Genomic analysis of the four ecologically distinct cactus host populations of *Drosophila mojavensis*

Carson W. Allan[1,2] and Luciano M. Matzkin[1,2,3,4*]

[1] Department of Biological Sciences, University of Alabama in Huntsville, 301 Sparkman Drive, Huntsville, AL 35899, USA

[2] Department of Entomology, University of Arizona, 1140 E. South Campus Drive, Tucson, AZ 85721, USA

[3] BIO5 Institute, University of Arizona, 1657 East Helen Street, Tucson, AZ 85721, USA

[4] Department of Ecology and Evolutionary Biology, University of Arizona, 1041 E. Lowell St., Tucson, AZ 85721, USA

* Corresponding author:

Luciano M. Matzkin

lmatzkin@email.arizona.edu

(520) 621-1955

Abstract

Background: Relationships between an organism and its environment can be fundamental in the understanding how populations change over time and species arise. Local ecological conditions can shape variation at multiple levels, among these are the evolutionary history and trajectories of coding genes. This study examines the rate of molecular evolution at protein-coding genes throughout the genome in response to host adaptation in the cactophilic *Drosophila mojavensis*. These insects are intimately associated with cactus necroses, developing as larvae and feeding as adults in these necrotic tissues. *Drosophila mojavensis* is composed of four isolated populations across the deserts of western North America and each population has adapted to utilize different cacti that are chemically, nutritionally, and structurally distinct.

Results: High coverage Illumina sequencing was performed on three previously unsequenced populations of *D. mojavensis*. Genomes were assembled using the previously sequenced genome of *D. mojavensis* from Santa Catalina Island (USA) as a template. Protein coding genes were aligned across all four populations and rates of protein evolution were determined for all loci using a several approaches.

Conclusions: Loci that exhibited elevated rates of molecular evolution tended to be shorter, have fewer exons, low expression, be transcriptionally responsive to cactus host use and have fixed expression differences across the four cactus host populations. Fast evolving genes were involved with metabolism, detoxification, chemosensory reception, reproduction and behavior. Results of this study gives insight into the process and the genomic consequences of local ecological adaptation.

Keywords: Genome evolution, adaptation, Drosophila, ecological genomics, genome sequencing, genome assembly

Background

Increasing availability of whole-genome sequencing data provides new insights into the complex relationship between an organism and its environment. By examining changes in the genetic code both at the level of individual genes and at the whole-genome level it is possible to gain a better understanding of how local ecological conditions can shape the pattern of variation within and between ecologically distinct populations [1, 2]. A comprehensive integrative approach combining genomic, phenotypic and fitness data has been identified as the gold standard in understanding the adaptation process [3, 4]. Yet, an examination of the genomic divergence of ecologically distinct populations can yield valuable insight into the adaptation process especially when the genomic data is placed in an ecological context [5]. This later approach can identify genomic regions and loci that exhibit a pattern of variation and evolution suggesting their role in local ecological adaptation. Furthermore, a consequence of the fixation of ecologically-relevant variants has been implicated in the evolution of barriers to gene flow and potentially the origins of reproductively isolated populations, i.e. species [6, 7].

While it has long been accepted that natural selection is a primary driver of change within species as a response to environmental pressures, understanding the mechanism of how this selection leads to speciation is unclear [8, 9]. More recently the idea of ecological speciation, where various mechanisms work to prevent gene flow between populations causing reproductive isolation and eventually speciation, has more directly shown how selection to local ecological conditions may affect the process of speciation [6, 7]. Reproductive isolation interrupts gene flow between populations and may potentially lead to the formation of new species [10]. When different populations of a species inhabits and/or utilizes distinct resources this opens many possibilities for local differentiation that can lead to obstacles of gene flow as these populations are likely to have differing environmental pressures [6, 7]. For example, in the leaf beetle *Neochlamisus bebbianae*, different populations have distinct host preferences and larvae perform significantly worse when growing on alternative host species [8]. Host preferences and performance in this system facilitates the genetic and genomic isolation observed between the host populations, as each prefers a different microenvironment and likely does not interact and hybridize with members of the other population [11, 12].

