi:10.1038/nrg3644

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., . . . Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol, 7*, 539. doi:10.1038/msb.2011.75

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res, 19*(6), 1117-1123. doi:10.1101/gr.089532.108

Stoletzki, N., & Eyre-Walker, A. (2011). Estimation of the neutrality index. *Mol Biol Evol, 28*(1), 63-70. doi:10.1093/molbev/msq249

Tamura, K., Subramanian, S., & Kumar, S. (2004). Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol, 21*(1), 36-44. doi:10.1093/molbev/msg236

Thomas, R. H., & Hunt, J. A. (1991). The molecular evolution of the alcohol dehydrogenase locus and the phylogeny of Hawaiian Drosophila. *Mol Biol Evol, 8*(5), 687-702.

Watson, G. F. (1979). On premating isolation between two closely related species of Hawaiian Drosophila. *Evolution, 33*, 771-774.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Fig 2.1

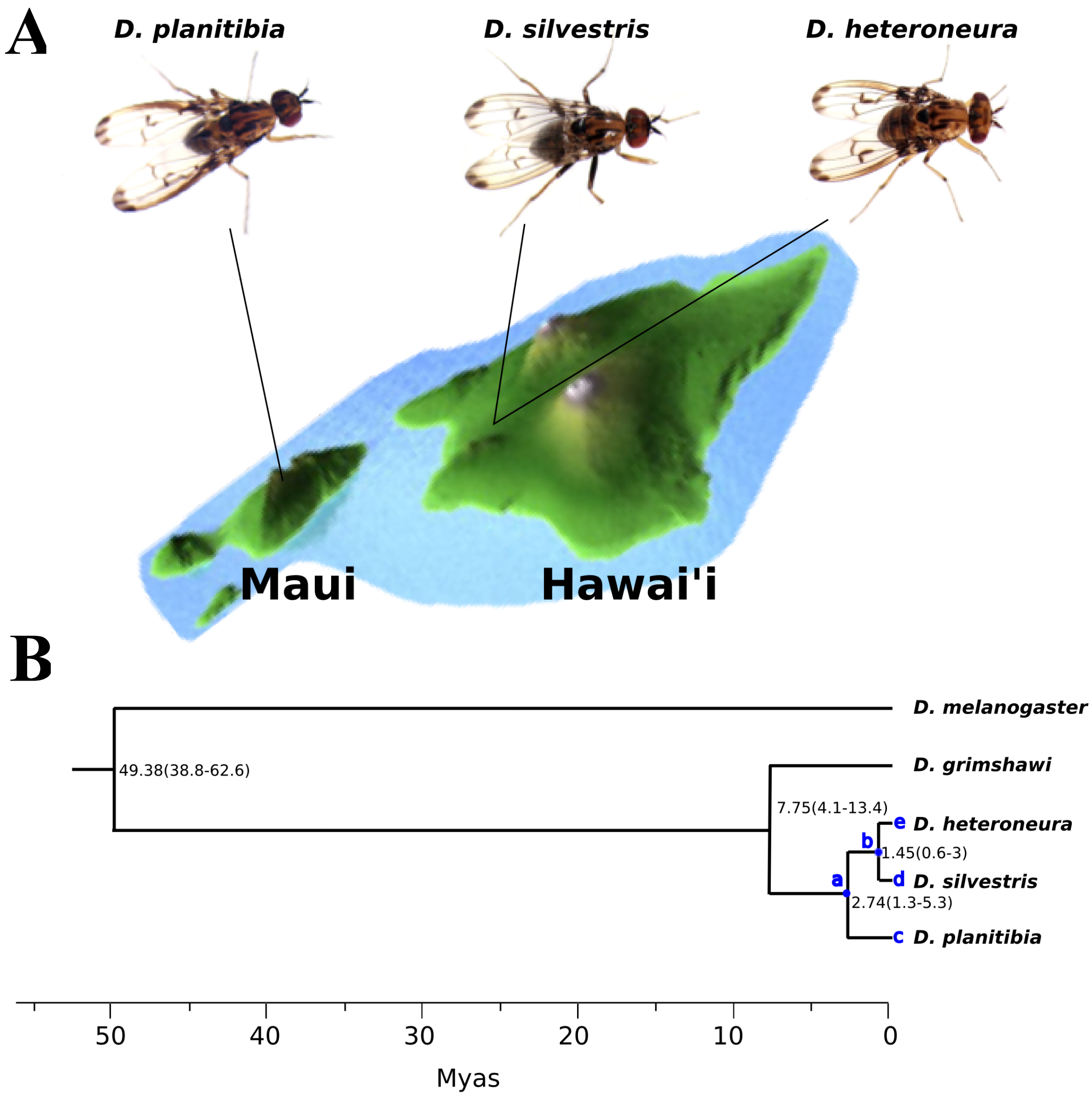

A
D. planitibia
D. silvestris
D. heteroneura
Maui
Hawai'i
B
D. melanogaster
D. grimshawi
D. heteroneura
D. silvestris
D. planitibia
49.38(38.8-62.6)
7.75(4.1-13.4)
1.45(0.6-3)
2.74(1.3-5.3)
a
b
c
d
e
50
40
30
20
10
0
Myas

Fig 2.2

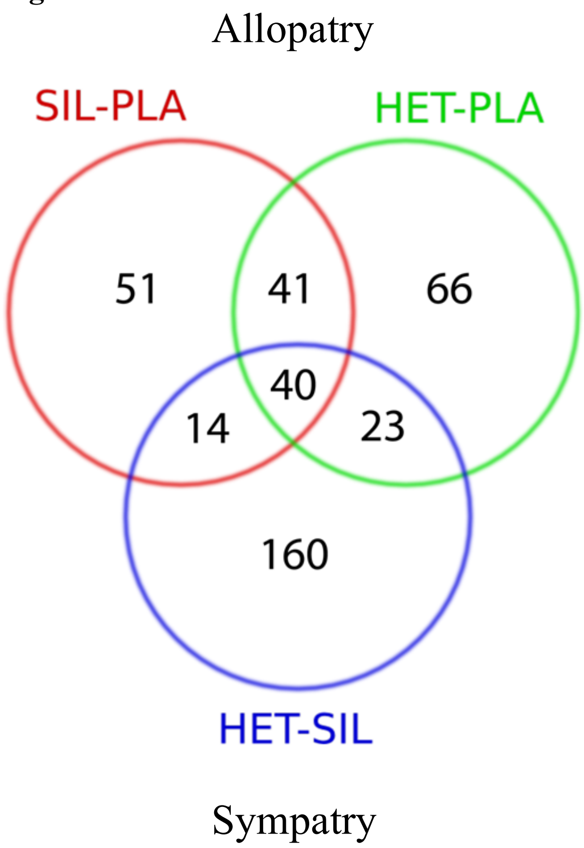
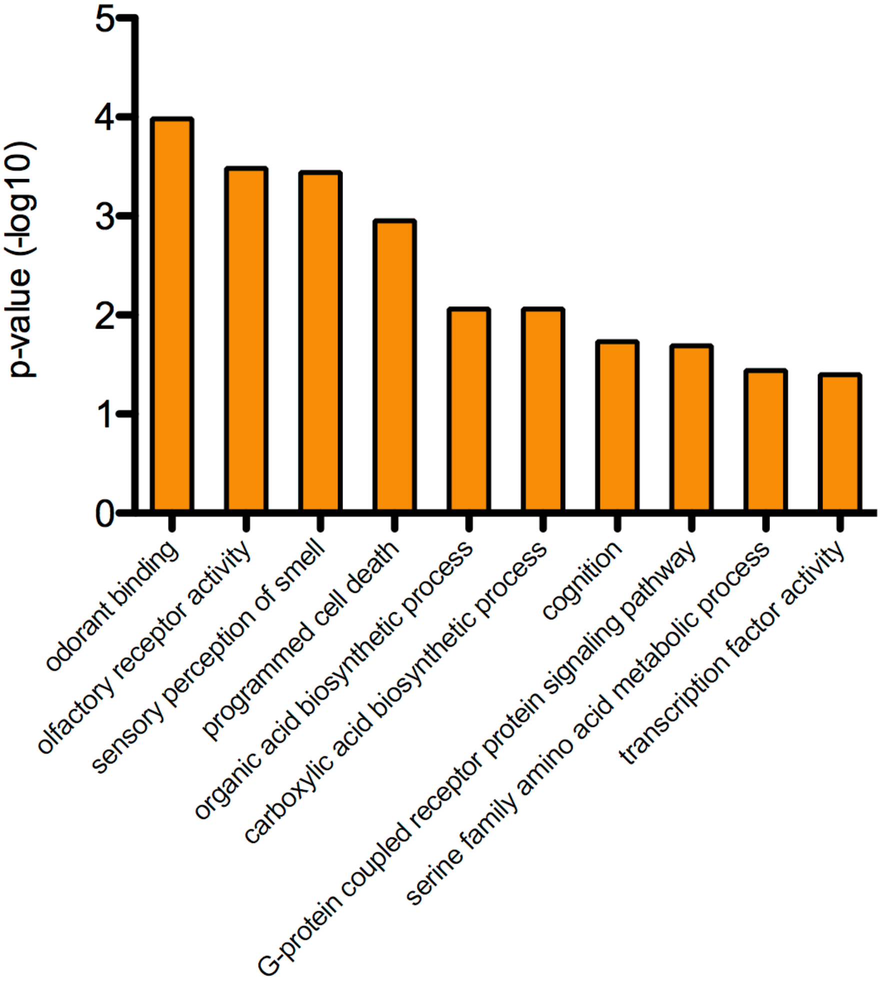

Allopatry
SIL-PLA
HET-PLA
51
41
66
40
14
23
160
HET-SIL
Sympatry

Fig 2.3

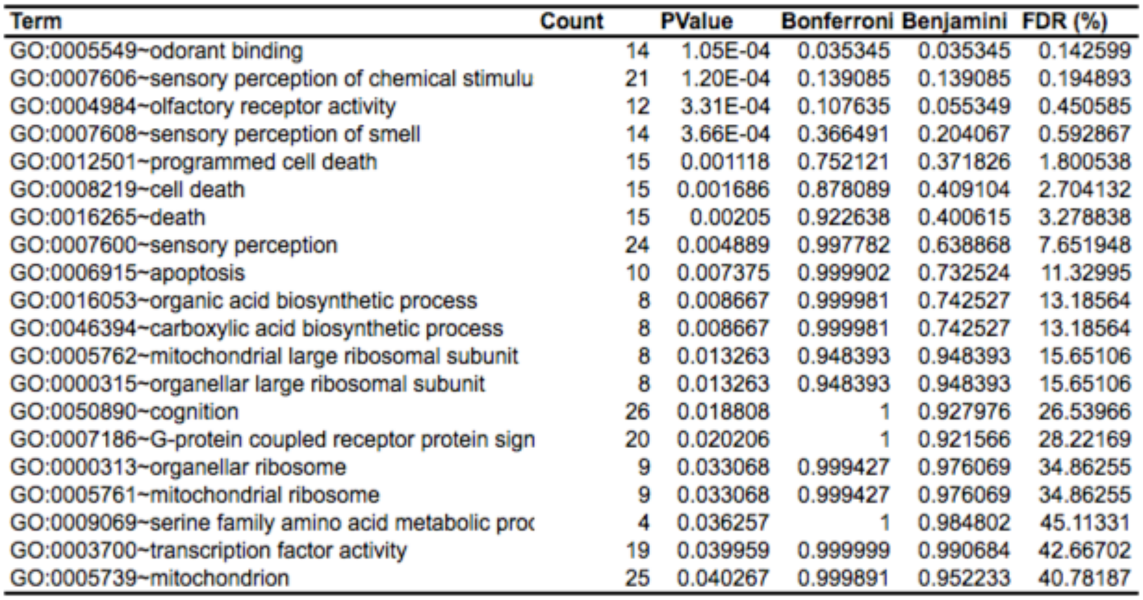

p-value (-log10)
5
4
3
2
1
0
odorant binding
olfactory receptor activity
sensory perception of smell
programmed cell death
organic acid biosynthetic process
carboxylic acid biosynthetic process
cognition
G-protein coupled receptor protein signaling pathway
serine family amino acid metabolic process
transcription factor activity

Appendix

Appendix Table 2.1. Mean Ks and Ka/Ks

	Ka/Ks	Ks
HET-PLA	0.234	0.03
HET-SIL	0.296	0.021
SIL-PLA	0.257	0.027

Appendix Table 2.2. GO and pathway enrichment for gene under positive selection (Ka/Ks tests)



Term	Count	PValue	Bonferroni	Benjamini	FDR (%)
GO:0005549~odorant binding	14	1.05E-04	0.035345	0.035345	0.142599
GO:0007606~sensory perception of chemical stimulu	21	1.20E-04	0.139085	0.139085	0.194893
GO:0004984~olfactory receptor activity	12	3.31E-04	0.107635	0.055349	0.450585
GO:0007608~sensory perception of smell	14	3.66E-04	0.366491	0.204067	0.592867
GO:0012501~programmed cell death	15	0.001118	0.752121	0.371826	1.800538
GO:0008219~cell death	15	0.001686	0.878089	0.409104	2.704132
GO:0016265~death	15	0.00205	0.922638	0.400615	3.278838
GO:0007600~sensory perception	24	0.004889	0.997782	0.638868	7.651948
GO:0006915~apoptosis	10	0.007375	0.999902	0.732524	11.32995
GO:0016053~organic acid biosynthetic process	8	0.008667	0.999981	0.742527	13.18564
GO:0046394~carboxylic acid biosynthetic process	8	0.008667	0.999981	0.742527	13.18564
GO:0005762~mitochondrial large ribosomal subunit	8	0.013263	0.948393	0.948393	15.65106
GO:0000315~organellar large ribosomal subunit	8	0.013263	0.948393	0.948393	15.65106
GO:0050890~cognition	26	0.018808	1	0.927976	26.53966
GO:0007186~G-protein coupled receptor protein sign	20	0.020206	1	0.921566	28.22169
GO:0000313~organellar ribosome	9	0.033068	0.999427	0.976069	34.86255
GO:0005761~mitochondrial ribosome	9	0.033068	0.999427	0.976069	34.86255
GO:0009069~serine family amino acid metabolic proc	4	0.036257	1	0.984802	45.11331
GO:0003700~transcription factor activity	19	0.039959	0.999999	0.990684	42.66702
GO:0005739~mitochondrion	25	0.040267	0.999891	0.952233	40.78187

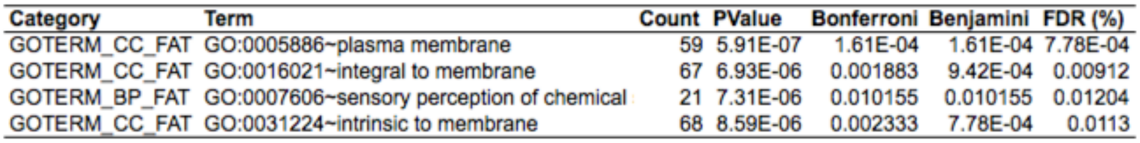
Appendix Table 2.3. Results of tests for positive selection (Ks and Ka/Ks from both pairwise comparisons and a single multispecies alignment (Ks(all) and Ka/Ks(all)) and McDonald-Kreitman tests, MKT) for odorant and gustatory receptor genes in D. planitibia (PLA), D. silvestris (SIL), and D. heteroneura (HET).

See (Kang et.al 2016) Supplementary Data (Supplementary Table S3)

Appendix Table 2.4. Results of tests for selection with McDonald-Kreitman tests and genome-wide MK approach SnIPRE for 3532 genes. Ds: the number of synonymous substitutions per gene; Dn: the number of non-synonymous substitutions per gene; Ps: the number of synonymous polymorphisms per gene; Pn: the number of non-synonymous polymorphisms per gene.

See (Kang et.al 2016) Supplementary Data (Supplementary Table S4)

Appendix Table 2.5. GO and pathway enrichment for genes under positive selection (McDonald-Kreitman tests).



Category	Term	Count	PValue	Bonferroni	Benjamini	FDR (%)
GOTERM_CC_FAT	GO:0005886~plasma membrane	59	5.91E-07	1.61E-04	1.61E-04	7.78E-04
GOTERM_CC_FAT	GO:0016021~integral to membrane	67	6.93E-06	0.001883	9.42E-04	0.00912
GOTERM_BP_FAT	GO:0007606~sensory perception of chemical	21	7.31E-06	0.010155	0.010155	0.01204
GOTERM_CC_FAT	GO:0031224~intrinsic to membrane	68	8.59E-06	0.002333	7.78E-04	0.0113

Chapter 3 : A Test for Gene Flow among Sympatric and Allopatric Hawaiian picture-winged *Drosophila*

Lin Kang[1], Robert Settlage[1], Wyatt McMahon[2], Katarzyna Michalak[1], Hongseok Tae[1], Harold R. Garner[1], Elizabeth Stacy[3], Donald K. Price[3] & Pawel Michalak[1]

[1]Biocomplexity Institution, Virginia Tech, Blacksburg, VA 24061, USA

[2]Howard Hughes Medical Institute, Johns Hopkins Medical Institutes, Baltimore, MD 21287, USA

[3] Tropical Conservation Biology and Environmental Science Graduate Program, University of Hawai`i at Hilo, Hilo, HI 96720, USA

Abstract

The Hawaiian *Drosophila* are one of the most species-rich endemic groups in Hawaii. *Drosophila silvestris* (SIL) and *D. heteroneura* (HET) are two closely related picture-winged *Drosophila* species that occur sympatrically on Hawai`i Island and are known to hybridize in nature, yet exhibit divergent behavioral and morphological traits driven largely through sexual selection. In contrast *D. planitibia* (PLA), their closest sister species from Maui, exhibits hybrid male sterility but reduced behavioral reproductive isolation with the two Hawai`i Island species. A modified four-taxon test for gene flow was examined among the three sequenced Hawaiian *Drosophila* species. The analysis indicated recent gene flow in sympatry but also, surprisingly, between allopatric species.

Introduction

The likelihood of speciation-with-gene-flow is still one of the most debated topics in evolutionary biology with the difficulty of isolating the effect of gene flow and the effect of time since population divergence (Nosil, 2008), while gene flow is frequently considered as a constraining force in evolution. Nevertheless, more and more examples of speciation-with-gene-flow were reported, and they advance our understanding of speciation in the absence of geographic obstacles (Jonsson et al., 2014; Keller et al., 2013; Martin et al., 2013). Several frameworks were also constructed for the comparison of speciation with and without gene flow (Feder et al., 2014; Smadja & Butlin, 2011). In the speciation-with-gene-flow process, the effective migration across the genome depends on many factors such as recombination rate, strength of selection, and number of selected loci. The whole genomic analysis can provide better empirical understanding of the genetic differentiation and speciation (Gagnaire, Pavey, Normandeau, & Bernatchez, 2013).

Due to the sequential geological formation of islands in the Hawaiian archipelago, founder events within the *Drosophila* appear to have occurred in sequential order from the oldest (northwest) to the youngest (southeast) islands (Price & Clague, 2002). *D. silvestris* (SIL) and *D. heteroneura* (HET) are the youngest species in the group, and *Drosophila planitibia* (PLA) is their closest sister species living on the Maui. Due to the lack of geographic barrier, although there is mate discrimination, *D. silvestris* and *D.*

heteroneura hybridize in nature (Carson, 1989). This indicates the likely gene flow between the two species. The test of gene flow between allopatric species would also be interesting in consideration of the 26-mile gap between Maui and the "Big island", Hawai`i. One method tests for a genome-wide excess of shared derived alleles between taxa using Patterson's *D* statistic. This four-taxon test or informal ABBA/BABA test was developed to determine whether there was any admixture between modern human and neanderthal populations (Green et al., 2010). Based on the data from (Kang et al., 2016), here we used a modified ABBA/BABA method which incorporates the fixation index (F_{ST}) value to test the gene flow of sympatry as well as allopatry. Obtained results will be used to address such fundamental question of gene flow in speciation.