Comparative genomic studies in mammals have shown clear evidence of positive selection both between humans, mice, and chimpanzees as well as between human populations [13-16]. Genes involved in the immune system, gamete development, sensory perception, metabolism, cell motility, and genes involved with cancer were those found to have signatures of positive selection. While in *Drosophila*, a genome level analysis of 12 species provided insight into the evolution of an ecological, morphological, physiological and behaviorally diverse genus [17]. Findings were relatively consistent with previously studies in other taxa with genes involving defense, chemosensory perception, and metabolism shown to be under positive selection [6, 13, 16, 18]. Since the *Drosophila* 12 genome project [17], several population genomics studies in *D. melanogaster* have examined variation within a single population, between clinal populations and between ancestral (African) and cosmopolitan populations to assess the consequence of population subdivision, evolution of quantitative trait variation and the adaptation to local ecological conditions [19-24]. These genome level analysis have been extended to other *D. melanogaster* species group flies with distinct life history and ecological strategies such as the *Morinda citrifolia* specialist *D. sechellia* [25] and the invasive agricultural pest *D. suzukii* [26].

Studying the sequence level constraints as well as functional categories and networks associated with genes under positive selection is paramount to understanding the process of evolutionary change. However, it is crucial to place patterns of variation and divergence in an ecological context to have a more complete view how selection shapes variation within and between populations. In this study we explore the link between ecology and patterns of genome-wide sequence variation in *D. mojavensis,* a fruit fly endemic to the southwestern United States and northwestern Mexico that has become a model for the understanding of the genetics of adaptation [27]. This species of *Drosophila* is a cactophile in that both larval and adult stages reside and feed in necrotic cactus tissues [28]. *Drosophila mojavensis* has four distinct host populations that are geographically separated (Fig. 1). In addition to geographic separation each population lives on a distinct cactus host species. The four populations are: Santa Catalina Island living on prickly pear cactus (*Opuntia littoralis*), Mojave Desert living on barrel cactus (*Ferocactus cylindraceus*), Baja California living on agria cactus (*Stenocereus gummosus*), and Sonoran Desert living on organpipe cactus (*S. thurberi*). *Drosophila mojavensis* diverged from its sister species *D. arizonae*, a cactus generalist, approximately half a million years ago [29-32] with the divergence between *D. mojavensis* populations being more recent (230,000 to 270,000 years ago) [33]. Differing host species provide different local environments for each *D. mojavensis* populations. The necrotic cactus environment in which these flies reside is composed not only of plant tissues, but a number of bacteria and yeast species [34-37]. In addition to nutritional differences between the necrotic cactus host, several of the compounds found therein have toxic properties [38-40]. This selective pressure has resulted in the fixation of variants that facilitate the survival of *D. mojavensis* and other cactophilic *Drosophila* species to their local necrotic cactus environment [28, 41].

Population genetics on individual candidate host adaptation genes in *D. mojavensis* has shown evidence for positive selection in loci involved with xenobiotic metabolism [31]. In addition, transcriptome-wide differences have been observed in *D. mojavensis* in response to host shifts [42, 43] as well as indicating fixed expression differences between the host populations [44]. Among the loci that are differentially expressed or constitutively fixed between populations many are involved in detoxification, metabolism, chemosensory perception and behavior, supporting the role of the local necrotic cactus conditions in shaping transcriptional variation [42-44]. Taking into consideration the breadth of ecological information of *D. mojavensis* this study highlights how selection pressures caused by local ecological environments differentially shape patterns of genomic variation across the host populations and provides further insight into how selection acts on organisms and its genome level consequences.

Results

Number of cleaned reads and the number assembled to the Catalina Island reference genome are shown in Table 1. All three populations had approximately 88 percent of paired-end reads successfully assembled. Mate pair reads had lower rates of mapping ranging from 27 percent to 63 percent. Of the 14,680 loci annotated in the reference genome the vast majority were also present in our template-based assemblies of the other three populations. Of these annotations, a common set of 12,695 were initially processed that did not lack any premature stop codons. From this common set of loci we filtered out those that among the four populations exhibited either less than five total, zero nonsynonymous, or zero synonymous substitutions. This yielded a working set of 9,087 loci for which all subsequent analyzes were performed. The list of all loci examined, summary data, test statistics, and *D. melanogaster* ortholog information can be found in Additional file 1: Table S1.

Characteristics and patterns of divergence of *D. mojavensis* loci

Estimates of ω (K_a/K_s) were calculated using both KaKs Calculator [45] and codeml in PAML [46]. Given the ω values were highly correlated ($r^2 = 0.88$, $P < 0.001$; see Additional file 2: Figure S1) all subsequent analyses were performed using the values from codeml. The distribution of $\log_2$ transformed ω are shown in Figure S2. Overall a total of 190 loci exhibited ω values greater than one. When examined per chromosome (Muller Element), we observed that the dot chromosome (Muller F) had the greatest mean ω, followed by the chromosomes for which segregate chromosomal inversions (Muller B and E) and than those chromosomes that lack inversions (Muller A, C and D) (Fig. 2, Additional File 2: Table S2).

To describe the characteristics of loci whose evolutionary trajectory could have been shaped by the adaptation of *D. mojavensis* populations to their respective ecological conditions we examined loci with ω values in the top 10% of the distribution, hereafter referred to as TOP10 loci. Furthermore, using codeml we performed a series of gene-wide tests of positive selection for each individual locus. Via a maximum likelihood rate test (model 7 vs. model 8) we identified 912 loci that exhibited a pattern of adaptive protein evolution. We used a smaller set of 244 loci, following an FDR correction, for all subsequent analyses, hereafter referred to as PAML-FDR loci. The set of TOP10, PAML significant loci and those with an FDR correction (PAML-FDR) can be found in Additional file 1: Table S1. The distribution of both the PAML-FDR and TOP10 loci was uniform across the *D. mojavensis* chromosomes (Additional file 2: Figure S3 and S4), with the exception that significantly fewer PAML-FDR genes were present in Muller E (Fisher's Exact test, *P* = 0.02).

Significant differences in ω values were observed across loci of differing protein coding lengths (Fig. 3). Loci smaller than 1 Kb exhibit significantly higher rate of molecular evolution, followed by those 1-2 Kb and then by gene categories of longer lengths (Additional file 2: Table S3). A similar pattern of ω values was observed for the TOP10 loci, where a significant excess of the smaller gene group (< 1 Kb) was composed of TOP10 loci, and a significantly fewer were observed in the greater than 4 Kb bin (Additional file 2: Figure S5). Although the overall ω was greater in shorter loci, the proportion of these loci who exhibited a significant pattern of positive selection was significantly less (Additional file 2: Figure S6). Similarly to what was observed for gene length, genome-wide, loci with fewer exons tended to have greater levels of ω, with the highest observed was from loci having two exons, then those with either only one or three exons, followed by those having four to six exons and lastly those with seven or more (Additional file 2: Figure S7, Table S4). TOP10 loci were overrepresented in the one and two exon categories and underrepresented in the more than seven exon category, whereas the PAML-FDR loci where uniformly distributed across all exon number categories (Additional file 2: Figures S8 and S9).

Relationship between expression and rate of molecular evolution

To assess the relationship between expression level and rate of molecular evolution we integrated our results with previous collected RNAseq data from *D. mojavensis* [47]. When examined genome-wide, genes with male-biased expression had significantly greater ω values than female-biased (Tukey HSD, P < 0.001) and unbiased (Tukey HSD, P < 0.001) expressed genes, and female-biased genes had the lowest rate (Tukey HSD, P < 0.001) of molecular evolution of all three expression categories (Additional file 2: Figure S10, Table S5). Among the TOP10 loci, there was a significant representation of them in the male-biased group of genes and a significant underrepresentation in the female-biased genes (Fig. 4). No significant over- or underrepresentation was observed among the PAML-FDR genes with respect to the sex biased expression categories (Additional file 2: Figure S11). Expression data was also used to assess the relationship between overall expression level and rate of molecular evolution. After removing both the female- and male-biased genes, we observed that of the 5,101 remaining loci those in the lowest expression category showed the greatest ω values (Additional file 2: Figure S12, Table S6). Similarly the TOP10 loci were overrepresented among the low expression category of loci and no differences were observed among the expression categories of the PAML-FDR loci (Additional file 2: Figures S13 and S14).

We also integrated our genomic data with two prior ecological transcriptional studies. We compare rates of molecular evolution of loci that are differentially expressed in response to cactus host utilization [43] as well as those loci who exhibit fixed significant expression differences between the four host populations in the absence of cactus compounds (i.e. constitutive differences) [44]. To remove the potential confounding effect of those loci that show a pattern of positive selection, we removed those loci from the subsequent expression analysis. For both datasets, loci that are either differentially expressed in response to necrotic cactus ($P < 0.001$ post FDR correction) or those that show constitutive differences between the populations ($P <$ 0.001 post FDR correction) have a significantly greater value of ω (ANOVA, $P < 0.001$, for both comparisons) (Additional file 2: Figures S15, Table S7).