Result and discussion

We used the ABBA/BABA method (Green et al., 2010) to contrast the two phylogenetic hypotheses, based on a total of 3,736,586 sites with fixed differences. To determine if paleo-PLA could be an ancestor of SIL and HET, we used a modified version of the ABBA/BABA method (Green et al., 2010) with *D. grimshawi* (GRI)*,* a more distantly related Hawaiian picture-winged *Drosophila* (Clark et al., 2007) as an outgroup. We estimated the frequency of the B·B·A·A (in the order: HET·SIL·PLA·GRI throughout the text) SNP pattern as a measure of the inter-node divergence time (**a-b**) between the most recent common SIL-HET ancestor and PLA ancestor (Fig 2.1B). The BBAA pattern was abundant. It was enriched beyond a frequency easily attributable to inter-node ancestral divergence, as it exceeded the divergence of the PLA (AABA) branch by 11% (jackknifed Z-score = 3.78, P < 0.0002). This also suggests that at least some SNPs shared by SIL and HET (BBAA) may have originated through introgression after sympatric interbreeding between SIL (ABAA) and HET (BAAA). This potential genetic admixture between SIL and HET is consistent with previous observations that up to 2% of F_1 and backcross SIL-HET hybrids were found in some localities of Hawai`i (Carson, 1989). After subtracting the estimates of recombinant BBAA from the total BBAA count (see Materials and Methods for details), we found that the length **a-b** was 16.5% of the PLA branch or less. This coalesces PLA very close to the most recent ancestral node for HET and SIL, indicating that the PLA lineage was likely ancestral to both HET and SIL.

The relatively high frequency of the PLA-specific AABA SNP pattern indicates divergence time almost twice as long as that in the PLA branch relative to the HET (BAAA) and SIL (ABAA) branches. This is roughly consistent with the difference in the geological ages of Maui and Hawai`i. Additionally, the HET-specific BAAA SNP pattern is enriched relative to the SIL-specific ABAA pattern (jackknifed Z-score = 10.54). This possibly reflects elevated genetic drift in HET due to the dramatically reduced population size that has led to its endangered species status (Muir & Price, 2008), or reflects the admixture between HET and an unknown species or population having occurred after the HET-SIL split. Using Tajima's estimates of nucleotide diversity π and Watterson's θ (Fu, 1994) in HET (π = 0.00647, θ = 0.0073), SIL (π = 0.00798, θ = 0.0101), and PLA (π = 0.0109, θ = 0.0125) (Fig 3.1), as well as mutation rate estimates $\mu$from *D. melanogaster* ($\mu = 3.5e^{-9}$ (Keightley et al., 2009)), we estimated effective population sizes N_e for HET, SIL, and PLA to be 504,286, 728,571, and 1,057,143, respectively. This smaller N_e of *D. heteroneura* suggests that drift effects in this species may indeed be stronger compared with the other species. From the shapes of the site frequency spectrum (or allele frequency spectrum) of the three species, SIL and HET have more similar demographic configurations compared to that of PLA (Fig 3.2). We also found a high frequency of both ABBA and BABA, two SNP patterns that are incongruent with phylogeny. Their presence can be explained by 1) hybridization between PLA and SIL (ABBA) and PLA and HET (BABA), 2) recurrent mutations, or 3) incomplete lineage sorting (Green et al., 2010). Remarkably, the ABBA/BABA test (Green et al., 2010) indicates significant hybridization between HET and PLA (mean=-0.087, SE=0.022, jackknifed Z-score=-4.03). However, caution is needed in interpreting these results, since BAAA is overrepresented relative to ABAA, which could be a violation of substitution or fixation rate constancy. We notice that the ABBA/BABA test is in fact no longer significant (mean=-0.006, SE=0.007, jackknifed Z =-0.92) when both polymorphic and fixed sites are counted, as opposed to fixed sites only; and BAAA and ABAA numbers are then less dissimilar (301,326 and 226,184, respectively). To explore hybridization as a source of ABBA/BABA enrichment, we reasoned that recently introgressed haplotypes have had insufficient time to be broken down by recombination, and that closely linked introgressed alleles should therefore occur in linkage-disequilibrium 'islands' (Martin et al., 2013; Sankararaman, Patterson, Li, Paabo, & Reich, 2012), which, in turn, should exhibit reduced sequence divergence in recipient-donor species comparisons relative to non-introgressed regions. To narrow down potential introgression 'islands,' we focused on the neighborhoods of all informative SNP patterns from the ABBA/BABA approach and their species-pairwise F_{ST} values (Table 3.1). A number of very specific predictions can be tested using this new method. For example, if ABBA and BABA reflect interbreeding between PLA and, respectively, SIL and HET, F_{ST} values from comparisons involving PLA should be significantly decreased. Such a pattern was indeed observed. In fact, ABBA and BABA produced the largest and most statistically significant differences between normalized F_{ST} values (jackknifed 7.05 < |Z-scores| < 17.58, Appendix 3.1), exceeding the estimates for BBAA. This is expected only if BBAA was a mixture of ancestral and HET-SIL post-hybridization introgressive genotypes, whereas ABBA and BABA were mostly post-hybridization-introgressed genotypes. For comparison, relatively low differences between normalized F_{ST} values were found within AABA, a genotype specific of PLA divergence with F_{ST} estimates least likely to have been affected by introgression/recombination on Hawai`i Island. We also predicted that if BBAA derives from recent interbreeding and introgression between HET and SIL, there should be a significant decrease in F_{ST} between HET and SIL, but not in PLA-SIL or PLA-HET, across the BBAA neighborhoods. Indeed, not only did we observe a dramatic (2.1 x) decrease in mean HET-SIL F_{ST} values, but we also found that the lowest F_{ST} values within the HET-SIL comparison were produced by BBAA, as expected under introgressive hybridization (Table 3.1; jackknifed |Z-scores| > 11.03).

Alternatively, F_{ST} values around BBAAs will necessarily be biased downwards if the speciation time is much more recent than the ultimate time to coalesce or, in other words, BBAA are surrounded by neighborhoods of LD spikes originating from the HET-SIL common ancestor rather than from recent introgressions. To minimize this bias, we focused on fixed SNPs, which under a neutral model require on average ~$4N_e$ generations for fixation, a long and likely sufficient time to erase most of the internal branches corresponding to a shallower coalescent history. Although this time-dependent homogenizing effect of recombination applies to both ancestral and newly introgressed linkage groups, our approach introduces an ascertainment bias in favor of recent LD spikes. Additional support for enrichment of the configurations BBAA, BABA, and ABBA due to introgressive hybridization comes from the analysis of species pairwise F_{ST} with respect to genomic window sizes used (Fig 3.3). Unlike HET-PLA and SIL-PLA, HET-SIL F_{ST} values from BBAA neighborhoods tend to grow with the window size, consistent with introgression erosion due to recombination (Fig 3.3A). Similarly, SIL-PLA F_{ST} values tend to increase only in ABBA neighborhoods (Fig 3.3B), while HET-SIL F_{ST} values rise only in BABA neighborhoods (Fig 3.3C). This suggest possible introgressive hybridization between PLA and SIL, as well as between PLA and HET, respectively. All other SNP patterns produce F_{ST} values slightly declining with the window size, which is a typical genome-wide tendency due to the greater sampling error in smaller windows (Beissinger, Rosa, Kaeppler, Gianola, & de Leon, 2015). Thus in addition to previously reported SIL-HET hybridization (Carson, 1989), there may have been occasional interbreeding between PLA and the other two species, after migrants had successfully crossed the 26-mile wide ‘Alenuihaha Channel separating Maui from Hawai`i. Although the direction of the inter-island migration and hybridization remains unknown, a Maui-Hawai`i flyover by PLA is a more parsimonious scenario as compared to HET and SIL both migrating back to Maui.

Materials and Methods

F_{ST} calculation

Sequences were mapped using BWA (Li & Durbin, 2010) with default parameters and *D. grimshawi* assembly as reference. Samtools (Li et al., 2009) was used to generate the pileup result. SNPs within 10 bp of indel were discarded, and Poopolation2 (Kofler, Pandey, & Schlotterer, 2011) was used to estimate the F_{ST} value for each SNP. All pairwise analyses used the maximum number of sites, i.e., F_{ST} estimates are based on sites that are polymorphic in at least one of the three species or divergent (if monomorphic) between at least two species.

ABBA/BABA approach

We used a modified ABBA/BABA method (Green et al., 2010) using *D. grimshawi* (GRI) as an outgroup. Reference *D. grimshawi* genome sequences were downloaded from FlyBase (r1.3, http://flybase.org), and PLA, HET, and SIL reads were mapped using BWA (Li & Durbin, 2010) with default parameters. We generated the genotypes for each species using GATK (DePristo et al., 2011) with default parameters except for setting heterozygosity to 0.01. Only sites with genotyping quality greater than 30 and minimal depth 10 were kept. After genotyping, we determined the $G_1G_2G_3G_4$ pattern for each SNP position, where G_1, G_2, G_3 and G_4 represent genotypes in HET, SIL, PLA and GRI, respectively. GRI alleles were considered ancestral (A), and sites with more than two alleles were filtered out. For each site, species were either assigned a "A" or "B" in the four taxa pattern according to the agreement of its genotype with GRI. The "AABA" pattern is the position in which PLA carries derived allele "B," but HET and SIL carry ancestral allele "A" as GRI. Similarly, the "BBAA" pattern indicates that HET and SIL carry derived allele, while PLA carries the ancestral allele, and so forth. We used similar *D* statistic (Green et al., 2010) for AABA and BBAA and other combinations to test if the two patterns were of equivalent abundance/counts using a jackknife block size of 2 Mb. To evaluate the linkage between the four taxa pattern and sequence diversity, we estimated F_{ST} values in the neighborhood of each pattern p ($p \in \{AABA, ABAA, ABBA, BAAA, BABA, BBAA, BBBA\}$). First, for each SNP from the pattern p we calculated *Rs* (where $s \in \{HS, HP, SP\}$ and H, S, P stands for HET, SIL and PLA, respectively), which is an average normalized F_{ST} of N downstream and N upstream SNPs (excluding the SNP itself). Normalized F_{ST} is obtained from the raw F_{ST} value divided by the mean F_{ST} for each species-pairwise comparison s. In each block of the size of n bases, the statistic *D* for s1 and s2 ($s1, s2 \in \{HS, HP, SP\}$) was calculated:

$$D_{s1,s2;p} = \frac{\sum_{i=1}^{n} B_{s1,s2}(i)}{\sum_{i=1}^{n} C_p(i)}$$

where $C_p(i)$ can be 1 or 0 according to the pattern p is seen in position i or not, $B_{s1,s2}(i)$ is given by:

$$B_{s1,s2}(i) = \begin{cases} 1, & if\ R_{s1}(i) > R_{s2}(i) \\ 0, & if\ R_{s1}(i) = R_{s2}(i) \\ -1, & if\ R_{s1}(i) < R_{s2}(i) \end{cases}$$

Other than Fig 3 with various Ns, results for N=100 only are reported throughout the text. Mean D and its variance were used to obtain Z-scores for jackknife tests, with the block size of 2 Mb, and a total of 100 blocks were generated. To reduce the effect of recent

HET-SIL ancestry as a confounding factor in F_{ST} calculations for the HET-SIL pair, we used only fixed SNPs because fixation permits more time (~$4N_e$) for eroding the association by HET-SIL ancestral LD under the neutral model. For comparison, we used the same method and tested for all sites including non-fixed genotypes (total 7,359,805 sites, including 5,216,097 sites found in SIL and 5,215,281 sites found in HET, thus indicating similar mutation rates in SIL and HET). We obtained very similar results, including F_{ST} dynamics (S11 File) and ancestral branch estimates (see below). We also simulated alternative demographic models on ABBA/BABA expectations using MSMS software (Ewing & Hermisson, 2010) and confirmed that our results are best explained by gene flow between species (Appendix 3.2 & 3.3).

The inference of ancestral relationships

Using a neutral coalescent model without gene flow, incomplete lineage sorting or recurrent mutations, let us assume that AABA (a total of 118,181) accumulate only along the PLA branch (**a-c**, Fig 2.1B). We also assume that XBAA (X can be either A or B) accumulate along the branch **a-b-d** leading to SIL (Let us ignore HET branch for now, as HET may have hybridized with an unknown species – see Results), and we note that lengths of branches **a-c** and **a-b-d** are the same (assuming equal mutation rates in all three species). Therefore, the number of XBAA should be equivalent to AABA. XBAA can be divided into 3 sets, with the first set accumulating divergence along the branch **a-b** belonging to the common SIL-HET ancestor, quantified as S_1, or the count of ancestral BBAA genotypes. The second set includes variants accumulated along the branch **b-d** in SIL and exchanged with HET through hybridization (S_2, the count of recombinant BBAA genotypes). The third set includes variants accumulated along the branch **b-d** and unique to SIL, i.e., not exchanged with HET (S_3, the count of ABAA genotype = 42,187). Then $S_1 + S_2 + S_3 = AABA$, and thus $S_2 = AABA - S_3 - S_1 = 75994 - S_1$. If we assume that SIL and HET had equal contributions to the number of recombinant BBAA, its count cannot exceed $2S_2$, that is, $< 151988 - 2S_1$. Since the total number of BBAA is 132558 and should not exceed $151988 - 2S_1 + S_1$, we obtain $132558 < 151988 - S_1 \Rightarrow S_1 < 19430$. The estimate indicates that the ancestral branch **a-b** does not exceed16.5% $(19430/118181)$ of the **a-c** PLA branch. We obtained a similar estimate (17.3%) including both fixed and non-fixed alleles.

ABBA/BABA simulations

MSMS software (Ewing & Hermisson, 2010) was used to simulate different demographic models and their consequences on SNP patterns related to ABBA/BABA tests. Gene flow was simulated by adding a migration event (-m). A total of 7 demographic models were simulated, each demographic model representing gene flow between

a) HET and SIL (m(HS));

b) SIL and PLA (m(SP));

c) HET and PLA (m(HP));

d) SIL and PLA, and between HET and PLA (m(SP+HS));

e) HET and PLA, and between HET and SIL (m(HP+HS));

f) SIL and PLA, between HET and PLA, and between HET and SIL (m(HP+SP+HS)); and

g) No gene flow (m(NULL));

We assumed Ne to be 100,000, adjusted across species according to their estimated effective population sizes (-en 0.28 1 2 -en 0.28 2 1.40), and we simulated the recent decrease of population size (-en 0.002 1 0.20 -en0.002 2 0.20 -en 0.002 3 0.04, (Foote, 1995)). A total of 377,627 segregated sites were simulated, corresponding to the number of fixed genotype differences relative to D. grimshawi. Estimated divergence times from mcmctree was used to adjust the time of joint events (-ej). Average numbers of genotypes were counted based on 100 simulation cycles. The command line for m(HS) simulation was given as follows:

```
java -jar msms.jar -N 100000 -ms 3 100 -s 377627 -I 3 1 1 1 -t 1400 -r 1400 -en 0.28 1 2 -en 0.28
2 1.40  -ej 0.145  3 2  -ej 0.274 2 1   -en 0.002 1 0.20 -en 0.002 2 0.20 -en 0.002 3 0.04  -em 0.05
3 2 5  -threads 12.
```

We found the model m(HS) to be most consistent with our empirical results, thus corroborating our main argument in the paper (Appendix 3.2 & 3.3).

Tables:

Table 3.1. Frequencies of trans-species single nucleotide polymorphisms (SNPs) with their corresponding F_{ST} values (normalized by mean F_{ST} value for all three pairwise comparisons). Species order in each genotype: Drosophila *heteroneur*a (HET), *D. silvestris* (SIL), *D. planitibia* (PLA), and *D. grimshawi*.

Genotype	**Normalized F_{ST} (divided by mean F_{ST})**			
	SNPs	**HET-SIL**	**HET-PLA**	**SIL-PLA**
BAAA	71620	1.31123	1.10063	1.03658
ABAA	42187	1.39469	1.05738	1.10928
AABA	118181	1.05111	1.19512	1.15716
BBAA	132558	0.94353	1.08849	1.0768
ABBA	5969	1.33074	1.08408	1.03209
BABA	7112	1.48349	1.02171	1.09017
BBBA	3,358,959	0.98716	0.98653	0.989

Figure legends:

Fig 3.1. (**A**) Distribution of the mean number of pairwise differences (Tajima's π) and (**B**) Watterson's θ for *D. heteroneura* (red), *D. silvestris* (blue), and *D. planitibia* (green). Both π and θ values were calculated based on a window size of 10 Kb.

Fig 3.2. Site frequency spectrum of three species.

Fig 3.3. Average F_{ST} values in the neighborhoods of SNP patterns with respect to species pairwise comparisons and window sizes (N). **(A)** BBAA. **(B)** ABBA. **(C)** BABA.

References

Beissinger, T. M., Rosa, G. J. M., Kaeppler, S. M., Gianola, D., & de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genetics Selection Evolution, 47*.

Carson, H. L., K. Y. Kaneshiro, F.C. Val. (1989). Natural hybridization between the sympatric Hawaiian species Drosophila silvestris and Drosophila heteroneura. *Evolution, 43*(1), 190-203.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., . . . MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature, 450*(7167), 203-218.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi:10.1038/ng.806

Ewing, G., & Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics, 26*(16), 2064-2065. doi:10.1093/bioinformatics/btq322

Feder, J. L., Nosil, P., Wacholder, A. C., Egan, S. P., Berlocher, S. H., & Flaxman, S. M. (2014). Genome-Wide Congealing and Rapid Transitions across the Speciation Continuum during Speciation with Gene Flow. *J Hered, 105*(S1), 810-820. doi:10.1093/jhered/esu038

Foote, D., Carson, H.L. . (1995). Drosophila as monitors of change in Hawaiian ecosystems. In E. T. LaRoe, Farris, G.S., Puckett, C.E., Doran, P.D. ,Mac, M.J. (Ed.), *Our living resources: a report to the nation on the distribution, abundance, and health of U.S. plants, animals, and ecosystems.* (pp. 368–372). Washington, DC: U.S. Department of the Interior, National Biological Service.

Fu, Y. X. (1994). Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics, 138*(4), 1375-1386.

Gagnaire, P. A., Pavey, S. A., Normandeau, E., & Bernatchez, L. (2013). The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. *Evolution, 67*(9), 2483-2497. doi:10.1111/evo.12075

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., . . . Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science, 328*(5979), 710-722. doi:10.1126/science.1188021

Jonsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., . . . Orlando, L. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc Natl Acad Sci U S A, 111*(52), 18655-18660. doi:10.1073/pnas.1412627111

Kang, L., Settlage, R., McMahon, W., Michalak, K., Tae, H., Garner, H. R., . . . Michalak, P. (2016). Genomic Signatures of Speciation in Sympatric and Allopatric Hawaiian Picture-Winged Drosophila. *Genome Biol Evol, 8*(5), 1482-1488. doi:10.1093/gbe/evw095

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res, 19*(7), 1195-1201. doi:10.1101/gr.091231.109

Keller, I., Wagner, C. E., Greuter, L., Mwaiko, S., Selz, O. M., Sivasundar, A., . . . Seehausen, O. (2013). Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol Ecol, 22*(11), 2848-2863. doi:10.1111/mec.12083

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics, 27*(24), 3435-3436. doi:10.1093/bioinformatics/btr589

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., . . . Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res, 23*(11), 1817-1828. doi:10.1101/gr.159426.113

Muir, C., & Price, D. K. (2008). Population structure and genetic diversity in two species of Hawaiian picture-winged Drosophila. *Mol Phylogenet Evol, 47*(3), 1173-1180.

Nosil, P. (2008). Speciation with gene flow could be common. *Mol Ecol, 17*(9), 2103-2106. doi:10.1111/j.1365-294X.2008.03715.x

Price, J. P., & Clague, D. A. (2002). How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proc Biol Sci, 269*(1508), 2429-2435. doi:10.1098/rspb.2002.2175

Sankararaman, S., Patterson, N., Li, H., Paabo, S., & Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genet, 8*(10), e1002947.