Functional gene groups analysis

Of our 9,087 genes in our filtered dataset, approximately 14% (1,238) genes did not have orthologous calls back to loci in the *D. melanogaster* reference genome (Additional file 2: Figure S16). Of the remaining set of genes with *D. melanogaster* orthologs, less than half of the genes (3,649) had at least one gene ontology (GO) term. The percentage of loci without *D. melanogaster* orthologous in the TOP10 and PAML-FDR genes was greater (40% and 23%, respectively). Overall only 336 and 144 loci had at least one GO term for the TOP10 and PAML-FDR datasets, respectively. Clustering of biological process and molecular function GO terms within the TOP10 and PAML-FDR dataset illustrated some distinct functional groups. Fig. 5 illustrates the biological process functional clusters for TOP10 genes, in which clusters associated with reproduction/development, detoxification and response to stimuli, and behavior are present. A network analysis of the same set of loci indicates similar functional networks as well as those associated with defense and chromatin regulation and remodeling (Fig. 6). Functional and network clustering for molecular function GO terms, KEGG and the PAML-FDR dataset can be found in Additional file 2: Figures S17-S20, Additional file 3: Table S11. Among molecular functions, in the TOP10 dataset, serine endopeptidase activity appeared to be overrepresented (Additional file 2: Table S8).

Discussion

In this study we sequenced, assembled and analyzed the genomes of each of the four cactus host populations of *D. mojavensis* for the purpose of assessing the genomic consequences of the adaptation to local ecological conditions. Overall we were able to analyze the sequence, pattern of divergence and structure of 9,087 genes. And although the four genomes examined diverged relatively recently [29-33], for several loci, sufficient number of substitutions occurred for us to begin to assess the changes associated with cactus host adaptation.

Unlike what is present in *D. melanogaster*, *D. mojavensis* chromosomes are all acrocentric and its karyotype is composed of six Muller elements [48]. In *D. melanogaster* element A is the X chromosome and elements B/C and D/E form large metacentric chromosomes (2L/2R and 3L/3R, respectively), while the F element or dot chromosome is reduced in sized and highly heterochromatic [49, 50]. In *D. mojavensis* we observed the highest rate of molecular evolution in the small F element, followed by elements B and E, and then the remaining autosomal elements and the X chromosome (Fig. 2).

Selection on the X chromosome has been examined in a number of studies with somewhat variable results [51]. Analysis of several melanogaster group species has shown significant elevated ω values for genes on the X chromosome [17]. From population genetics theory it is generally predicted that the X chromosome would show elevated rates of evolution due to its reduced population size and level of recommendation [51]. A subsequent genomic analysis of the X chromosome across more distant *Drosophila* species (*D. melanogaster, D. pseudoobscura, D. miranda* and *D. yakuba*) failed to find evidence of increased protein evolution on the X chromosome [52]. It is difficult to make any conclusions about the lack of a pattern of accelerated X chromosome evolution found here, it may be possible that there has not been enough divergence time between these populations for influences such as effective population size to have a measurable effect. The greatest ω values were present in the dot chromosome which in *D. mojavensis* is heterochromatic and has a highly reduced level of recombination [53], which would make it highly susceptible to sweeps and hence higher rates of molecular evolution.

Within *D. mojavensis* there are polymorphic inversions in Muller elements B and E [54], both exhibited overall higher chromosomal-wide levels of ω (Fig. 3). Lower levels of recombination and higher divergence rates have been known to occur around the inversion breakpoint regions in *Drosophila* [55]. One possible explanation for the elevated rates of molecular evolution in these chromosomes is the distinct karyotypes of the sequenced lines (Additional file 2: Table S9). One consequence of a template-based assembly as performed in this study, is that chromosomal structural differences can be largely wiped away. A more detailed analysis of the consequence of chromosomal inversion on the evolutionary trajectories of associated loci will be performed in future analyses of *de novo* assemblies of *D. mojavensis* genomes from all host populations as well as from sibling species (*D. arizonae* and *D. navojoa*) (unpublished data, Matzkin).

Genes across the genome as well as those evidence of positive selection or in the top 10 percent of ω values were assessed for a number of characteristics. Genome-wide loci exhibiting greater ω values tended to be shorter, have fewer exons (3 or less), have low expression, be differentially expressed in response to cactus host use and have fixed expression differences across the four cactus host populations of *D. mojavensis* (Fig. 3; Additional file 2: Figures S7, S12, S15). Overall this pattern of divergence was similar when examining the TOP10 or PAML-FDR loci.   Previous genomic analyses in *D. melanogaster* and related species have observed similar characteristics of loci with elevated ω values. This indicates that although the phylogenetic scale of the present study is limited (within *D. mojavensis*) the forces shaping genome evolution between diverged species can also be observed between recently isolated populations within species.