Smadja, C. M., & Butlin, R. K. (2011). A framework for comparing processes of speciation in the presence of gene flow. *Mol Ecol, 20*(24), 5123-5140. doi:10.1111/j.1365-294X.2011.05350.x

Fig 3.1

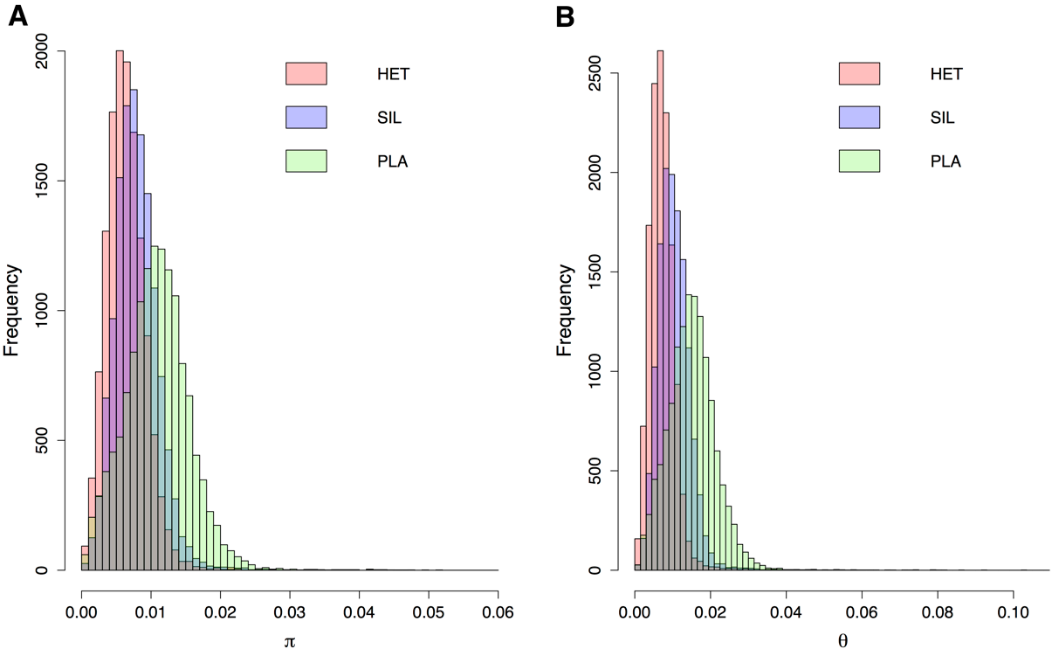

A
Frequency
2000
1500
1000
500
0
HET
SIL
PLA
0.00
0.01
0.02
0.03
0.04
0.05
0.06
π
B
Frequency
2500
2000
1500
1000
500
0
HET
SIL
PLA
0.00
0.02
0.04
0.06
0.08
0.10
θ

Fig 3.2

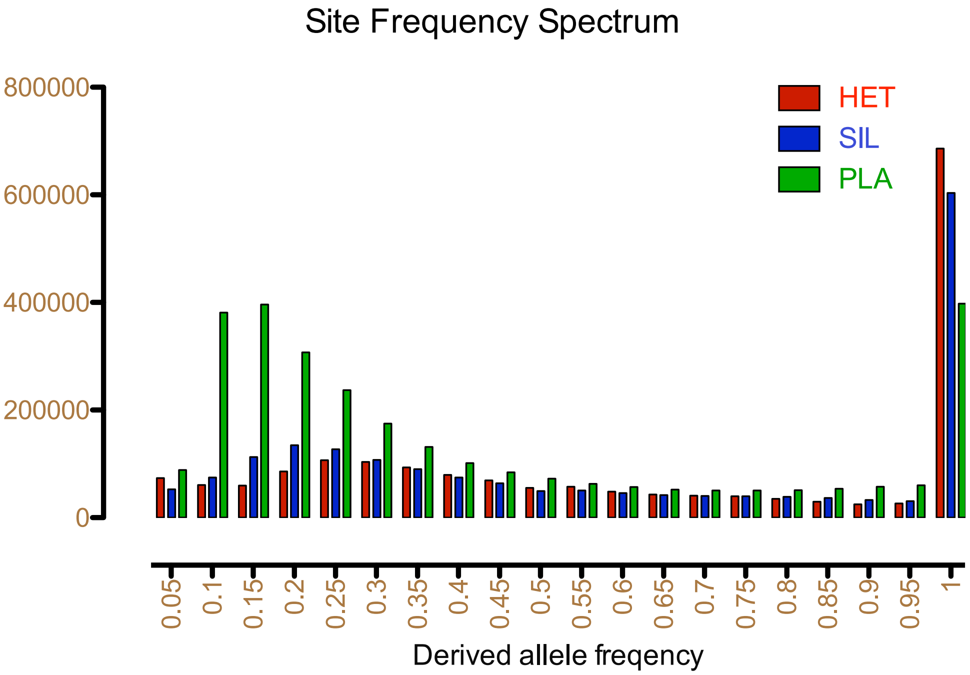
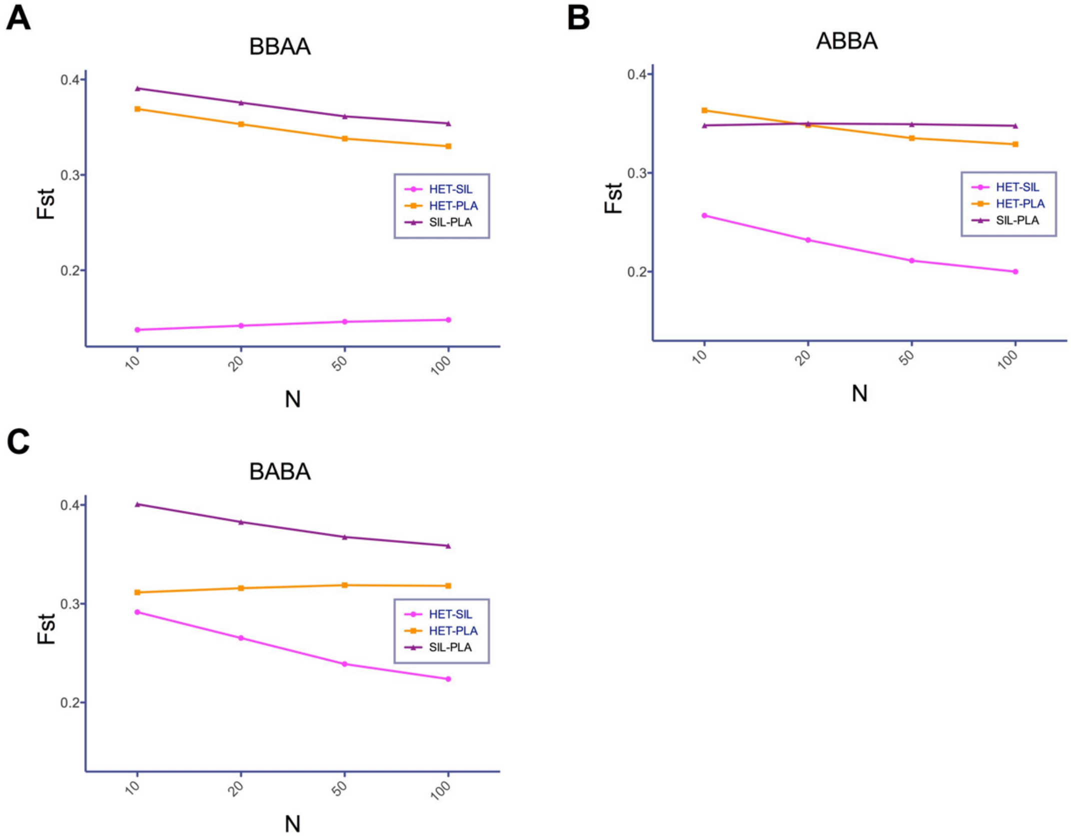
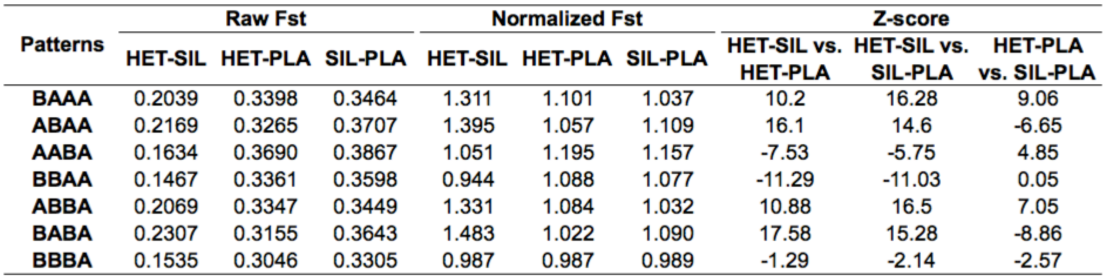

Site Frequency Spectrum
800000
600000
400000
200000
0
HET
SIL
PLA
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Derived allele freqency

Fig 3.3


A
BBAA
Fst
0.4
0.3
0.2
N
10
20
50
100
HET-SIL
HET-PLA
SIL-PLA
B
ABBA
Fst
0.4
0.3
0.2
N
10
20
50
100
HET-SIL
HET-PLA
SIL-PLA
C
BABA
Fst
0.4
0.3
0.2
N
10
20
50
100
HET-SIL
HET-PLA
SIL-PLA

Appendix

Appendix 3.1. Table of Z-scores for species-pairwise F_{ST} comparisons

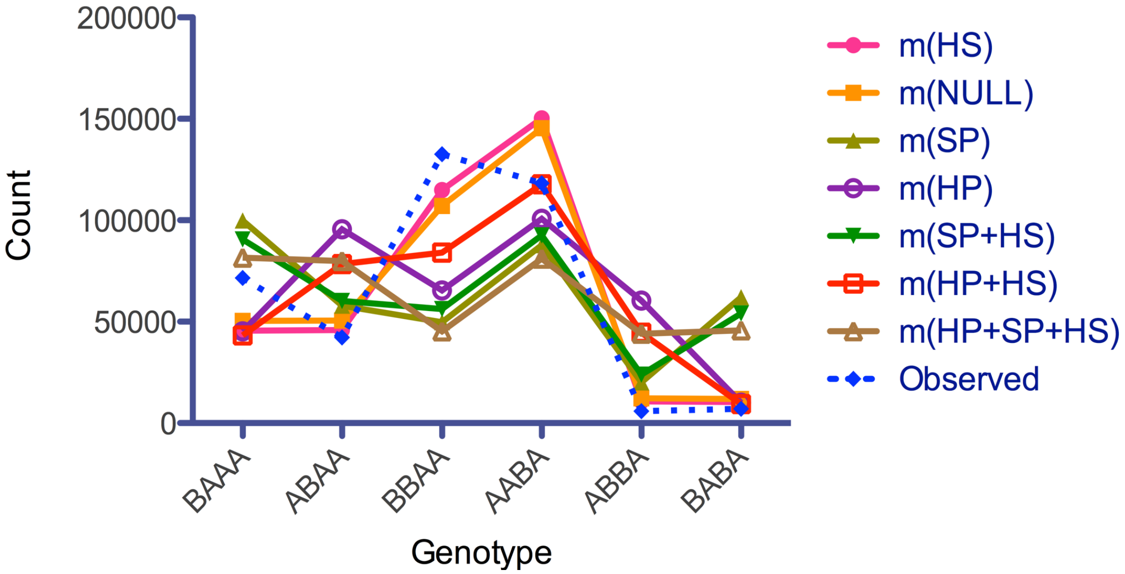


Patterns	Raw Fst			Normalized Fst			Z-score		
	HET-SIL	HET-PLA	SIL-PLA	HET-SIL	HET-PLA	SIL-PLA	HET-SIL vs. HET-PLA	HET-SIL vs. SIL-PLA	HET-PLA vs. SIL-PLA
BAAA	0.2039	0.3398	0.3464	1.311	1.101	1.037	10.2	16.28	9.06
ABAA	0.2169	0.3265	0.3707	1.395	1.057	1.109	16.1	14.6	-6.65
AABA	0.1634	0.3690	0.3867	1.051	1.195	1.157	-7.53	-5.75	4.85
BBAA	0.1467	0.3361	0.3598	0.944	1.088	1.077	-11.29	-11.03	0.05
ABBA	0.2069	0.3347	0.3449	1.331	1.084	1.032	10.88	16.5	7.05
BABA	0.2307	0.3155	0.3643	1.483	1.022	1.090	17.58	15.28	-8.86
BBBA	0.1535	0.3046	0.3305	0.987	0.987	0.989	-1.29	-2.14	-2.57

Appendix 3.2. Table for sum of genotype differences between observed and different demographic models in ABBA/BABA simulations

	m (HS)	m (NULL)	m (SP)	m (HP)	m (SP+HS)	m (HP+HS)	m (HP+SP+HS)	Observed
BAAA	45583	50407	99931	45313	90584	43252	81548	71620
ABAA	45917	50617	57984	95592	60251	78452	79888	42187
BBAA	114843	107063	49789	65468	56234	84119	45293	132558
AABA	150273	145416	87821	100695	92589	117786	81155	118181
ABBA	10670	12170	20022	60445	23838	44636	44021	5969
BABA	10340	11954	62079	10114	54132	9381	45722	7112
Sum of difference to observed	87503	93416	226259	221766	203834	154404	248581	0

Appendix 3.3. Figure for different demographic models in ABBA/BABA simulations



Chapter 4 : Mapping Genomic Scaffolds to Chromosomes Using Laser Capture Microdissection in Application to Hawaiian Picture-winged *Drosophila*

Manuscript in preparation

Lin Kang[1], Phillip George[2], Ellie E. Armstrong[3], Stefan Prost[3], Donald K. Price[4], Igor Sharakhov[2], Pawel Michalak[1]

[1] Biocomplexity Institute, Virginia Tech, Blacksburg, VA 24061, USA
[2] Department of Entomology, Virginia Tech, Blacksburg, VA 24061, USA
[3] Department of Biology, Stanford University, Stanford, CA 94305, USA
[4] School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA

Abstract

Next generation sequencing technologies have led to a decreased cost and an increased throughput in genome sequencing. Yet, many genome assemblies based on short reads have only been assembled to the scaffold level due to lack of sufficient mapping information. Traditional ways of mapping scaffolds to chromosomes require large amount of laboratory work and time to generate genetic and/or physical maps. To address this problem, we demonstrate a rapid approach that uses laser capture microdissection (LCM) for mapping scaffolds of *de novo* genome assemblies to chromosomes in Hawaiian picture-winged *Drosophila*. We isolated and sequenced intact chromosome arms from larvae of *Drosophila differens.* By mapping the reads of each chromosome to the recently assembled scaffolds from six Hawaiian *Drosophila* species, at least 67% of the scaffolds were successfully assigned with chromosome label. Even though the scaffolds are not ordered within a chromosome, the fast-generated sketchy chromosome information allows chromosome-related analyses after genome assembling. We used the assigned chromosome information to test the faster-X evolution effect, and the result supports this phenomenon in Hawaiian *Drosophila*.

Introduction

Along with the rapid development of massively parallel sequencing, or next generation sequencing (NGS), the DNA sequencing capability grows exponentially. This allows researchers to perform *de novo* genome assemblies for thousands of species from virus and insects to mammals (e.g. i5K, Genome 10K (G10KCOS, 2009; i5K Consortium, 2013)). However, the complexity of the assembly task by using shorts reads still poses a major challenge. Genome assemblies are hierarchical, and a typical *de novo* assembly process, especially for large eukaryotic genomes, often ends at the level of scaffolds, assembled from shorter assembly components, contigs (Ensembl database, http://www.ensembl.org/). With sufficient mapping information, scaffolds can be assembled into chromosomes. The mapping information may include genetic, physical, or optical maps (Chamala et al., 2013; R. A. Dean et al., 2005; Dong et al., 2013), as well as sequence data from long insert size or mate pair reads. All of these are highly laborious before the genomic location can be obtained.

An alternative approach for mapping scaffolds to chromosomes can be synteny-based, where a closely related species with assembled chromosome information is used as reference for mapping (Clark et al., 2007; Kim et al., 2013; Schaeffer et al., 2008; Stark et al., 2007), or synteny across species inferred from a comparative analysis is utilized to order the scaffolds (Aganezov, Sitdykova, Consortium, & Alekseyev, 2015; Husemann & Stoye, 2010). However, chromosome synteny alone should be treated with caution since structural and chromosomal rearrangements may have occurred even between closely related species. Assembly errors from the reference species assembly may also be transferred to the new species assembly (Ekblom & Wolf, 2014), thus affecting the accuracy of the mapping.

Alternatively, relatively quick chromosome mapping information can be obtained through the isolation of euchromatic segments from distinct chromosome arms using laser capture microdissection (LCM), coupled with whole genome amplification (WGA) and NGS (George, Sharma, & Sharakhov, 2014). Here we test this approach in application to polytenic chromosomes of Hawaiian picture-winged *Drosophila* species, an iconic example of sequential colonization, speciation, and adaptive radiation (H. L. Carson, 1997; H. L. Carson, Clague, D.A.,, 1995). Although the Hawaiian picture-winged *Drosophila* played a large role as a model system in substantiating the modern evolutionary synthesis and the Biological Species Concept (Mayr, 1963, 1982), to date genome information for this group has been limited. Only one species of Hawaiian *Drosophila*, *D. grimshawi*, was sequenced by the Drosophila 12 Genomes Consortium (Clark et al. 2007), but this genome remains unfinished and the scaffolding was partly guided by synteny with distantly related species. We have recently sequenced, pre-assembled, and analyzed genomes of three other species of Hawaiian picture-winged *Drosophila* (Kang et al., 2016), here followed by three additional species. Chromosomal variation in Hawaiian picture-winged *Drosophila* was of great taxonomical and biogeographical value, as the polytene karyotypes had been arranged into lineages reflecting a series of inversions within chromosome arms and species divergence (H. L.

Carson, 1983). Other than inversions, all these species possess six pairs (2n=12) of highly syntenic chromosomes.

Based on LCM and NGS from polytenic chromosomes, we assigned scaffolds to chromosomes from the following six species with pre-assembled genomes: *D. silvestris, D. heteroneura, D. planitibia, D. murphyi, D. ochracea, and D. sproati.* Although these scaffolds are still not localized or ordered within a chromosome, the initial chromosome information can be useful for further mapping, as well as certain population and evolutionary analyses. For the latter, we exemplify the utility of LCM-based mapping in application to a preliminary test for faster X chromosome evolution (Charlesworth, Coyne, & Barton, 1987) among the Hawaiian picture-winged *Drosophila*.

Result and discussion

Genome completeness and annotation

The pre-assembled genomes were assessed for their completeness with two methods, CEGMA (Parra, Bradnam, & Korf, 2007) and BUSCO (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) (Appendix Table 1 & 2). The numbers of contigs for the 6 assemblies ranged from 1,659 (*D. ochracea*) to 15,471 (*D. planitibia*), with the average completeness of 98.86% and 91.93% by CEGMA and BUSCO, respectively. Among the 6 species, *D. planitibia* had the largest assembled genome size of 189 Mb, while *D. silvestris* had the smallest assembled genome size of 146 Mb. The numbers of predicted protein-coding gene ranged between 10,919 (*D. heteroneura*) and 12,746 (*D. planitibia*), correlated with their genome sizes (Appendix Table 3). More than 97% of predicted gene could be annotated against the non-redundant protein database.

Chromosomes isolation and sequencing

A total of 24 chromosome-samples from 3rd instar larvae of *D. differens*, an endemic species from Molokai (sister species to *D. planitibia* from Maui), were isolated and sequenced (Fig 4.1). An average of 10,410,225 reads or 1,405,892,022 bps were generated for each sample. After a stringent quality control, including tests for inter-chromosomal contamination, as well as foreign DNA contamination, only 12 samples were used for chromosome-mapping and further analysis (Table 4.2, see Materials and

Methods). The majority of reads from each of the 12 samples were mapped to a single Muller element using *D. melanogaster* as a reference, with the ratio of the highest count and the second highest count greater than 10 (Table 4.2 and Fig 4.2).

Mapping scaffolds to chromosomes

The 12 samples were then merged into five groups corresponding to Muller elements, based on chromosomal synteny. Reads from each group were merged and mapped against assembly from each species separately. Scaffolds from each assembly were assigned a Muller element label if stringency criteria were met (see Materials and Methods). More than 67% (from 67.09% to 89.59%) scaffolds were successfully assigned with Muller element labels (Table 4.3). Merging scaffolds into LCM-derived chromosomes according to their Muller element labels, and mapping the chromosomes from each species against the *D. melanogaster* chromosomes generated a general chromosomal alignment (Fig 4.3). A large proportion of each LCM-derived chromosome successfully mapped to a single *D. melanogaster* chromosome, confirming the validity of the method, as well as high chromosomal synteny between Hawaiian *Drosophila* and *D. melanogaster*. Presumably only a small proportion of "aberrant" alignments across chromosomes does indicate chromosomal and structural rearrangement events after the split from the common ancestor with *D. melanogaster*.

A preliminary test for faster X chromosome evolution

Even though the scaffolds within LCM-derived chromosomes were not ordered, information on approximate chromosome mapping is useful for addressing a variety of questions, such as those related to sex chromosomes versus autosomes. We asked a question about the relative rate of inter-species sequence divergence of X chromosome. Faster-X evolution was proposed to occur due to selection on recessive or partially recessive mutations being more efficient on the X chromosome than on the autosomes (Charlesworth et al., 1987). We tested whether X-linked genes accumulate more substitutions than autosomal genes by calculating the Ks/Ks ratio and using *D. grimshawi* as an outgroup. The average synonymous substitution rate Ks values of X-linked genes were greater than that of autosomal genes in the five out of six species with the exception of *D. heteroneura* (Fig 4.4). As Ks is usually considered to reflect non-adaptive changes, a greater Ks of X-linked genes is expected due to the smaller effective population size (Ne) of X-linked relative to autosomal genes. The Ka/Ks values were also greater in X-linked genes compared to autosomal genes in these five species, supporting the faster-X chromosome divergence. More robust approaches, such as McDonald–Kreitman tests, will be required to further examine this phenomenon in Hawaiian *Drosophila* species, based on multiple individuals being sequenced for each species.

Discussion

Laser capture microdissection is a powerful technique that permits rapid dissection and isolation of tissues, cells, organelles, chromosomes, and their fragments. It allows accurate investigation of subcellular or tissue-specific profiles, including disease-associated ones, especially when combined with polymerase chain reaction amplification, gene expression assays (Emmert-Buck et al., 1996; Nakazono, Qiu, Borsuk, & Schnable, 2003), proteomic analysis (Xu, Caprioli, Sanders, & Jensen, 2002), enzyme recovery from LCM transferred tissue (Emmert-Buck et al., 1996), or even profiling at single-cell level (Kamme et al., 2004; Keays, Owens, Ritchie, Gilden, & Burgoon, 2005). The proposed approach of mapping scaffolds to chromosomes using laser capture microdissection is a fast and cost-efficient way for acquiring chromosomal information for draft genome assemblies, avoiding likely errors of reference- or synteny-based approaches. The LCM-based approach enables rapid downstream chromosome-related analyses, essentially along with the de novo genome assembly construction, and it can aid in chromosome mapping that integrates genetic and physical maps by pre-classifying scaffolds into different categories.

We exemplify the utility of the approach in application to a test for faster X chromosome evolution (Charlesworth, Coyne, & Barton, 1987). Although the X chromosome is usually similar to the autosomes in size and cytogenetic appearance, theoretical models predict that the hemizygosity of males may cause unusual patterns of evolution on X chromosome (Vicoso & Charlesworth, 2006). Tests for faster-X chromosome evolution in *Drosophila* and mammals have been largely inconclusive, necessitating more research (Meisel & Connallon, 2013). Early studies of the faster-X evolution in coding sequences of *Drosophila* species led to contradictory results (Betancourt, Presgraves, & Swanson, 2002; Thornton, Bachtrog, & Andolfatto, 2006; Thornton & Long, 2002), but increasingly more studies tend to support this phenomenon (Charlesworth & Campos, 2014; Veeramah, Gutenkunst, Woerner, Watkins, & Hammer, 2014). Newer studies also include a wider taxonomical spectrum, and include human, mouse, and birds (R. Dean, Harrison, Wright, Zimmer, & Mank, 2015; Kousathanas, Halligan, & Keightley, 2014; Wang, Ekblom, Bunikis, Siitari, & Hoglund, 2014; Wright et al., 2017). In addition to sequence divergence, gene expression patterns of divergence in *Drosophila* and mammals are also pronounced among X-linked genes (Hu, Eisen, Thornton, & Andolfatto, 2013; Kousathanas et al., 2014). The analysis of the relative divergence rates between X-linked genes and autosomal genes in Hawaiian *Drosophila*, a group that has undergone rapid adaptive radiation, is thus particularly of interest.

Materials and Methods

Genome assemblies

The genome assemblies of *D. silvestris, D. heteroneura* and *D. planitibia* were obtained as described elsewhere (Kang et al., 2016). That study was followed by whole genome sequencing of single non-inbred individuals from the other three species, *D. sproati, D. ochracea* and *D. murphyi*, all from Hawai'i Island. Genomic DNA extractions from these species were sent to the UC Davis Genome center for library preparation and sequencing. Gel-free mate-paired libraries and paired-end 180bp overlap libraries (20bp overlap) were selected to minimize the required input DNA amount, as well as to maximize ease of assembly. All libraries were sequenced on a single lane of the Illumina Hiseq 3000 platform using the 100bp paired-end setting. After preprocessing the raw-read data, draft genomes were assembled with Allpaths-LG (Gnerre et al., 2011), and improved with GapCloser (Luo et al., 2012) to produce N50 values of 275 Kb (*D. murphyi*), 364 Kb (*D. sproati*), and 476 Kb (*D. ochracea*). Genes and repeats in the genomes sequence were initially annotated as described previously (Kang et al., 2016).

Polytene chromosome preparation and sequencing

Intact chromosome arms were isolated from larvae of *Drosophila differens* using a published LCM protocol for isolating and amplifying the euchromatic segments of individual polytene chromosome arms (George et al., 2014). Microdissected chromosomes were lysed to release gDNA. Whole genome amplification was used to amplify the DNA from the microdissected material and to create libraries with the PicoPLEX™ DNA-seq kit. The 250bp single-end libraries were further amplified exponentially with primers containing unique Illumina dual barcodes suitable for Illumina Miseq. A total of 24 single-chromosome samples were generated (Fig 4.1). Each sample was assigned a putative chromosome ID according to its karyotype/chromosome banding pattern (H. L. Carson, 1983).

Quality control and mapping

Raw reads from sequencing were quality-controlled and filtered through FastqMcf (Aronesty, 2013). Briefly, adaptors were clipped and reads with average quality less than 30 or with more than 5 N (bad) bases were discarded. To reduce the possible contamination, reads were mapped to the bacteria database downloaded from NCBI (http://www.ncbi.nlm.nih.gov/) and unmapped reads were kept. To assess the quality of the samples, the clean reads were next mapped to *D. planitibia* (Kang et al., 2016), *D. grimshawi* genome reference (downloaded from http://flybase.org/) and human genome reference (hg19, downloaded from UCSC http://genome.ucsc.edu/) using BWA (Li & Durbin, 2009) with default parameters (Table S1). Samples with reads mapping against the human reference at a rate greater than 60% or mapping against the *D. planitibia* reference at a rate lower than 40% were removed (7 samples). To obtain more valid chromosome information, reads from the remaining 17 samples were then mapped to repeat-masked *D. melanogaster* reference (dm3). The sizes of regions with at least 5 reads mapped were counted for each *D. melanogaster* chromosome. For each sample, the ratio between the highest count and the second highest count was calculated. Samples with this ratio less than 10 were considered possible chromosome cross-contaminations and were thus discarded (there were 5 such samples, Table S2). The remaining 12 samples were then merged into 5 groups with samples sharing the same best chromosome hit against *D. melanogaster*, and synteny with Muller element arm (for *D. melanogaster*, chrX, chr2L, chr2R, chr3L, chr3R and chr4 correspond to Muller elements A, B, C, D, E and F, respectively).

Scaffold chromosome annotation

Reads of samples within the same group were combined together to maximize the data utilization. Grouped reads were mapped to draft assemblies from 6 *Drosophila* species (*D. silvestris, D. heteroneura, D. planitibia, D. murphyi, D. ochracea, D. sproati*) using BWA with default parameters, while the draft assemblies were repeat-masked before mapping by RepeatMasker (Open-4.0, http://www.repeatmasker.org). For each scaffold in an assembly, mapping information for all 5 groups are collected and the lengths of region covered by at least 5 reads from each group were counted. Scaffold was assigned with chromosome/group information if the largest length is greater than 1,000 and the second largest length is less than one tenth of the largest one (e.g. scaffold10 has 10,000bp region with each base covered by at least 5 reads from group MullerA, and it also has 500 bp region covered by at least 5 reads from group MullerB, then we assigned MullerA to scaffold10). The remaining scaffolds were labeled as unassigned. In general, more than 65% of the total size of scaffolds could be assigned with group information for each species (Table S3).

Gene Prediction and Annotation

Protein-coding genes were predicted using MAKER2 (Holt & Yandell, 2011), which used *D. melanogaster* protein sequences from FlyBase (r6.02, http://flybase.org) as protein homology evidence, and integrated with prediction methods including BLASTX, SNAP (Korf, 2004) and AUGUSTUS (Stanke & Waack, 2003). In order to increase further the accuracy of gene prediction, we included the RNA-seq data to the prediction process. RNA-seq data for *D. silvestris, D. heteroneura,* and *D. planitibia* were downloaded from NCBI Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo) under accession number GSE80124 and from our previous study (Brill, Kang, Michalak, Michalak, & Price, 2016). Adapters were removed from the raw sequencing reads. Duplicated and low-quality reads were discarded using FastqMcf (Aronesty, 2013) with default parameters. To minimize the possible contamination, reads were aligned to a NCBI bacterial database, and only unmapped reads were kept to assemble the transcriptome. The processed reads from each species were merged together and assembled with Trinity (Grabherr et al., 2011) (with parameter “--trimmomatic”). The assembled RNA-seq data were input to MAKER2 pipeline as EST evidence. Predicted genes were subsequently used as query sequences in a blastx database search of NR database (non-redundant database, http://www.ncbi.nlm.nih.gov/). Blastx alignments with e-value greater than 1e-10 were discarded, and the top hit (or top hit from *Drosophila* species if existed) was used to annotate the query genes.

Genome completeness

Two methods were used for genome completeness estimation. CEGMA (Parra et al., 2007) examines the existence of 248 core eukaryotic gene in assembly. BUSCO (Simao et al., 2015) was used to assess universal single-copy orthologs of eukaryote in the assembly.

Ka/Ks ratio

In order to reduce possible impact of gene mis-annotations on the Ka/Ks (or dN/dS) ratio, we used only annotations against Swissprot (http://www.ebi.ac.uk/uniprot). Blastx alignments with e-value greater than 1e-40 or identity less than 40% were discarded. *Drosophila grimshawi* was used as an outgroup, sequences with the same annotation were grouped together, and Clustal-omega (Sievers et al., 2011) was used to conduct the multiple sequence alignments. Nucleotide sequences were parsed to amino acid sequences before carrying multiple-sequence alignments to avoid possible frame-shifts, and the amino acid sequences of the alignment were changed back to nucleotide sequences for Ka/Ks calculations. PAML (Yang, 2007) (version 4.7) was used to calculate the pair-wise Ka/Ks ratio values, setting the model = 0 in the control file of codeml. To further minimize the possible effect of wrong annotations and grouping, Ks values greater than 2 were excluded from further analyses, and the maximal Ka/Ks value was set to be 3. Models M7/M8 along with likelihood ratio tests were applied to test for the significance of positive selection, with p-values generated from chi-square distribution (Nielsen & Yang, 1998).

Tables

Table 4.1. Quality control for mapping reads against Hawaiian Drosophila species and human genome references.

#Sample ID	NumOfReads	TotalBase	Q20(%)	Q30(%)	GC Content(%)	MappingRate		
						D. planitibia	*D. grimshawi*	Human
2*	8,643,203	1,288,795,937	97.05%	95.04%	44.21%	**14.31%**	**4.42%**	**68.26%**
3*	9,965,841	1,289,067,003	97.60%	95.91%	45.75%	**18.94%**	**8.91%**	**85.98%**
5	10,098,983	1,361,545,667	97.32%	95.43%	44.77%	**47.68%**	**39.30%**	**59.15%**
6*	10,968,595	1,484,765,509	97.36%	95.49%	44.95%	**20.48%**	**12.40%**	**60.74%**
7	10,225,920	1,255,423,051	97.66%	95.94%	45.02%	**65.61%**	**60.82%**	**30.09%**
9	9,429,412	1,238,139,403	97.43%	95.55%	44.75%	**66.36%**	**63.26%**	**22.90%**
10	8,843,571	1,172,290,243	97.46%	95.65%	44.58%	**50.27%**	**44.05%**	**46.23%**
11	10,677,556	1,334,570,886	97.71%	96.01%	44.83%	**70.56%**	**67.22%**	**20.96%**
12	9,846,903	1,270,347,221	97.48%	95.66%	44.64%	**69.53%**	**65.44%**	**31.37%**
13	11,632,471	1,626,606,120	97.29%	95.32%	43.80%	**94.76%**	**92.75%**	**17.29%**
14	10,494,444	1,393,026,007	97.49%	95.68%	44.57%	**66.80%**	**62.39%**	**30.80%**
15	12,056,935	1,506,469,114	97.46%	95.59%	45.39%	**79.14%**	**74.65%**	**25.88%**
16	10,244,193	1,303,887,668	97.50%	95.66%	44.71%	**86.22%**	**84.03%**	**19.49%**
17	10,747,167	1,403,344,374	97.58%	95.80%	44.00%	**76.73%**	**73.18%**	**21.06%**
19	10,866,170	1,444,046,227	97.40%	95.53%	44.10%	**70.38%**	**65.89%**	**29.37%**
20	9,957,740	1,284,186,408	97.58%	95.76%	44.45%	**90.78%**	**88.51%**	**17.38%**
21*	13,352,093	2,076,865,360	96.97%	94.81%	40.66%	**4.25%**	**4.22%**	**6.30%**
22	14,159,655	1,918,046,964	97.32%	95.41%	44.29%	**62.48%**	**57.04%**	**36.67%**
23*	9,501,479	1,430,814,118	97.03%	95.01%	44.01%	**28.27%**	**18.53%**	**69.33%**
24	10,323,650	1,487,264,963	97.30%	95.35%	45.45%	**46.02%**	**41.10%**	**33.83%**
25*	9,922,001	1,355,634,235	97.47%	95.65%	45.58%	**14.68%**	**7.80%**	**51.13%**
26*	8,751,880	1,339,402,154	97.04%	95.03%	44.19%	**22.83%**	**12.27%**	**73.06%**
27	8,855,911	1,246,340,475	97.14%	95.15%	44.31%	**53.86%**	**46.03%**	**50.66%**
28	10,279,633	1,230,529,432	97.72%	96.02%	44.84%	**50.88%**	**46.16%**	**31.02%**

*: filtered sample

Table 4.2. Count of mapping for each sample against masked *D. melanogaster* genome reference.

ID	1st Best		2nd Best		Ratio^	Group
	CHR	Count	CHR	Count		
5	chr2R	207475	chr3R	1482	140.00	MullerC
12	chr2R	255165	chr2L	1139	224.03	
16	chr2R	286372	chr2L	1073	266.89	
28	chr2R	129207	chr3R	2105	61.38	
7	chr3L	364033	chr3R	1128	322.72	MullerD
10	chr3L	212133	chr3R	767	276.57	
11	chr3L	332434	chr2L	2182	152.35	
14*	chr3L	382687	chr3R	87431	4.38	
9	chr2L	231771	chr3R	836	277.24	MullerB
19	chr2L	395533	chr3L	9116	43.39	
15	chrX	251583	chr2R	1463	171.96	MullerA
27	chrX	225643	chr2L	937	240.81	
20	chr3R	416570	chrX	3482	119.64	MullerE
13*	chr3R	284731	chr2L	237464	1.20	
17*	chr3R	188318	chr2L	112463	1.67	
22*	chr2L	191315	chr3R	128616	1.49	
24*	chr2R	95415	chr2L	51737	1.84	

*: discarded sample;

^: Count of the best divided by count of the second best.

Table 4.3. Summary of scaffolds group information assignment for Hawaiian *Drosophila* species.

Species	**Size of assmebly (Mb)**	**Size of assigned scaffolds (Mb)**	**%**
D. silvestris	145.88	122.85	85.84%
D. heteroneura	146.62	122.94	83.14%
D. planitibia	188.99	129.39	67.09%
D. murphyi	164.95	147.78	89.59%
D. ochracea	162.55	139.18	85.62%
D. sproati	163.64	141.36	86.38%
D. grimshawi	207.72	144.76	69.69%

Figure legends:

Fig 4.1. Examples of microscopic image of LCM chromosomes (A-C).

Fig 4.2. Mapping of reads from LCM samples against *D. melanogaster* chromosomes.

Fig 4.3. Circos diagrams showing genomic synteny of Hawaiian *Drosophila* with *D. melanogaster*. Links were created based on the alignment results between *D. heteroneura* (**A**), *D. silvestris* (**B**), *D. planitibia* (**C**), *D. murphyi* (**D**), *D. ochracea* (**E**), *D. sproati* (**F**), and *D. melanogaster* (reference in **A-F**) using Blastn, and alignments spanning less than 100 bp were discarded.

Fig 4.4. Relative rate of Ks and Ka/Ks between autosomal genes and X-linked genes.

Reference

Aganezov, S., Sitdykova, N., Consortium, A. G. C., & Alekseyev, M. A. (2015). Scaffold assembly based on genome rearrangement analysis. *Comput Biol Chem, 57*, 46-53. doi:10.1016/j.compbiolchem.2015.02.005

Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal, 7*, 1-8. doi:10.2174/1875036201307010001

Brill, E., Kang, L., Michalak, K., Michalak, P., & Price, D. K. (2016). Hybrid sterility and evolution in Hawaiian Drosophila: differential gene and allele-specific expression analysis of backcross males. *Heredity (Edinb)*. doi:10.1038/hdy.2016.31

Carson, H. L. (1983). Chromosomal sequences and interisland colonizations in hawaiian Drosophila. *Genetics, 103*(3), 465-482.

Chamala, S., Chanderbali, A. S., Der, J. P., Lan, T., Walts, B., Albert, V. A., . . . Barbazuk, W. B. (2013). Assembly and validation of the genome of the nonmodel basal angiosperm Amborella. *Science, 342*(6165), 1516-1517. doi:10.1126/science.1241130

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., . . . Consortium, D. G. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature, 450*(7167), 203-218. doi:10.1038/nature06341

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., . . . Birren, B. W. (2005). The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature, 434*(7036), 980-986. doi:10.1038/nature03449

Dong, Y., Xie, M., Jiang, Y., Xiao, N., Du, X., Zhang, W., . . . Wang, W. (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotechnol, 31*(2), 135-141. doi:10.1038/nbt.2478

George, P., Sharma, A., & Sharakhov, I. V. (2014). 2D and 3D Chromosome Painting in Malaria Mosquitoes. *Journal of Visualized Experiments : JoVE*(83), 51173. doi:10.3791/51173

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol, 29*(7), 644-652. doi:10.1038/nbt.1883

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics, 12*, 491. doi:10.1186/1471-2105-12-491

Husemann, P., & Stoye, J. (2010). r2cat: synteny plots and comparative assembly. *Bioinformatics, 26*(4), 570-571. doi:10.1093/bioinformatics/btp690

Kang, L., Settlage, R., McMahon, W., Michalak, K., Tae, H., Garner, H. R., . . . Michalak, P. (2016). Genomic Signatures of Speciation in Sympatric and Allopatric Hawaiian Picture-Winged Drosophila. *Genome Biol Evol, 8*(5), 1482-1488. doi:10.1093/gbe/evw095

Kim, J., Larkin, D. M., Cai, Q., Asan, Zhang, Y., Ge, R. L., . . . Ma, J. (2013). Reference-assisted chromosome assembly. *Proc Natl Acad Sci U S A, 110*(5), 1785-1790. doi:10.1073/pnas.1220349110

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Meisel, R. P., & Connallon, T. (2013). The faster-X effect: integrating theory and data. *Trends Genet, 29*(9), 537-544. doi:10.1016/j.tig.2013.05.009

Nielsen, R., & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics, 148*(3), 929-936.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23*(9), 1061-1067. doi:10.1093/bioinformatics/btm071

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O'Grady, P. M., . . . Kaufman, T. C. (2008). Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics, 179*(3), 1601-1655.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., . . . Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol, 7*, 539. doi:10.1038/msb.2011.75

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212. doi:10.1093/bioinformatics/btv351

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., . . . Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature, 450*(7167), 219-232. doi:10.1038/nature06340

Vicoso, B., & Charlesworth, B. (2006). Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet, 7*(8), 645-653. doi:10.1038/nrg1914

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Fig 4.1

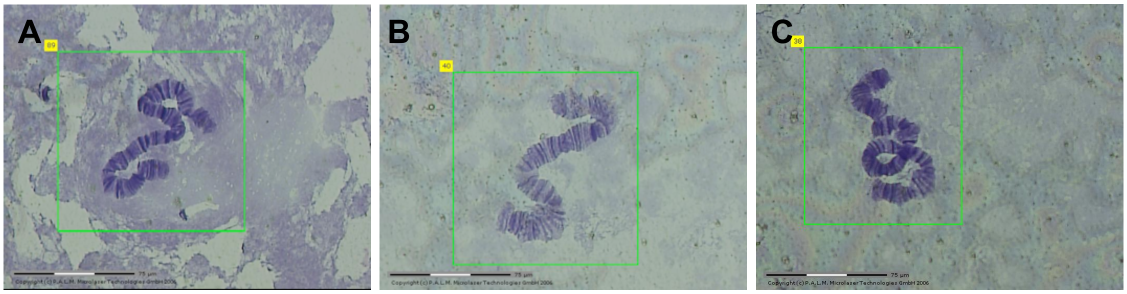

A
89
75 µm
Copyright (c) P.A.L.M. Microlaser Technologies GmbH 2006
B
40
75 µm
Copyright (c) P.A.L.M. Microlaser Technologies GmbH 2006
C
38
75 µm
Copyright (c) P.A.L.M. Microlaser Technologies GmbH 2006