The first comparative genomic study within the *D. melanogaster* group species [56] observed an association between coding length and ω, which they partially attributed to a positive correlation between K_s and protein length. Longer genes have more of these mutations and this may explain in part why genes with high ω values are likely to be shorter. In this study we did not observe such correlation, in fact the relationship is negative (P < 0.001), but explains very little of the variation in K_s (r^2 = 0.004) (Additional file 2: Figure S21). Therefore, it is difficult to infer the effect of the association between K_s and protein length, and the lack of positive correlation might be a function the close relationship between the genomes studied here. The negative association between intron number and rate of molecular evolution has been previously suggested to be due to the presence of exonic splice site enhancers which help in the correct removal of introns from the transcription sequence. As mutations in these regions are more likely to be conserved changes here could cause an intron to not be removed or part of an exon to be removed instead [57]. The link between intron presence and ω values may also help explain why TOP10 genes tend to be shorter as long genes are more likely to have introns [58]. The correlation between gene length and rate of molecular evolution could also be explained as a result of the increased level of interactions between sites in larger exons [59]. In this study a negative correlation between ω and exon length (r^2 = 0.08, P < 0.001) was observed (Additional file 2: Figure S22). These interactions between residues of a protein, commonly refer to as Hill-Robertson interference [60], have a tendency to buffer against the accumulation of amino acid substitutions.

Highly expressed genes tend to have a higher level of constraint as indicated by the tendency of having lower rates of molecular evolution. This has been previously explained as being a result of selection against mutations that alter transcriptional and translational efficiency as well as selection for the maintenance of correct folding (translational robustness) [56, 61-65]. Given our coarse transcription data we were not able to tease apart which of the above-mentioned forces might more strongly shape the rate of molecular evolution. Nonetheless we observed a clear negative relationship across the four *D. mojavensis* genomes between transcriptional level and ω. In addition to overall expression, both tissue and sex-bias expression have been known shape the evolutionary trajectories of genes [66-68]. Male, or more specifically testes expressed genes have been associated with elevated rates of molecular evolution in Drosophila and across many taxa [69]. Many of these loci are believed to be under strong sexual selection, which would explain their accelerated rate of molecular evolution. As predicted we observed an overall higher rate of molecular evolution in male-biased genes. Even female-biased loci exhibited a significant greater ω than unbiased genes. Previous behavioral and molecular studies in *D. mojavensis* have shown that this species is experiences strong and recurrent bouts of sexual selection [70-77].

Loci indicating a pattern of positive selection and those with elevated ω appear to be associated with a wide range of metabolic processes. These changes are likely a result of the distinct nutritional and xenobiotic environment the distinct *D. mojavensis* populations experience. The chemical composition of the cacti and the species of yeast found in each rot varies [34-41] and thus the populations have likely needed to optimize the recognition, avoidance and processing of these necrosis-specific compounds through changes in metabolism, physiology and behavior.

One aspect of metabolism that has likely been shaped by cactus host adaptation is the detoxification of cactus compounds, as the distinct cactus hosts have different chemical compositions. Expression studies have shown that genes involved in detoxification are enriched when flies develop in an alternative necrotic cactus species. Fitness costs of living on the alternative cactus have also been shown to be quite high with those flies having low viability (< 40%) [43, 78, 79]. Out of all GO terms examined in this study, the only ones that were consistently overrepresented were those associated with serine-type endopeptidase activity. These type of proteins perform a number of function within organisms, among them is their targeting of organophosphorus toxins [80]. These compounds are often used in pesticides and are found to inhibit serine hydrolase function in both insects and vertebrates [80]. While the apparent positive selection on these genes may be directly due to development of resistances to pesticides they might experience in the field, but more likely they may be evolving in response to the effects of toxic or nutritional compounds found in cactus rots.