Fig 4.2

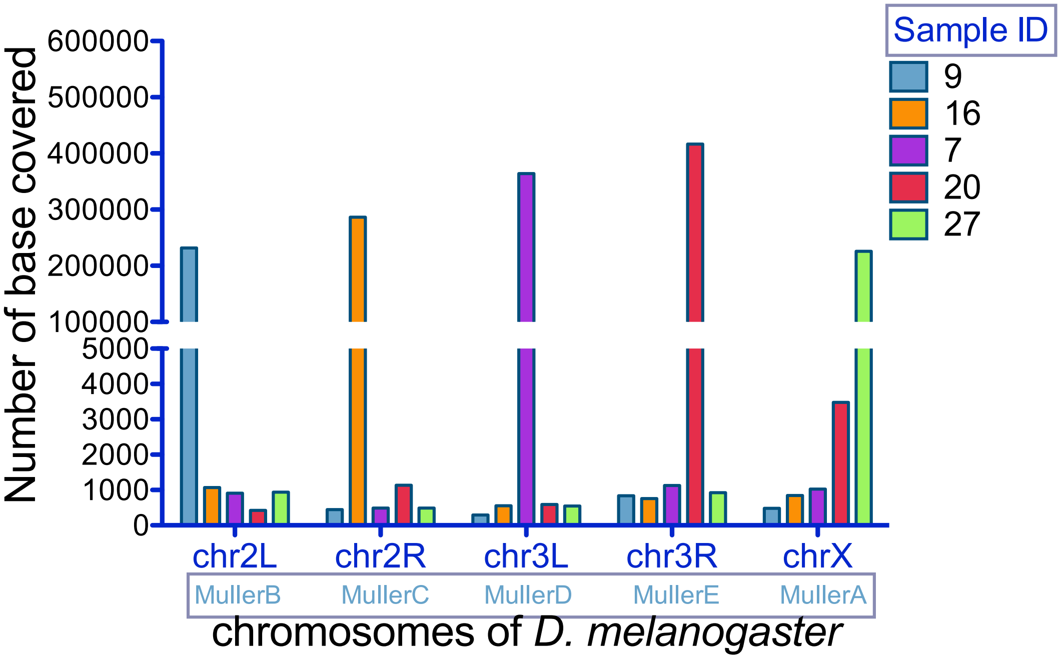

Number of base covered
600000
500000
400000
300000
200000
100000
5000
4000
3000
2000
1000
0
chr2L
chr2R
chr3L
chr3R
chrX
MullerB
MullerC
MullerD
MullerE
MullerA
chromosomes of *D. melanogaster*
Sample ID
9
16
7
20
27

Fig 4.3

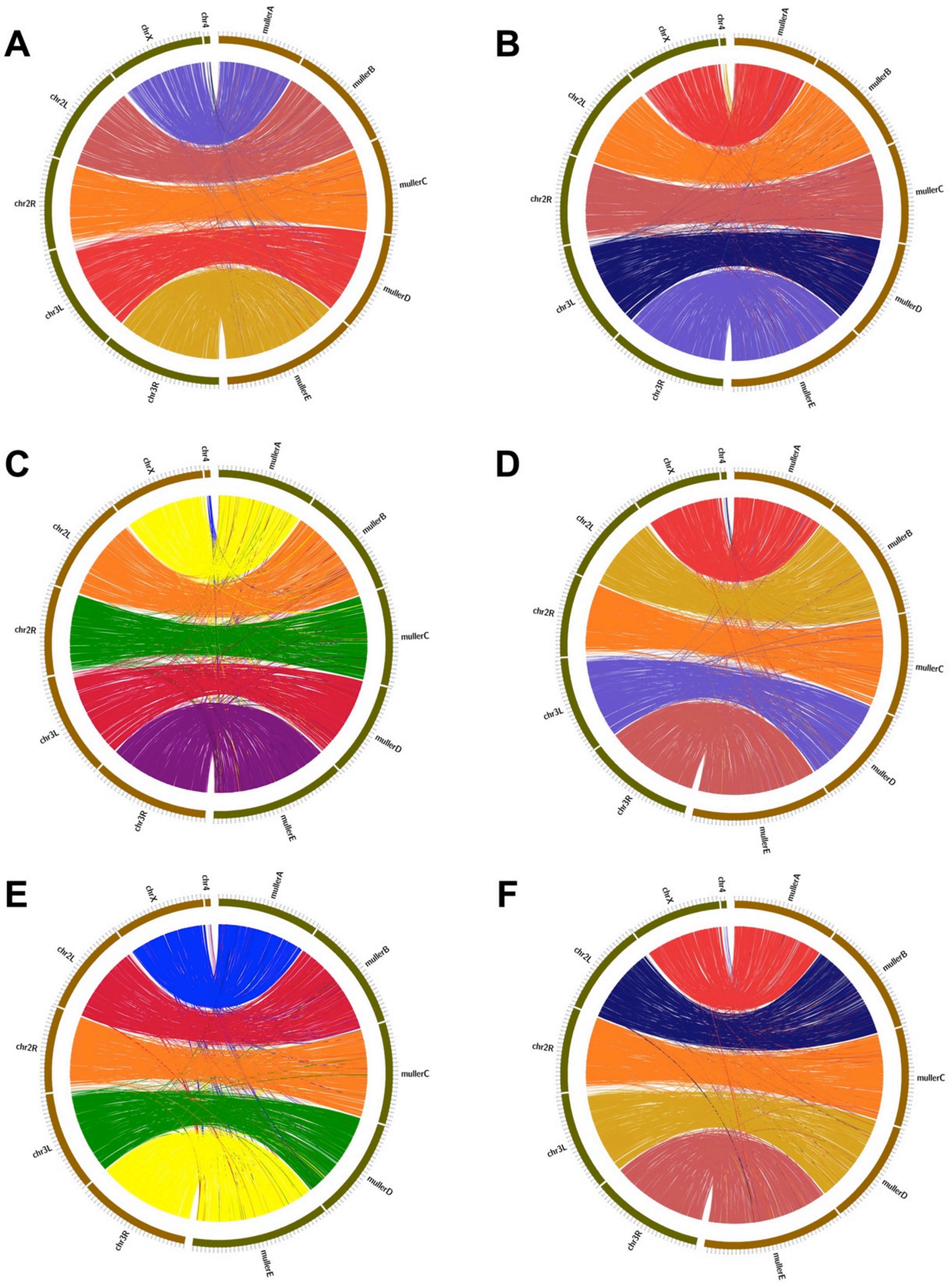
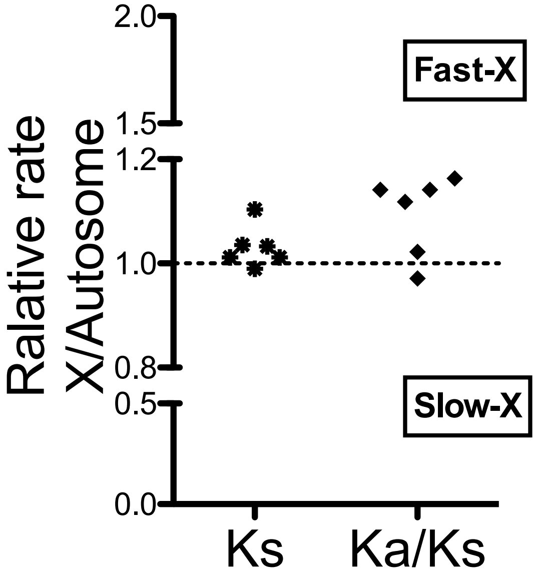
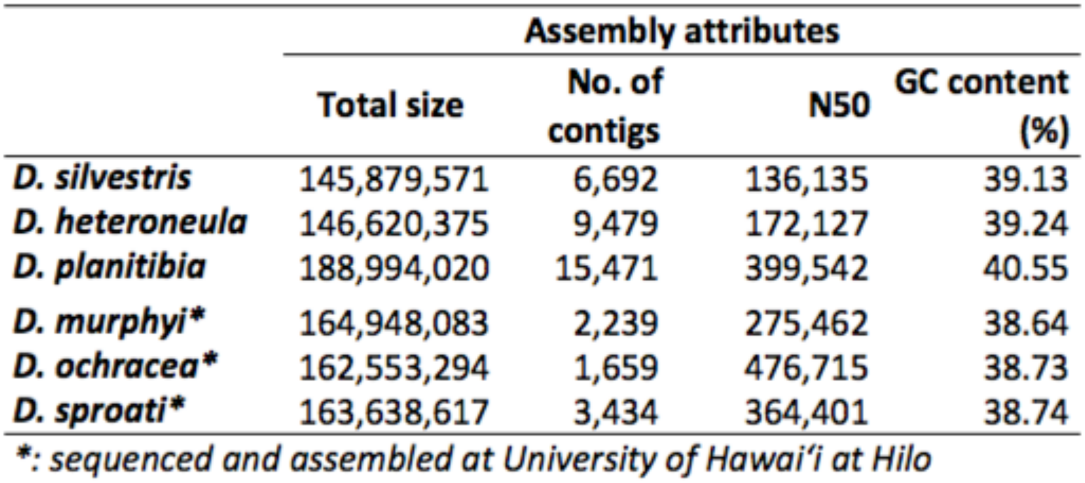

A
chr4
mullerA
chrX
mullerB
chr2L
mullerC
chr2R
mullerD
chr3L
chr3R
mullerE
B
chr4
mullerA
chrX
mullerB
chr2L
mullerC
chr2R
mullerD
chr3L
chr3R
mullerE
C
chr4
mullerA
chrX
mullerB
chr2L
mullerC
chr2R
mullerD
chr3L
chr3R
mullerE
D
chr4
mullerA
chrX
mullerB
chr2L
chr2R
mullerC
chr3L
mullerD
chr3R
mullerE
E
chr4
mullerA
chrX
mullerB
chr2L
chr2R
mullerC
chr3L
mullerD
chr3R
mullerE
F
chr4
mullerA
chrX
mullerB
chr2L
chr2R
mullerC
chr3L
mullerD
chr3R
mullerE

Fig 4.4


Ralative rate
X/Autosome
2.0
1.5
1.2
1.0
0.8
0.5
0.0
Fast-X
Slow-X
Ks
Ka/Ks

Appendix

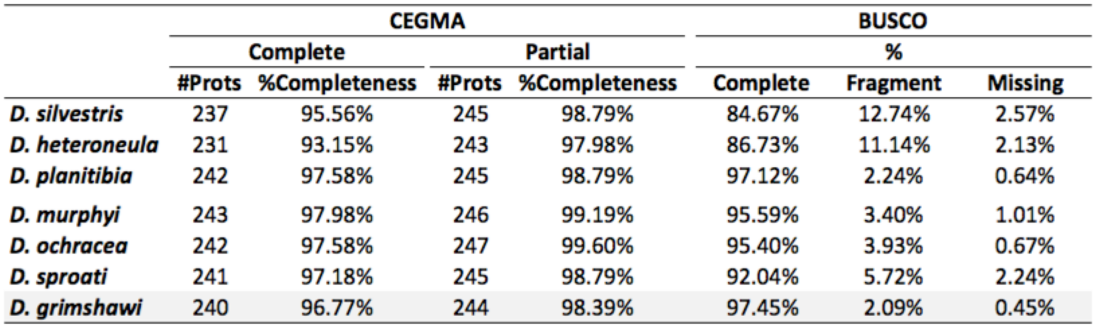
Appendix Table 1. Assembly summary of 6 Hawaiian Drosophila.



	Assembly attributes			
	Total size	No. of contigs	N50	GC content (%)
D. silvestris	145,879,571	6,692	136,135	39.13
D. heteroneula	146,620,375	9,479	172,127	39.24
D. planitibia	188,994,020	15,471	399,542	40.55
D. murphyi*	164,948,083	2,239	275,462	38.64
D. ochracea*	162,553,294	1,659	476,715	38.73
D. sproati*	163,638,617	3,434	364,401	38.74

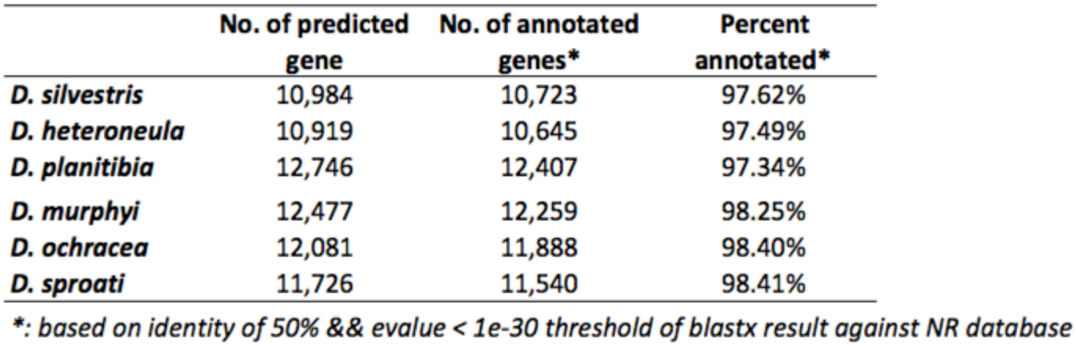
**: sequenced and assembled at University of Hawai'i at Hilo*

Appendix Table 2. Assessment of genome completeness of 6 Hawaiian Drosophila by two methods.



	CEGMA				BUSCO		
	Complete		Partial		%		
	#Prots	%Completeness	#Prots	%Completeness	Complete	Fragment	Missing
D. silvestris	237	95.56%	245	98.79%	84.67%	12.74%	2.57%
D. heteroneula	231	93.15%	243	97.98%	86.73%	11.14%	2.13%
D. planitibia	242	97.58%	245	98.79%	97.12%	2.24%	0.64%
D. murphyi	243	97.98%	246	99.19%	95.59%	3.40%	1.01%
D. ochracea	242	97.58%	247	99.60%	95.40%	3.93%	0.67%
D. sproati	241	97.18%	245	98.79%	92.04%	5.72%	2.24%
D. grimshawi	240	96.77%	244	98.39%	97.45%	2.09%	0.45%

Appendix Table 3. Gene prediction and annotation summary of 6 Hawaiian *Drosophila*.



	No. of predicted gene	No. of annotated genes*	Percent annotated*
D. silvestris	10,984	10,723	97.62%
D. heteroneula	10,919	10,645	97.49%
D. planitibia	12,746	12,407	97.34%
D. murphyi	12,477	12,259	98.25%
D. ochracea	12,081	11,888	98.40%
D. sproati	11,726	11,540	98.41%

****: based on identity of 50% && evalue < 1e-30 threshold of blastx result against NR database***

Chapter 5 : Hybrid sterility and evolution in Hawaiian *Drosophila*: Differential gene and allele-specific expression analysis of backcross males

Eva Brill[1§], Lin Kang[2§], Katarzyna Michalak[2], Pawel Michalak[2], Donald K. Price[1]

[1]Tropical Conservation Biology and Environmental Science Graduate Program,
University of Hawai`i at Hilo, Hilo, HI 96720, USA

[2] Biocomplexity Institute, Virginia Tech, Blacksburg, VA 24061, USA

[§]These authors contributed equally to the paper

Heredity 117, 100-108 (August 2016) | doi:10.1038/hdy.2016.31

Abstract

The Hawaiian *Drosophila* is an iconic example of sequential colonization, adaptive radiation and speciation on islands. Genetic and phenotypic analysis of closely related species pairs that exhibit incomplete reproductive isolation can provide insights into the mechanisms of speciation. *D. silvestris* from Hawai'i Island and *D. planitibia* from Maui are two closely-related allopatric Hawaiian picture-winged *Drosophila* that produce sterile F_1 males but fertile F_1 females, a pattern consistent with Haldane's rule. Backcrossing F_1 hybrid females between these two species to parental species gives rise to recombinant males with three distinct sperm phenotypes despite a similar genomic background: motile sperm, no sperm (sterile), and immotile sperm. We found that these three reproductive morphologies of backcross hybrid males produce divergent gene expression profiles in testes, as measured with RNA sequencing. There were a total of 71 genes significantly differentially expressed between backcross males with no sperm compared to those backcross males with motile sperm and immotile sperm, but no significant differential gene expression between backcross males with motile sperm and backcross males with immotile sperm. All of these genes were underexpressed in males with no sperm, including a number of genes with previously known activities in adult testes. An allele-specific expression analysis showed overwhelmingly more *cis*-divergent than *trans*-divergent genes, with no significant difference in the ratio of *cis*- and *trans*-divergent genes amongst the sperm phenotypes. Overall, the results indicate that the regulation of gene expression involved in sperm production likely diverged relatively rapidly between these two closely-related species.

Key words: *Drosophila planitibia*, *Drosophila silvestris*, Hawai'i, reproductive isolation, gene expression, hybrid male sterility

Introduction

Reduced hybrid fitness in the form of hybrid sterility can play an important role in speciation by acting as a post-zygotic isolating barrier (Coyne and Orr 2004). Haldane (Haldane 1922) first documented that in crosses between species sterility is more likely to occur in hybrid individuals of the heterogametic sex. Since then this observation has been shown to occur in almost all animals, especially in *Drosophila* spp., and is called "Haldane's rule" (Coyne and Orr 2004). In *Drosophila* species, hybrid male sterility (HMS) has been shown to function as an evolutionarily early limiting factor to introgression between species and consequently maintain species integrity (Noor and Feder 2006). Hybrid incompatibility and sterility are thought to occur when epistatic interactions of alleles from different species are dysfunctional (Johnson 2000), creating incompatible developmental pathways, ecological detriments, or altered (and therefore, unsuccessful) mating behavior (Coyne and Orr 2004). Genetic models have been developed that indicate that this type of fitness reduction may require at least two genetic changes – one from each species – but can be much more complex and be the result of multiple gene interactions (Johnson 2000, Coyne and Orr 2004). Multiple studies of HMS in the well-studied allopatric species pair *D. mauritiana* and *D. simulans* showed interactions of at least three genes or more (Johnson 2000), and the identification of over 100 genes contributing to HMS on the X chromosome (Wu, Johnson et al. 1996), as well as many genes on the autosomes (Tao, Zeng et al. 2003, Araripe, Montenegro et al. 2010, Dickman and Moehring 2013).

In addition to extensive X- and autosomal HMS loci studies, gene expression studies in *Drosophila* have shown a number of spermatogenesis genes differentially expressed between hybrids and parental species (Landry, Hartl et al. 2007). Theoretical models suggest that the effects of accumulation of regulatory incompatibilities in the architecture of transcriptional networks can be a part of hybrid incompatibilities directly influencing the process of speciation (Porter and Johnson 2002, Johnson and Porter 2007). Techniques such as microarrays have been used to characterize gene expression, and they have uncovered several genes that are deregulated in hybrids between species, especially in *D. mauritiana* and *D. simulans* (Michalak and Noor 2003, Moehring, Teeter et al.

2007). *Drosophila mauritiana/D. simulans* sterile male F_1 hybrids are more likely to succumb to down-regulation of genes associated specifically with male reproduction, inferring a possible genetic cause to their sterility (Michalak and Noor 2003, Michalak and Noor 2004). Closely related species pairs can provide further insights into the genetic mechanisms of speciation due to incomplete reproductive isolation in the form of sterile F_1 males but fertile F_1 females that can be backcrossed to parental species. Analyzing the differences in gene expression amongst backcrossed (BC) individuals will provide insights into the variation in expression of testis-specific genes and also into potential candidate genes that lead to reproductive isolation, by comparing transcriptomes between sterile and fertile BC siblings. Such a comparison between backcross hybrids having similar genomic backgrounds but distinct fertility phenotypes provides a higher resolution of the association between genes and phenotypes than that between F_1 hybrids and parental species, as every subsequent generation of backcrossing on average halves the amount of HMS-unrelated heterozygosity (Michalak and Noor, 2004).

HMS is generally observed to occur earlier in the divergence of species (i.e. more closely related species) than hybrid inviability, suggesting stronger selection on sterility-causative genes (Orr, Masly et al. 2004). DNA sequence divergence and expression levels of sex-related genes in many studies support the idea that genes involved in male fertility diverge faster between species than other types of genes (Orr, Masly et al. 2004). Hybrid sterility-causative genes, such as *Odysseus site homeobox* (*OdsH*) in *Drosophila* (Ting, Tsaur et al. 1998), *Meisetz* (*Prdm9*) in mice (Oliver, Goodstadt et al. 2009), and *AEP2/OLI1* in yeast (Lee, Chou et al. 2008), are often characterized by rapid sequence evolution and distinct expression patterns. Candidate genes for speciation, therefore, include genes responsible for spermatogenesis, sperm motility, and other genes that cause reproductive incompatibilities in hybrids.

HMS is an important post-zygotic reproductive isolating mechanism in many eukaryotic organisms, but only recently has genome-wide gene expression analysis been used to investigate the full suite of genes involved in the expression of this complex trait (Gomes and Civetta 2015). Here we analyze the differential gene expression and allele-specific expression of testis-level fertility in BC males between two closely related endemic Hawaiian *Drosophila* species in the picture-wing clade under *planitibia* group and IVβ subgroup, *D. planitibia* and *D. silvestris* (Spieth 1986). *D. silvestris* is endemic to Hawai'i Island and *D. planitibia* is endemic to the island of Maui, and they diverged about 0.7 Mya (O'Grady, Lapoint et al. 2011, Magnacca and Price 2015). Both species are bark breeders, whereby females oviposit eggs on and larvae develop in the decaying bark of the endemic Hawaiian flowering plant, *Clermontia* spp. ('oha wai) (Magnacca, Foote et al. 2008). The subgroup is known for selecting leks in more open spaces as opposed to other subgroups that prefer a hidden location. Current phylogenetic analyses show that *D. planitibia* shares a close ancestor with both *D. silvestris* and *D. heteroneura* as the latter two species were established on Hawai'i Island (Magnacca and Price 2015).

Previous studies have shown that successful mating can occur between the two species, creating sterile F_1 hybrid males, but fertile F_1 hybrid females (Craddock, 1974). Fertile F_1 females can be used to create a backcross generation whereby males exhibit different sperm phenotypes. Using the RNA-seq platform and a *de novo* transcriptome assembly, we identified 71 differentially expressed genes across three BC phenotypes, showing a clear underexpression of key functional genes in BC individuals that lack sperm compared with those individuals who possess large numbers of motile sperm. These results demonstrate a directional gene expression change correlated with HMS, providing important insight to the mechanisms of reproductive isolation.

Materials and Methods

Drosophila stocks

The *D. planitibia* and *D. silvestris* populations used in this study as parental population were initiated from individuals recently captured from wild populations: ~20 *D. planitibia* individuals captured in Waikamoi Preserve, east Maui (GPS coordinates 20.811286, -156.241901), in December 2012, and ~20 *D. silvestris* individuals collected in South Kona Forest Reserve, Kukuiopae unit (GPS coordinates 19.297281, -155.811710), Hawai'i Island in June 2013. All parental, F_1 and BC populations were maintained in a controlled-environment room maintained at a constant 18°C, 70% RH with a 12:12 light:dark cycle. This is a standard environment for rearing Hawaiian *Drosophila* (Uy, LeDuc et al. 2015) with a generation time of approximately 3-4 months. Adults were housed in 4-L glass jars with a damp fine sand floor and containing 3 to 5 vials of adult food. The 25x95mm adult food vials contained a yeastless Wheeler-Clayton medium and a small tissue moistened with a tea made with the leaves of *Clermontia* spp. on which females oviposited as described by Carson (Carson 1987). Adult food vials in which females deposited eggs on tissue were replaced weekly and placed in larvae-rearing trays. Larvae food, a yeast-cornmeal-molasses medium, was provisioned to each vial with active larvae several times per week. At four weeks maturing larvae were transferred to emergence jars containing moistened large-grain sand for burial pupation. Emerging adult flies were aspirated from emergence jars twice a week and separated into male or female jars to ensure that flies were virgins for future experiments.

Adult *D. silvestris* and *D. planitibia* were used to produce F_1 and BC offspring were approximately 14-21 day post-eclosion when the flies reach reproductive maturity and were placed in mating vials (28.5x95mm) that contained adult food and a tissue moistened with *Clermontia* spp. tea. Each cross was labeled, dated, and numbered to ensure identification. Adults were transferred to new mating vials each week to ensure ideal mating and egg-laying conditions. The old mating vials were placed in larvae trays labeled with type of cross, date, and identification number and fed larvae food. As larvae reached the 3rd instar the larval vials were placed in 2-L pupation/emergence jars, and emerging adult male and female flies were removed weekly from the pupation/emergence jars, prior to becoming sexually mature, and they were segregated into jars by each species or hybrid and sex. Hybrids and backcrosses were attempted during parental generations F_4-F_7. The production of SPS BC individuals was accomplished by mating F_1 females of parental cross SP (*D. silvestris* females x *D. planitibia* males) to emerged *D. silvestris* males of the parental species after flies reached sexual maturity (Fig 5.1). Due to the difficult nature of obtaining hybrids and backcrosses, all BC males used in this study were full siblings from one successful F_1 hybrid female/*D. silvestris* parental male pair.

Testes dissection and RNA collection

D. planitibia, *D. silvestris,* F_1 hybrid males, and BC adult male flies that had reached sexual maturity (4 weeks post-eclosion) were dissected at room temperature under similar housing and stabilization conditions using sterilized Dumont No. 5 fine forceps, tungsten needles, and a 100x20mm glass dissection dish that were sprayed with RNaseZAP™ (Sigma-Aldrich) before each dissection to remove RNase contamination. Males were housed in their own individual glass vials with food and dissected after a 24-hour stabilization period in the laboratory. Testes were dissected under a compound light microscope. One testis and its accessory glands were bifurcated and placed into a RNase-free 2mL round-bottomed tube (Eppendorf) filled with RNALater™ (LifeTechnologies). The other testis was prepared for a live testis squash by placing the tissue on a 20x20mm cover slip with a 10uL drop of testis buffer and covered with a clean glass slide. The prepared slide was inverted and placed under a compound light microscope at 10x and 40x magnification for observation.

BC males were categorized into three phenotype groups based on testis and sperm observations and cross-referenced with Craddock's (Craddock 1974) classifications: Males that exhibited complete absence of mature sperm (BC-NS) and were thus sterile, males that had sperm that was non-motile and often clumped inside the testes (BC-NM), and males that had motile sperm (BC-MS). Due to the low numbers of mature BC males in the laboratory, we did not have extra BC males to conduct fertility experiments to determine the ability of BC-MS and BC-NM males to father progeny.

Sample preparation and sequencing

Twenty-four testis samples were prepared for RNA extraction: 3 *D. planitibia*, 3 *D. silvestris*, 2 *D. silvestris* x *D. planitibia* F_1 hybrids, 7 BC-MS, 4 BC-NM, and 5 BC-NS. RNA extraction and sequencing were conducted at the Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA and University of Hawai‘i at Hilo. Total RNA was extracted using Trizol Reagent (Life Technologies) following the manufacturer's instructions. Using TruSeq RNA sample preparation kit (Illumina, FC-122-1001/1002), mRNA from 1 μg of total RNA with RIN ≥ 8.0 was converted into a library of template molecules suitable for subsequent cluster generation and sequencing with Illumina HiSeq

2500. The libraries generated were validated using Agilent 2100 Bioanalyzer and quantitated using Quant-iT dsDNA HS Kit (Invitrogen) and qPCR. Individually indexed cDNA libraries were pooled, clustered onto a flow cell using Illumina's TruSeq SR Cluster Kit v3 (GD-401-3001), and sequenced 101 cycles using two TruSeq SBS Kit -HS (FC-401-1002) on HiSeq 2500.

De novo transcriptome assembly

Adapter sequences were removed from the raw sequencing reads. Duplicated and low-quality reads were discarded using FastqMcf (Aronesty 2013) with default parameters. To exclude possible contamination, all reads were aligned to a bacterial database downloaded from NCBI, and only unmapped reads were used to assemble the transcriptome. The processed reads from all samples were merged together and assembled with Trinity (Grabherr, Haas et al. 2011) (with parameter --trimmomatic). TransDecoder (Broad institute) was used to identify candidate coding regions within assembled transcripts, and transcripts with open reading frame (ORF) lengths less than 300 (100 amino acids) were filtered out. The final transcriptome assembly was used as a reference for gene annotation and expression calculation.

Gene annotation

Transcripts/genes were mapped to NR database (Non-redundant protein database from NCBI) using BLAST (v. 2.2.28). Alignments with threshold e-values greater than 1e-20 or identity less than 50% were discarded. E-values determine significant matches to the database by describing the number of hits by chance. The top hit from *Drosophila* species was used to annotate the query genes (giving priority to *Drosophila melanogaster*, if present), and genes without *Drosophila* hits were discarded to further exclude possible contamination.

Gene expression

The clean reads were mapped to the reference assembly using Bowtie (Langmead, Trapnell et al. 2009) (v. 1.0.0) with parameters set to '-l 25 -I 1 -X 1000 -a -m 200' (seed length 25, the minimum and maximum insert sizes for paired-end alignment of 1 and 1000, respectively, and report alignments less than 200). RSEM (Li and Dewey 2011) was used to calculate the gene expression with default parameters. The differential expression of genes was calculated using the DEseq (Anders and Huber 2010) package in R software (http://www.r-project.org/), with Benjamini-Hochberg adjusted p-values less than 0.05 considered to be significant.

Allele-specific expression (ASE)

RNA-Seq reads were mapped to the reference transcriptome using BWA (v. 0.7.6) with default parameters. Genotypes were identified by UnifiedGenotyper from GATK package (McKenna, Hanna et al. 2010) with default parameters except setting heterozygosity to 0.01 and indel heterozygosity to 0.00125. Genotypes with genotyping quality less than 40 or mapping quality less than 30, or covered depth less than 10 were discarded. Only homozygous species-specific loci from parental *D. planitibia* and *D. silvestris* samples were chosen as ASE candidate sites. To reduce the bias of reference preference during the mapping process, a masked reference was generated by changing the reference to a third genotype different from the homozygous genotypes in *D. planitibia* and *D. silvestris* in all ASE candidate positions. Genotypes for other samples were generated as above, using the masked sequence as reference. Loci of heterozygous genotypes in F_1 and BC samples with a minimal genotyping score of 40, mapping score of 30 and coverage depth of 10 were tested with a binomial test, and candidates with Benjamini-Hochberg adjusted p-values less than 0.05 were kept. Simulated reads around the candidate ASE sites were generated with half of the reads carrying the *D. silvestris* genotype and the other half carrying the *D. planitibia* genotype. Simulated reads that mapped against the masked reference and the simulated genotypes generated with binomial test p-value less than 0.05 were excluded from the ASE candidate pool as false positives. Since the analysis was confined to heterozygous loci, we were unable to detect extreme cases of ASE in which only one allele was expressed. This bias against extreme ASE leads to an underestimate of the total ASE levels but at the same time makes our analysis more conservative.

Using a general framework of *cis-trans* divergence with respect to allelic ratios (Wittkopp, Haerum et al. 2004, Landry, Wittkopp et al. 2005), we formulated *cis*-score (S_{cis}) and *trans*-score (S_{trans}) as follows:

$$S_{cis} = 1 - \frac{1}{4} \cdot (|p - f| + |p - b|)$$

$$S_{trans} = 1 - \left(\frac{1}{4} \cdot |f - b| + \frac{1}{4} \cdot (|f| + |b|)\right)$$

where p, f, b are allelic ratios for parental, F_1, and BC samples, respectively. For each gene, $p = (E_{pla} - E_{sil})/(E_{pla} + E_{sil})$ was calculated as the ratio based on the gene expressions of *D. planitibia* (E_{pla}) and *D. silvestris* (E_{sil}); $f = (C_{pla} - C_{sil})/(C_{pla} + C_{sil})$ was calculated as a ratio based on the number of reads corresponding to the *D. planitibia* allele (C_{pla}) and the number of reads corresponding to the *D. silvestris* (C_{sil}) allele in F_1 samples; b was calculated in a similar way as f but in BC samples. The gene was called either *cis*- or *trans*-divergent in each BC group when one score was larger than the other. If the difference of the two scores were less than 0.05, the gene was classified as *cis-trans* synergy. This formalization is consistent with the idea that *cis* divergence pattern is characterized by equal parental (*p*) and hybrid (*f*) ratios, whereas *trans* divergence pattern is characterized by the *f* ratio (but not necessarily *p* ratio) approaching zero. We expanded the framework to include heterozygous backcross hybrids, assuming that sequence divergence exerts the same ASE effects in F_1 and backcross hybrids, and hence *f* and *b* should be convergent. In addition to each phenotypic BC group classified separately, all BC individuals (combined BC) were analyzed as one group for a combined BC *cis-trans* divergence measure. A subset of 800 genes containing informative single nucleotide polymorphism between *D. planitibia* and *D. silvestris* were used for this analysis.

GO enrichment

All functional enrichment analyses were carried out on the Database for Annotation, Visualization and Integrated Discovery (DAVID) site (v. 6.7) (http://david.abcc.ncifcrf.gov/). The list of DE genes was uploaded to the DAVID website which cross-referenced the gene IDs amongst many databases to provide gene ontology annotations as well as categorizing the genes into broader functional groups.

Interaction network

Interaction network for *cis*- and *trans*-divergent genes was generated using GeneMania (Mostafavi, Ray et al. 2008), while DroID (Yu, Pacifico et al. 2008) data were incorporated as an additional data source.

Results

Testes and sperm morphologies

Parental species males (N=42 *D. silvestris,* N=44 *D. planitibia*) consistently exhibited fully-formed testes with dense, coiled sperm that was motile after the testis squash (Fig 5.2A-B). F_1 hybrid males (N=23) were sterile, containing no sperm inside an intact testis (Fig 5.2C). Full-sib BC males showed a range of sperm morphologies (Fig 5.2D-F) consistent with the results reported in Craddock (Craddock 1974). BC-MS males (N=19) exhibited a large number (>80%) of highly motile sperm inside an intact testis (Fig 5.2D); BC-NM males (N=15) exhibited non-motile sperm in intact testes (Fig 5.2E); and BC-NS males (N=15) exhibited a completely empty testis where no sperm were present similar to F_1 hybrid males (Fig 5.2F). Of the 16 BC individuals dissected in this study the distribution of the three sperm phenotype categories were as follows: 7 were BC-MS (motile sperm), 4 were BC-NM (non-motile sperm present), and 5 were BC- NS (completely sterile, absence of sperm).

Transcriptome assembly and annotation

After quality control for raw sequencing reads, an average of 2.84 Gb clean data per sample was generated. Clean reads from all samples were used in the *de novo* transcriptome assembly. Trinity output yielded 102,981 genes (170,887 transcripts) with an average contig/transcript length of 554 base pairs (bp). To assure enrichment of mRNAs, contigs were post-processed by filtering out ORFs less than 300 bp. A remaining total of 24,320 genes (55,038 transcripts) with average length of 991 bp and total length of 54.5Mb were used for mapping. Out of the 24,320 genes, 21,307 (87.6%) mapped to the NR database under our criteria, with 16,761 (68.9%) of them with hits to *Drosophila* Gene IDs, were assigned according to the hits, with the majority coming from the well annotated *D. melanogaster* (13,655), and *D. grimshawi* (2,933), the only Hawaiian *Drosophila* species sequenced to date.

Analysis of differential gene expression and candidate genes

Gene expression was compared among BC groups. There were 65 genes significantly differentially expressed (DE) between BC-MS and BC-NS and 33 between BC-NM and BC-NS, but no significantly differential gene expression was found between BC-MS and

BC-NM (Fig 5.3, Appendix Fig 5.1 and Appendix Table 5.1). Interestingly, all significantly differentially expressed genes found between BC-MS and BC-NS, as well as BC-NM and BC-NS were down-regulated in BC-NS. For comparison, 41 genes showed higher expression in BC-MS and 30 genes had higher expression in BC-NM for the comparison between BC-MS and BC-NM, although differential expression was not statistically significant for any of these genes (Appendix Table 5.1 & Appendix Fig 5.1).

Among the 33 DE genes between BC-NM and BC-NS, 27 were shared in DE genes between BC-MS and BC-NS groups. The most significant shared DE gene was CG31467, known to be expressed at moderate levels in D. melanogaster adult testis according to FlyAtlas (http://flyatlas.org/). No significant GO enrichment was found for DE genes, presumably due to the relatively low level of functional annotation. Cellular components, biological processes, and molecular functions for each gene hit were recorded if the data were available in Flybase (Table 5.1). Nevertheless, we have analyzed GO term enrichments for the top 5% misexpressed genes in the three comparisons between BC groups (Appendix Table 5.2). The analysis showed high similarities between the comparisons, especially between BC-MS/BC-NS and BC-NM/BC-NS, in which GO terms related to microtubule cytoskeleton were among the most overrepresented.

Of the 65 DE genes between BC-MS and BC-NS, four genes have known spermatogenesis-related function(s) in D. melanogaster (Table 5.1). Of the 33 DE genes between BC-NM and BC-NS, at least three genes have spermatogenesis-related function: kl-5, pendulin, and CG15161 (Table 5.1). Other important functions represented by the remaining DE genes include microtubule activities (LP11180p, kl-2, Dgri\GH10382, gb:AAN71272.1), phosphatase binding (CG31467, Dgri\GH13925), oxidation-reduction processes (Dgri\GH19779, Dgri\GH22053, Dgri\GH17535), and hydrolase activities (Dgri\GH11455, PpV). Of the 71 DEs found among BC groups, all of them also showed differential expression between parental species and BC-NS, and 42 of them showed differential expression between parental species and both BC-MS and BC-NM.

Allele-specific expression analysis

There were more *cis*-divergent than *trans*-divergent genes with no significant difference in the ratio of *cis*- and *trans*-regulated genes amongst three BC male groups (X^2 = 1.29, df = 2, P = 0.52, Table 5.2). BC-MS had 335 *cis*-regulated, 97 *trans*-regulated, and 169 synergistic genes; BC-NM had 283 *cis*-regulated, 67 *trans*-regulated, and 125 synergistic genes; BC-NS had 333 *cis*-regulated, 90 *trans*-regulated, and 143 synergistic genes; and the combined BC groups had 333 *cis*-regulated, 108 *trans*-regulated, and 180 synergistic genes (Table 5.2). Out of the 71 significantly misexpressed genes (Table 5.1 and Appendix Table 5.1), only four had polymorphism (species-specific SNPs in transcribed sequences and heterozygosity in backcrosses) permitting computations of *cis-trans* scores and *cis-trans* divergence classification. Two out of the four genes were *cis-trans* synergistic (*Octbeta3R* and CG5196), and two were *cis*-regulated (Dgri\GH10450 and Dgri\GH13801). To test for a possible contribution of compensatory *cis-trans* evolution (manifested as misregulation of *cis-trans* synergistic genes in hybrids (Landry, Wittkopp et al. 2005)) to backcross hybrid sterility, we estimated Spearman's rank correlations between *cis-trans* score differences (being low for *cis-trans* synergy) and expression fold-changes between the BC groups. The correlation was positive but not significant (r = 0.04, p = 0.34), inconsistent with the scenario that compensatory *cis-trans* evolution significantly contributed to differences between the BC groups. In addition, an interactome network of *cis*- and *trans*-divergent genes was generated (Appendix Fig 5.2).

Discussion

Sperm morphologies

Our analysis of sperm production and motility in this study is consistent with Craddock's (Craddock 1974) study showing similar proportions of the three fertility phenotypic groups in the BC males. The results show that the underlying physiological processes in the BC males of these two species results in three distinct phenotypic classes with one group lacking sperm in the testes (BC-NS), a second group with non-motile sperm (BC-NM), and a third group with testes filled with motile sperm (BC-MS). The F_1 hybrid males are similar to the sterile BC males (BC-NS), and both parental species are similar to the fertile BC males (BC-MS). These results also suggest that HMS may be caused by two processes, with one disrupting sperm production and the other disrupting sperm motility.

Differential gene expression

Our RNA-Seq analyses on BC adult male testes also showed highly significant differential gene expression between two of the three phenotypic groups of BC males. There were 65 differentially expressed genes observed between BC-MS and BC-NS and 33 DE genes between BC-NM and BC-NS. Three potential candidate genes of interest are identified as *D. melanogaster* genes *lost boys,* male fertility factor *kl-2,* and male fertility factor *kl-5. lost boys* is a gene involved in ciliar motility. It encodes a conserved flagellar protein CG34110, which is localized along fly sperm flagella and is highly expressed in ciliated respiratory epithelia and sperm (Yang, Cochran et al. 2011). Phenotypic analysis in *D. melanogaster* showed that *lost boys* specifically affected sperm movement into the female storage receptacle (Yang, Cochran et al. 2011). Therefore, it is a gene that is involved directly in sperm motility and sperm storage.

Male fertility factor kl-2 and *male fertility factor kl-5* are two genes that reside in the long arm of the Y chromosome in *D. melanogaster*. They are two of seven fertility factors identified in *D. melanogaster* (Carvalho, Lazzaro et al. 2000). Deletion studies showed that the lack of *kl-5* results in the loss of the outer arm of the sperm tail axoneme; sperm in males lacking *kl-2* and *kl-5* were missing important heavy chain proteins, and therefore the individuals produced immotile sperm (Carvalho, Lazzaro et al. 2000). *kl-5* is known to code for an axonemal beta-dynein heavy chain expressed in the testis; these heavy chains are known to be responsible for the motility of flagella and cilia (Carvalho, Lazzaro et al. 2000). It is interesting to note that all of the BC males in this study have the same *D. silvestris* Y-chromosome. If *kl-2* and *kl-5* genes are on the Y-chromosome of *D. silvestris* and *D. planitibia,* it suggests that in order for there to be DE of these genes between the backcross males in this experiment, that all have the same Y-chromosome, there could have been interactions of the *kl-2* and *kl-5* genes with genes or regulatory factors located elsewhere in the genome on either an autosome or X-chromosome. The construction of chromosomal maps for these species would be valuable in determining if these genes are located on the Y-chromosome.

Other functionally annotated genes of interest in this set include *CG31467* and *Octbeta3R*. *CG31467,* the top DE gene, plays a role in phosphatase binding and is expressed moderately in the adult testis of *D. melanogaster* (FlyBase). Protein phosphatases have been known to modulate sperm motility in mammals (Fardilha, Esteves et al. 2011), and a gene encoding acylphosphatase (*Acyp*) has been found to be associated with HMS in F_1 and backcross hybrids between *D. simulans* and *D. mauritiana* (Michalak and Noor 2004, Michalak and Ma 2008). *Octbeta3R*, homologous to beta-adrenergic receptors in vertebrates, is an octopamine, which play multifunctional roles in insects (Farooqui 2012). In *D. melanogaster*, octopamines have been detected in pathways relating to different behaviors such as olfactory learning and memory, aggression, locomotion and grooming, and conditional courtship (Farooqui 2012). *Octbeta3R* specifically has been known to partially restore ovulation and fecundity in sterile females.

DE genes between BC-MS and BC-NS and between BC-NM and BC-NS were all down-regulated in the sterile phenotype that lacked sperm (BC-NS) compared to the fertile phenotype (BC-MS) and the non-motile sperm phenotype (BC-NM). This result suggests, as expected, that the sterile phenotype does not express testis function genes compared to the purported fertile phenotype. It should be noted that many of the down-regulated genes in BC-NS do not necessarily need to be responsible for, or even associated with, HMS. Indeed, many of these genes are also down-regulated in BC-MS and BC-NM relative to the parental species, suggesting that this misexpression may be a more general feature of backcross hybrids regardless of their fertility phenotype. On the other hand, it would also be premature to claim that for this reason such genes cannot be related to fertility, as we cannot rule out a possibility that even fertile backcross males are subfertile relative to parental species. Another complication in this and other studies of gene expression in HMS is due to tissue structural alterations, including gonadal atrophies and sperm deficiencies, leading to spurious expression changes among genes with tissue-specific activity, difficult (if possible at all) to distinguish from true gene silencing. Since BC-NS males lack sperm, genes with sperm-specific expression will necessarily be underrepresented in the analysis. To minimize the effect of sperm absence on the expression profile, we excluded genes with zero-level expression.

Interestingly, we did not find significant differential expression between BC-MS and BC-NM. Both BC-MS and BC-NM contain sperm, but BC-MS contains many motile sperm, and BC-NM contains less dense, non-motile sperm. Although such sperm motility-related genes as *lost boys, male fertility factor kl-2,* and *male fertility factor kl-5*, were significantly underexpressed in BC-NS, their expression alterations were subtler in BC-NM. Indeed, there were a number of genes that exhibited differential expression between the BC-MS and BC-NM phenotypic groups that did not reach statistical significance with the sample sizes in this study (Fig 5.3). Both BC-MS/BC-NS and BC-NM/BC-NS top DE gene was CG31467, which is moderately expressed in the testis. Alternatively, BC-NM phenotype could be conveyed through regulatory changes at postranscriptional, translational or post-translational levels, including protein-protein interactions. Lastly, we did not capture inter-family variation, since all BC males used in this study were offspring from a single F_1 hybrid female/*D. silvestris* parental male pair, which has its advantages (genetically related brothers with distinct phenotypes) but also disadvantages (difficulty with extrapolations to the population levels).

Allele specific expression

The interaction of *cis*- and *trans*- regulatory factors during transcription can affect gene expression (reviewed in (Bell, Kane et al. 2013)), and both factors are subject to mutational changes that may result in transcriptional alterations. Sensitivity of gene expression to mutations increases with both increasing trans-mutational target size and the presence of a TATA box (Landry, Lemos et al. 2007). These regulatory networks are primarily composed of regulatory and structural genes (Wittkopp, Haerum et al. 2004). We found overwhelmingly more *cis*-divergent genes than *trans*-divergent genes in all BC groups. This result is consistent with those from other *Drosophila*, including closely related *Drosophila p. pseudoobscura* and *D. p. bogotana* (Gomes and Civetta 2015), as well as more distantly related *D. melanogaster* and *D. simulans* (Wittkopp, Haerum et al. 2004, Wittkopp, Haerum et al. 2008). A review of regulatory experiments by Bell et al. (Bell, Kane et al. 2013) concludes that generally *cis*-regulatory changes account for more divergent expression between more genetically divergent parents (e.g. interspecific) than *trans*-effects, which account for a higher proportion of variation in gene expression between less divergent parents (e.g. intraspecific).

Wittkopp et al.'s (2004) models of regulatory divergence suggest structural genes tend to be more *cis*-regulatory than *trans*-regulatory, due to their proximity to terminal nodes of the network where expression of genes is not regulated within the network. In contrast, some of the *cis*-genes of interest had regulatory functions. However, we did not observe an excess of DNA-binding gene activities (or other significant GO term enrichments) among *trans*-divergent genes (Appendix Tables 5.3-5.6). The most abundant groups of genes among *cis*-divergent genes were functionally related to cytoskeleton (12 genes) and reproductive cellular processes (7 genes). Unlike Landry *et al.* (2005), we did not observe increased misexpression of *cis* x *trans* synergistic genes. This is likely related to the fact that comparisons between backcross hybrids display less extensive misexpressions than comparisons between F_1 hybrids and parental species (Michalak and Noor, 2004), while *cis* x *trans* synergy does not disproportionally contribute to backcross sterility.

Evolutionary implications of regulatory divergence

The combination of *cis*- and *trans*-regulatory elements in these BC groups allows further insight into the gene expression within and between groups, and the potential genetic architecture contributing to their divergence. Gene regulation has been shown to be heritable (Pavey, Collin et al. 2010, Yang, Liu et al. 2014) and an important part of divergence among species and variation within populations, but *cis*- and *trans*-acting elements differ in their evolutionary influence (Meiklejohn, Coolon et al. 2014). For example, if *cis*-elements are stronger than *trans*-elements, then changes in a species pair may have evolved one gene at a time, instead of through a broad sweep as would be inferred by a *trans*-dominant regulation, which affects many genes (Wittkopp, Haerum et al. 2004). Understanding regulatory networks can lead to further insight into the potential associations amongst gene expression, adaptive genetic divergence, and reproductive isolation, as this link is not conclusive in many experiments and needs further testing (Pavey, Collin et al. 2010). The combination of molecular techniques and fitness assays will provide a more robust analysis of the relationship between isolation and divergence relative to gene expression (Pavey, Collin et al. 2010). These studies could also be used as a proxy for comparing and contrasting the ecological divergence of other pairs of closely related Hawaiian *Drosophila* species, such as the sympatric, strongly reproductively-isolated *D. heteroneura* and *D. silvestris*. Potential differences in *cis*- and *trans*-acting factors may be important in providing a more robust analysis and unraveling the evolutionary histories amongst species.

Overall, the HMS between the closely related *D. silvestris* and *D. planitibia* species highlight the relatively rapid divergence of gene regulation for testis development and sperm production in Hawaiian *Drosophila*. The physiological and developmental processes that are involved in testis formation and sperm production appears to have diverged to such an extent that the two systems have become incompatible within approximately 0.7 million years. The evolution of these species on separate islands also suggests that selection operating separately on male gamete production within each species is an important contributor to divergence between these two species. Furthermore, the potential for interactions between genes on the Y-chromosome with genes or genetic factors located either on the X-chromosome or autosomes is consistent with the proposition that this sterility occurs, at least in part, through epistatic interactions of alleles from the two species that lead to incompatible systems (Johnson 2000). As gonad formation begins in the larval stage and is maintained into the adult stage (Williamson and Lehmann 1996) it will be important to examine gene expression patterns throughout the developmental process to determine where the compatibility of the genomes breaks down in the formation of testes and the production of sperm. Further research is also necessary to examine the potential for a smaller number of genes to create the initial incompatible developmental pathways that then leads to other changes in developmental and gene expression patterns to create hybrid male sterility (Johnson 2000, Coyne and Orr 2004). The comparison of gamete formation between more species in a phylogenetic context could shed light into how gene expression patterns and the gamete physiological and developmental systems evolve over time within and between species (Johnson and Porter, 2007; Porter and Johnson, 2002).

Acknowledgements

We thank Renee Corpuz for assisting in raising the Hawaiian picture wing *Drosophila* and Anne Veillet at the University of Hawaii at Hilo Core Genomics Facility for assisting with the dissections and RNA extractions. We thank Betsy Gagné, Cynthia King and Lyman Perry of the Hawaii Department of Land and Natural Resources, Division of Forestry and Wildlife for permission to collect Hawaiian *Drosophila* and Mark Vaught of the East Maui Irrigation Co. for permission to collect in the Waikamoi Forests in East Maui. Three anonymous reviewers provided valuable suggestions for improving the manuscript. This research is based upon work supported by the National Science Foundation CREST Center award EHR-0833211 to DKP.

Tables

Table 5.1. Differentially expressed genes among the testes of BC-MS, BC-NM, and BC-NS and their known biological processes and molecular functions. Only genes with known functions are included in this table.

Gene ID^	F/S*	NM/S*	biological processes	molecular function
^{12}*CG31467*	S	S	negative regulation of phosphatase activity	phosphatase binding
^{12}*kl-5*	S	NS	microtubule-based movement	ATPase activity, coupled; motor activity
^{2}*pendulin*	S	NS	centrosome duplication; lymph gland development; sperm individualization	protein transmembrane transporter activity
^{12}*Octbeta3R*	S	NS	G-protein coupled receptor signaling pathway	G-protein coupled amine receptor activity
^{12}*LP11180p*	S	NS	microtubule-based movement	motor activity; ATPase activity, coupled
^{2}*Dgri\GH10450*	S	NS	metabolic processes; transmembrane transport	catalytic activity
^{12}*pen*	S	S	apposition of dorsal and ventral imaginal disc-derived wing surfaces	RNA binding
^{2}*Dgri\GH13925*	S	S	negative regulation of phosphatase activity	phosphatase binding
^{12}*Dgri\GH19779*	S	S	oxidation-reduction process	oxidoreductase activity
^{12}*Dgri\GH20554*	S	S	rRNA transcription	endoribunuclease activity
^{12}*Dgri\GH11455*	S	S	metabolic process	hydrolase activity
^{2}*CG5196*	S	S	golgi organization; protein palmitoylation	protein-cysteine S-palmitoyltransferase activity; zinc ion binding
^{12}*CG5718*	S	S	electron transport chain; tricarboxylic acid cycle	succinate dehydrogenase (ubiquinone) activity
^{12}*lobo (lost boy)*	S	NS	sperm motility; sperm storage	N/A
^{12}*Dgri\GH10382*	S	NS	microtubule nucleation	N/A
^{2}*PpV*	S	S	mitotic cell cycle; protein dephosphorylation	hydrolase activity
^{2}*Prosbeta7*	S	S	cell proliferation; cellular response to DNA damage stimulus; mitotic spindle	endopeptidase activity
^{2}*Dgri\GH11042*	S	NS	protein phosphorylation	ATP binding; protein kinase activity
^{2}*CG15161*	S	S	cilium assembly; intraciliary transport	N/A
^{12}*kl-2*	S	NS	microtubule-based movement	ATP binding; ATPase activity, coupled; motor activity
^{2}*ZnT33D*	S	S	cellular zinc ion homeostasis	zinc ion transmembrane transporter activity
^{2}*Dgri\GH23359*	S	S	protein glycosylation	fucosyltransferase activity
^{12}*CG9173*	S	NS	regulation of cell cycle	N/A
^{12}*Dgri\GH22053*	S	NS	N/A	cytochrome-c oxidase activity
^{12}*Dgri\GH10111*	S	NS	lipid metabolic process	phosphoric diester hydrolase activity; starch binding
^{2}*Dgri\GH15386*	S	NS	N/A	lysozyme activity
^{2}*Tpc2*	S	NS	transmembrane transport	transmembrane transporter activity

^{12}*Dgri\GH17535*	S	NS	oxidoreductase activity	oxidation-reduction process
^{2}*Vha16-1*	S	NS	dsRNA transport; imagine disc-derived wing morphogenesis	hydrogen ion transmembrane transporter activity
^{2}*Dgri\GH25085*	S	NS	metabolic process	catalytic activity
^{2}*brv3*	NS	S	calcium ion transport	calcium channel activity; calcium ion binding
^{2}*Dgri\GH16238*	NS	S	protein glycosylation	fucosyltransferase activity
^{2}*gb:AAN71272.1*	S	S	microtubule-based movement	motor activity; ATPase activity, coupled
^{12}*unc80*	NS	S	locomotor rhythm	cation channel activity

*: F/S and NM/S indicate comparison between BC-MS and BC-NS, and between BC-NM and BC-NS, respectively. S indicates significant differential expression; NS indicates non-significant differential expression.

^: Superscript 1 indicates differential expression between parental species and both BC-MS and BC-NM; superscript 2 indicates differential expression between parental species and BC-NS.

Table 5.2. Total number of *cis*-divergent, *trans*-divergent, and *cis/trans* synergistic genes expressed in the testes of each BC group and all BC groups combined.

BC group	***cis***	***trans***	***cis/trans***
BC-MS	335	97	169
BC-NM	283	67	125
BC-NS	333	90	143
All BC	333	108	180

Figure legends

Fig 5.1. Schematic diagram of backcrossing of F_1 *D. silvestris x D. planitibia* female to *D. silvestris* with three distinct sperm phenotypes among male offspring.

Fig 5.2. Photographs of representative testis phenotypes post-testis squash exhibited by A) *D. silvestris,* B) *D. planitibia*, C) F_1 hybrid, D) BC-MS (motile sperm), E) BC-NM (sperm present but non-motile), and F) BC-NS (no sperm present). All dissections and photographs were taken under a light compound microscope at either 10x or 40x magnification.

Fig 5.3. Volcano plot for differential expression analysis among the testes of backcross male groups. The x-axis is the log2 fold change of the differences in gene expression **A:** BC-MS > BC-NM, **B:** BC-MS > BC-NS, **C**: BC-NM > BC-NS. The y-axis is the negative of the log10 FDR-values of the comparison between the two BC male groups with red dots indicating significant DE genes.

References

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol, 11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Araripe, L. O., Montenegro, H., Lemos, B., & Hartl, D. L. (2010). Fine-scale genetic mapping of a hybrid sterility factor between Drosophila simulans and D. mauritiana: the varied and elusive functions of "speciation genes". *BMC Evol Biol, 10*, 385.

Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal, 7*, 1-8. doi:10.2174/1875036201307010001

Bell, G. D., Kane, N. C., Rieseberg, L. H., & Adams, K. L. (2013). RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biol Evol, 5*(7), 1309-1323. doi:10.1093/gbe/evt072

Carson, H. L. (1987). High Fitness of Heterokaryotypic Individuals Segregating Naturally within a Long-Standing Laboratory Population of Drosophila silvestris. *Genetics, 116*(3), 415-422.

Carvalho, A. B., Lazzaro, B. P., & Clark, A. G. (2000). Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. *Proc Natl Acad Sci U S A, 97*(24), 13239-13244. doi:10.1073/pnas.230438397

Coyne, J. A., & Orr, H. A. (2004). *Speciation* (Vol. 37): Sinauer Associates Sunderland, MA.

Craddock, E. M. (1974). Reproductive relationships between homosequential species of Hawaiian Drosophila. *Evolution*, 593-606.

Dickman, C. T., & Moehring, A. J. (2013). A novel approach identifying hybrid sterility QTL on the autosomes of Drosophila simulans and D. mauritiana. *PLoS One, 8*(9), e73325.

Fardilha, M., Esteves, S. L., Korrodi-Gregorio, L., Pelech, S., da Cruz, E. S. O. A., & da Cruz, E. S. E. (2011). Protein phosphatase 1 complexes modulate sperm motility and present novel targets for male infertility. *Mol Hum Reprod, 17*(8), 466-477.

Farooqui, T. (2012). Review of octopamine in insect nervous systems. *Open access insect physiol, 4*, 1-17.

Gomes, S., & Civetta, A. (2015). Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci Rep, 5*, 11976.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol, 29*(7), 644-652. doi:10.1038/nbt.1883

Haldane, J. B. (1922). Sex ratio and unisexual sterility in hybrid animals. *J Genet, 12*(2), 101-109.

Johnson, N. (2000). Gene interactions and the origin of species. *and Wade MJ (eds) Epistasis and the Evolutionary Process. Oxford University Press, New York*, 197-212.

Johnson, N. A., & Porter, A. H. (2007). Evolution of branched regulatory genetic pathways: directional selection on pleiotropic loci accelerates developmental system drift. *Genetica, 129*(1), 57-70. doi:10.1007/s10709-006-0033-2

Landry, C. R., Hartl, D. L., & Ranz, J. M. (2007). Genome clashes in hybrids: insights from gene expression. *Heredity (Edinb), 99*(5), 483-493. doi:10.1038/sj.hdy.6801045

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J., & Hartl, D. L. (2007). Genetic properties influencing the evolvability of gene expression. *Science, 317*(5834), 118-121.

Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G., & Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics, 171*(4), 1813-1822. doi:10.1534/genetics.105.047449

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Lee, H. Y., Chou, J. Y., Cheong, L., Chang, N. H., Yang, S. Y., & Leu, J. Y. (2008). Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell, 135*(6), 1065-1073.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Magnacca, K. N., Foote, D., & O'Grady, P. M. (2008). A review of the endemic Hawaiian Drosophilidae and their host plants. *Zootaxa, 1728*, 1-58.

Magnacca, K. N., & Price, D. K. (2015). Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing Drosophila (Diptera: Drosophilidae). *Mol Phylogenet Evol, 92*, 226-242. doi:10.1016/j.ympev.2015.06.014

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Meiklejohn, C. D., Coolon, J. D., Hartl, D. L., & Wittkopp, P. J. (2014). The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Res, 24*(1), 84-95. doi:10.1101/gr.156414.113

Michalak, P., & Ma, D. (2008). The acylphosphatase (Acyp) alleles associate with male hybrid sterility in Drosophila. *Gene, 416*(1-2), 61-65.

Michalak, P., & Noor, M. A. (2003). Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol Biol Evol, 20*(7), 1070-1076. doi:10.1093/molbev/msg119

Michalak, P., & Noor, M. A. (2004). Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. *Journal of Molecular Evolution, 59*(2), 277-282. doi:10.1007/s00239-004-2622-y

Moehring, A. J., Teeter, K. C., & Noor, M. A. (2007). Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Molecular Biology and Evolution, 24*(1), 137-145.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol, 9 Suppl 1*, S4. doi:10.1186/gb-2008-9-s1-s4

Noor, M. A., & Feder, J. L. (2006). Speciation genetics: evolving approaches. *Nat Rev Genet, 7*(11), 851-861.

O'Grady, P. M., Lapoint, R. T., Bonacum, J., Lasola, J., Owen, E., Wu, Y., & DeSalle, R. (2011). Phylogenetic and ecological relationships of the Hawaiian Drosophila inferred by mitochondrial DNA analysis. *Mol Phylogenet Evol, 58*(2), 244-256.

Oliver, P. L., Goodstadt, L., Bayes, J. J., Birtle, Z., Roach, K. C., Phadnis, N., . . . Ponting, C. P. (2009). Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet, 5*(12), e1000753. doi:10.1371/journal.pgen.1000753

Orr, H. A., Masly, J. P., & Presgraves, D. C. (2004). Speciation genes. *Curr Opin Genet Dev, 14*(6), 675-679. doi:10.1016/j.gde.2004.08.009

Pavey, S. A., Collin, H., Nosil, P., & Rogers, S. M. (2010). The role of gene expression in ecological speciation. *Ann N Y Acad Sci, 1206*, 110-129. doi:10.1111/j.1749-6632.2010.05765.x

Porter, A. H., & Johnson, N. A. (2002). Speciation despite gene flow when developmental pathways evolve. *Evolution, 56*(11), 2103-2111.

Spieth, H. T. (1986). Behavioral characteristics of Hawaiian *Drosophila*. *Proc Hawaiian Entomol Soc, 26*, 101-108.

Tao, Y., Zeng, Z. B., Li, J., Hartl, D. L., & Laurie, C. C. (2003). Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana. II. Mapping hybrid male sterility loci on the third chromosome. *Genetics, 164*(4), 1399-1418.

Ting, C. T., Tsaur, S. C., Wu, M. L., & Wu, C. I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science, 282*(5393), 1501-1504.

Uy, K. L., LeDuc, R., Ganote, C., & Price, D. K. (2015). Physiological effects of heat stress on Hawaiian picture-wing Drosophila: genome-wide expression patterns and stress-related traits. *Conservation Physiology, 3*(1), 1-14.

Williamson, A., & Lehmann, R. (1996). Germ cell development in Drosophila. *Annu Rev Cell Dev Biol, 12*, 365-391. doi:10.1146/annurev.cellbio.12.1.365

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature, 430*(6995), 85-88. doi:10.1038/nature02698

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet, 40*(3), 346-350.

Wu, C. I., Johnson, N. A., & Palopoli, M. F. (1996). Haldane's rule and its legacy: Why are there so many sterile males? *Trends Ecol Evol, 11*(7), 281-284.

Yang, S., Liu, Y., Jiang, N., Chen, J., Leach, L., Luo, Z., & Wang, M. (2014). Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. *BMC Genomics, 15*, 13. doi:10.1186/1471-2164-15-13

Yang, Y., Cochran, D. A., Gargano, M. D., King, I., Samhat, N. K., Burger, B. P., . . . Lu, X. (2011). Regulation of flagellar motility by the conserved flagellar protein

CG34110/Ccdc135/FAP50. *Mol Biol Cell, 22*(7), 976-987. doi:10.1091/mbc.E10-04-0331

Yu, J., Pacifico, S., Liu, G., & Finley, R. L., Jr. (2008). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics, 9*, 461. doi:10.1186/1471-2164-9-461

Fig 5.1

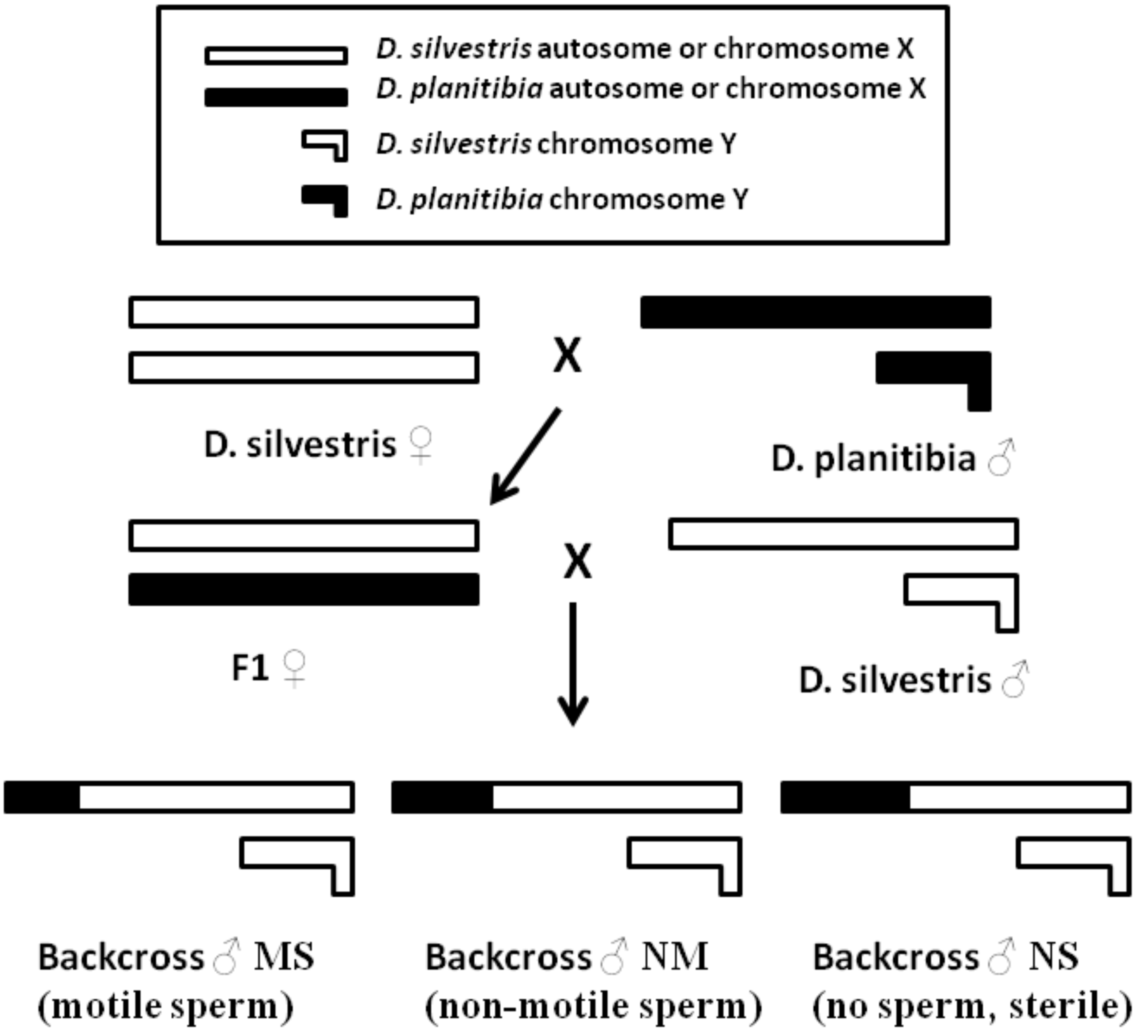

D. silvestris autosome or chromosome X
D. planitibia autosome or chromosome X
D. silvestris chromosome Y
D. planitibia chromosome Y
X
D. silvestris ♀
D. planitibia ♂
X
F1 ♀
D. silvestris ♂
Backcross ♂ MS (motile sperm)
Backcross ♂ NM (non-motile sperm)
Backcross ♂ NS (no sperm, sterile)

Fig 5.2

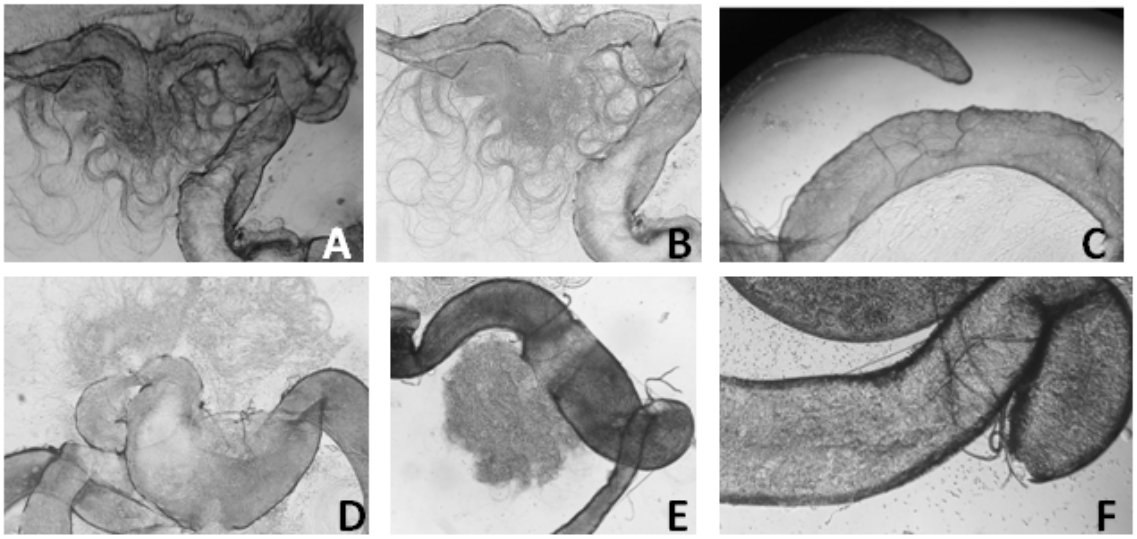
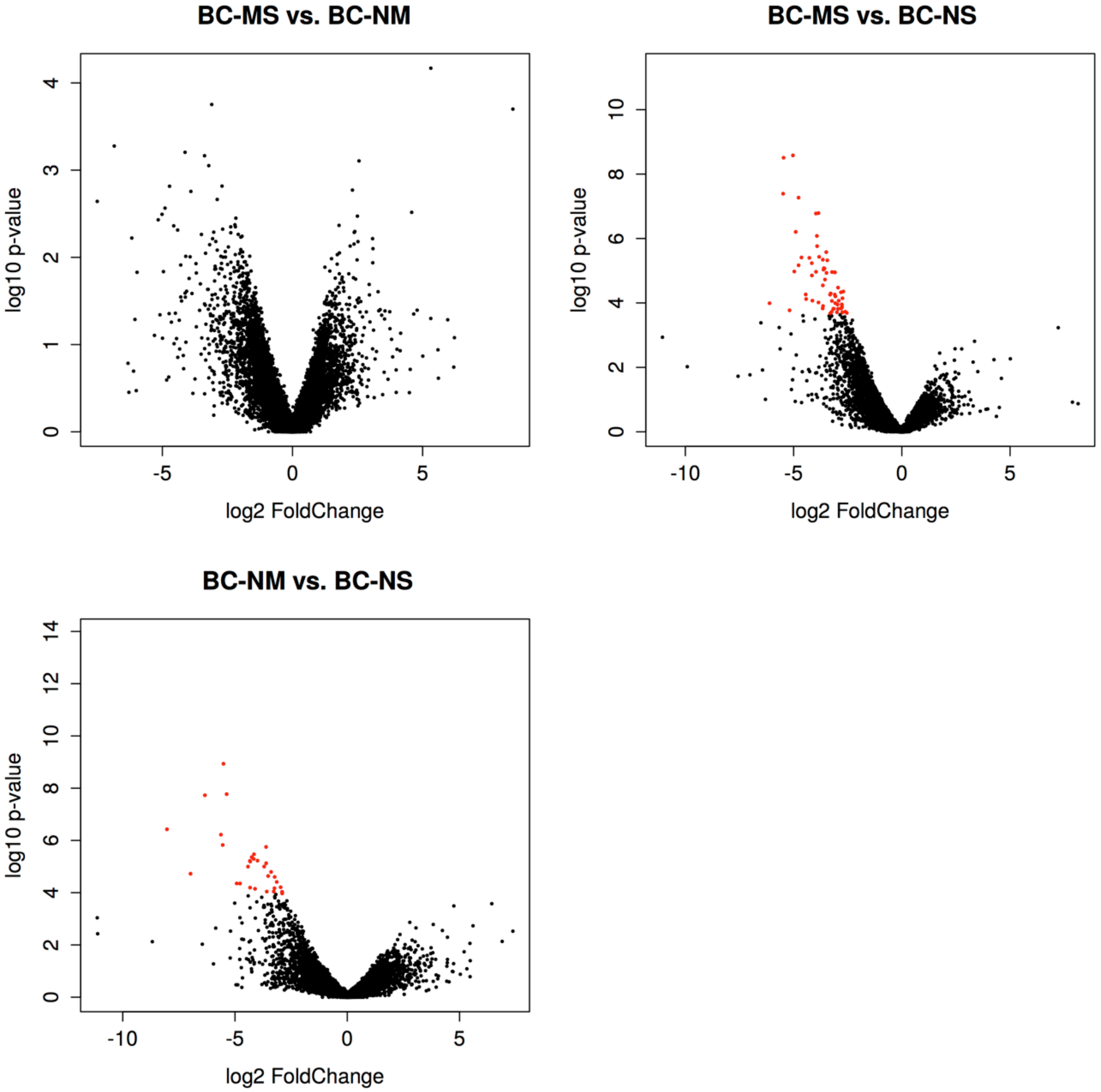


Fig 5.3


BC-MS vs. BC-NM
log10 p-value
0
1
2
3
4
-5
0
5
log2 FoldChange
BC-MS vs. BC-NS
log10 p-value
0
2
4
6
8
10
-10
-5
0
5
log2 FoldChange
BC-NM vs. BC-NS
log10 p-value
0
2
4
6
8
10
12
14
-10
-5
0
5
log2 FoldChange

Appendix

Appendix Table 5.1: 71 DEs and p-values with raw FPKM in BC groups

See (Brill et.al 2016) Supplementary information (Table S1)

Appendix Table 5.2: GO enrichment of top 5% most differentially expressed genes among BC comparisons

See (Brill et.al 2016) Supplementary information (Table S2)

Appendix Table 5.3: GO enrichment for cis-regulatory genes in BC groups

See (Brill et.al 2016) Supplementary information (Table S3)

Appendix Table 5.4: GO enrichment for trans-regulatory genes in BC groups

See (Brill et.al 2016) Supplementary information (Table S4)

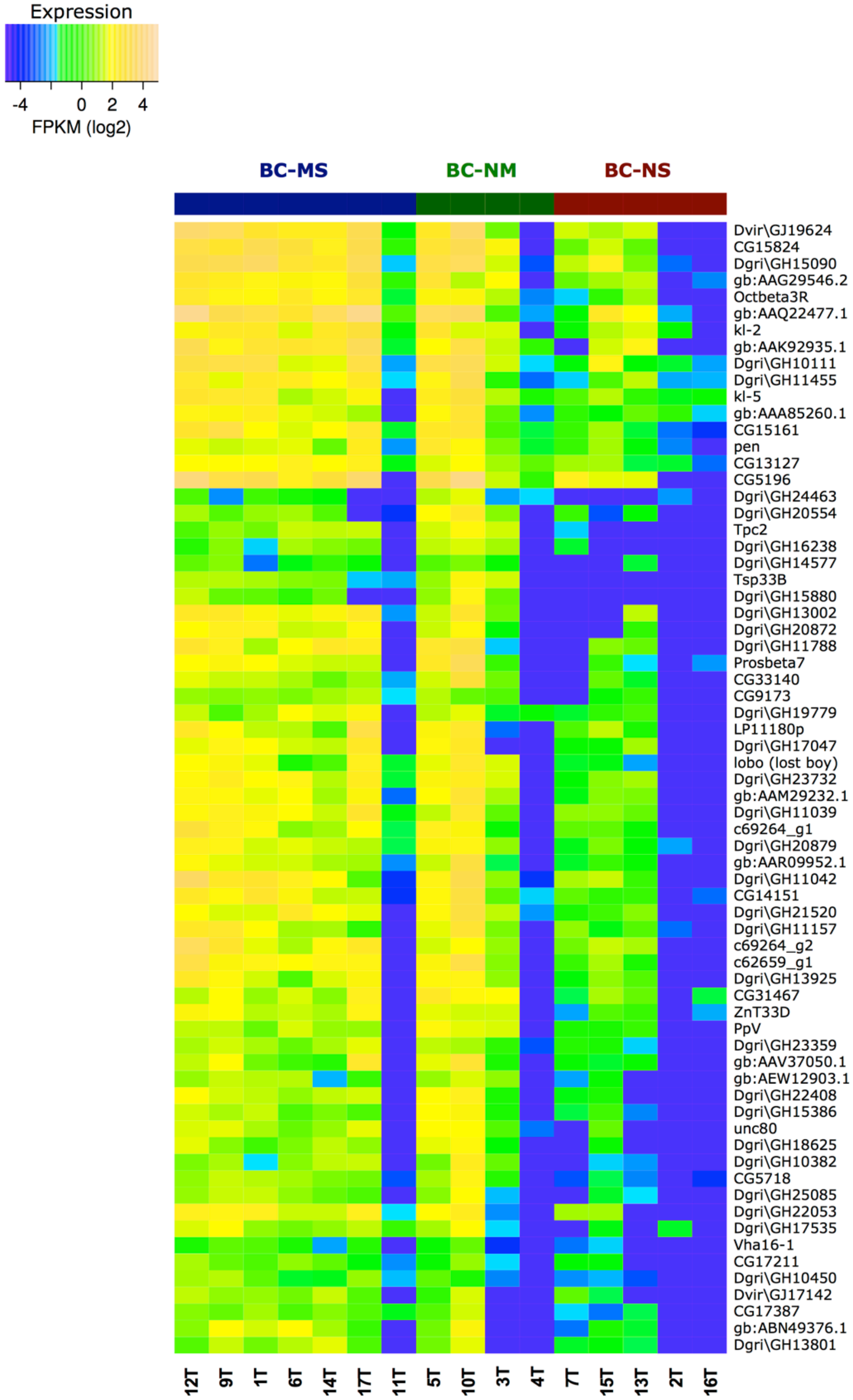
Appendix Table 5.5: GO enrichment for synergetic-regulatory genes in BC groups

See (Brill et.al 2016) Supplementary information (Table S5)

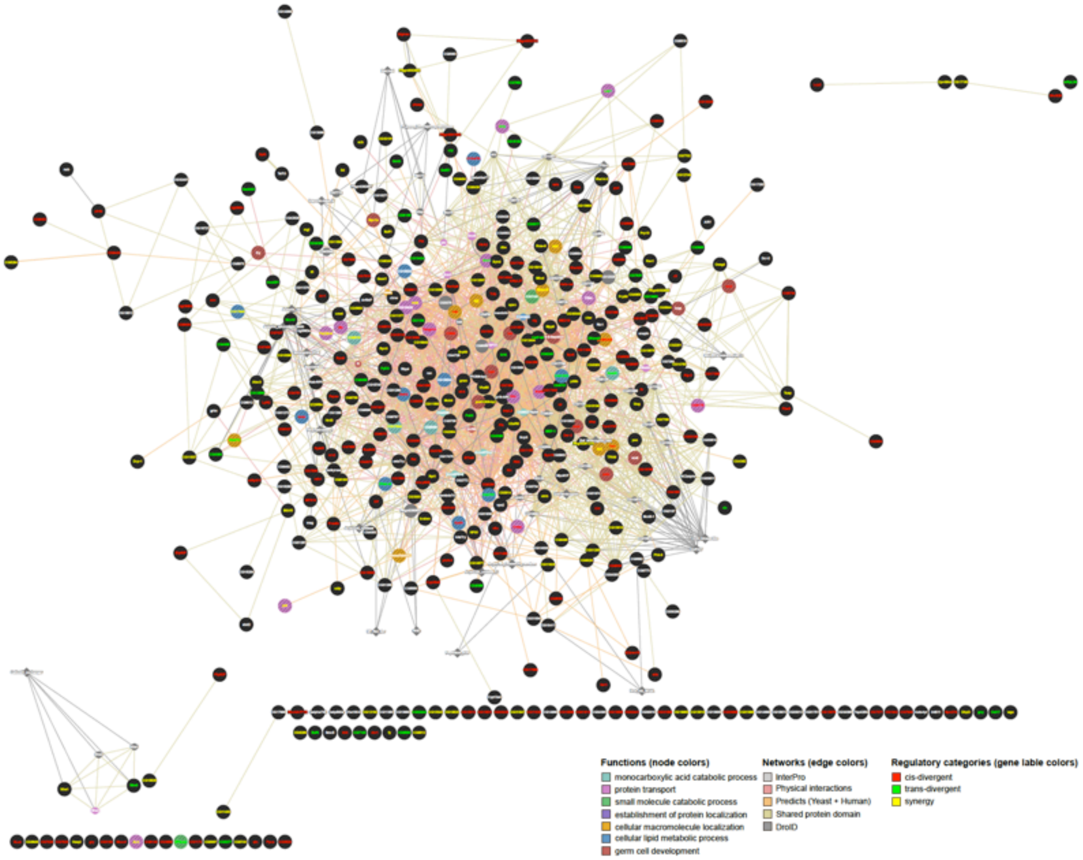
Appendix Table 5.6: detailed categories of cis- and trans- regulatory genes (NA indicates unclassified)

See (Brill et.al 2016) Supplementary information (Table S6)

Appendix Fig 5.1 Heatmap of all DE genes in BC-MS, BC-NM, and BC-NS. Yellow color indicates high expression, and blue color indicates low expression. The columns along x-axis from left to right are samples from BC-MS (n=7), BC-NM (n=4), and NC-NS (n=5).



Appendix Fig 5.2 An interactome network of *cis*- and *trans*-divergent genes (a high resolution version can be found here.)


Functions (node colors)
monocarboxylic acid catabolic process
protein transport
small molecule catabolic process
establishment of protein localization
cellular macromolecule localization
cellular lipid metabolic process
germ cell development
Networks (edge colors)
InterPro
Physical interactions
Predicts (Yeast + Human)
Shared protein domain
DroID
Regulatory categories (gene lable colors)
cis-divergent
trans-divergent
synergy