Cactophilic Drosophila have been shown to deploy a number of enzymatic strategies to ameliorate the deleterious consequences of ingesting cactus necrosis-derived compounds. Many of the previously identified proteins playing a role in detoxification in cactophiles (Glutathione S-transferases, Cytochrome P450s, Esterases and UDP-glycosyltransferase) have been associated with detoxification in a broad number of taxa [81-85]. In fact, in recent comparative genomic analysis of the cactophilic *D. buzzatii* [86] and *D. aldrichi* [87], a number of metabolic genes, including those associated with detoxification were shown to be under positive selection. In the present genomic analysis of the *D. mojavensis* genome we observed the largest functional cluster (Fig. 5) was composed of several genes belonging to known detoxification protein families, such as Cytochrome P450 and Glutathione S-transferases (Gst). Furthermore, previous transcriptional studies have indicated that these same categories of detoxification loci are differentially expressed when *D. mojavensis* are utilizing necrotic cactus tissues [42, 43]. A population genetics analysis of *GstD1* has indicated a pattern of adaptive amino acid evolution at this locus in the Sonora and Baja California populations [31]. The location of the fixed residue fixed in the lineages leading to these two populations indicated potential functional consequences and a recent kinetic analysis of these proteins have support this prediction (Matzkin, unpublished data).

The diversity of bacterial species found on each necrotic cactus provides, directly or indirectly, nutritional resources for the fly populations, but also are composed of potentially distinct pathogenic organisms [88, 89]. A number of genes with elevated rates of molecular evolution in this study are linked to a range of processes involved with the immune response. As each population is faced with a different composition of threats, the evolutionary arms race between flies and their pathogens creates further divergence between the populations as they face different pathogenic landscapes. Studies in other species, such as *D. simulans*, have found that genes with immune related functions were found to have higher rates of positive selection than the genome average [90]. Exposure to bacterial pathogens in *D. mojavensis* could occur while utilizing the necrotic cactus substrate, but as has been previously suggested [91], via sexual transmission.

A number of the TOP10 loci in this study perform functions associated with sensory perception and behavior (Fig. 6). *Drosophila mojavensis* larvae actively seek out patches of preferred yeast species [92] and across the four host populations there are distinct larval foraging strategies [93]. More specifically genes involved in chemosensory behavior were observed to have elevated ω values in these genomes. Across Drosophilids, there have been a number of studies indicating the links between the evolution of chemosensory genes and host specialization [94-96]. In *D. sechellia,* a specialist species, was found to be losing olfactory receptor genes at a faster rate than its sibling generalist species *D. simulans* [97]. In *D. mojavensis* each cactus species rot contains different compounds and thus have a different set of volatiles emanating from the necroses [39, 40]. These chemical differences have shaped the feeding and oviposition behavior of flies as has been shown by the exposure of adults to cactus volatiles [98-100]. Recent analysis of populations differentiation in odorant and gustatory receptors have shown that unlike what might be initially predicted a number of the changes in these receptors suggests that effects at the level of signal transduction in addition to odorant recognition [101]. Further functional analysis is needed to better understand the evolution and functional changes of chemosensory pathways associated with the adaptation to necrotic cacti.

In addition to their role in xenobiotic metabolism, serine proteases have been shown to be involved in the network of proteins associated with reproductive interactions in several taxa. In *D. melanogaster* accessory gland proteins (ACP), such as sex peptide, are found to perform a wide range of functions ranging from stimulating ovulation and reducing a female's remating rate to helping to defend against infections [102-104]. Knockouts of serine proteases have been shown to interfere with the behavioral and physiological effects of the male-derived sex peptide [104]. In *D. mojavensis* and its sister species *D. arizonae* a large number of proteases are expressed in female reproductive tracts and several have been shown to be under strong positive selection [73, 105-107]. In addition to ACPs being transferred via the ejaculate, gene transcripts have been found to be deposited by males into females during copulation [72]. Some of these male-derived transcripts could alter the female's transcriptional response, while other may potentially be translated within females. Furthermore, the loci of several of these male-transferred transcripts show a pattern of strong and continuous positive selection, likely as the result of persistent sexual selection [71]. While there seems to be no postzygotic effects of sexual isolation within the *D. mojavensis* populations there is some evidence of prezygotic isolation, where certain populations prefers to mate with members of its own population [76]. The pattern of positive selection and/or elevated rate of molecular evolution for proteases and reproductive loci in the present study may highlight the continuing genomic consequence of sexual selection in this species.

Methods

Drosophila mojavensis lines and sample preparation

Fly lines MJBC 155 collected in La Paz, Baja California in February 2001, MJ 122 collected in Guaymas, Sonora in 1998, and MJANZA 402-8 collected in ANZA-Borrego Park, California in April 2002 were used as the source lines for the sequencing of three *D. mojavensis* populations. These lines were highly inbred to reduce the heterozygosity of their DNA. Summary of the karyotype of each of the lines sequenced as well as the Catalina Island template genome stock (15081-1352.00) can be found in Additional file 2: Table S9. The flies were grown for two generations in banana molasses media [93] supplemented with ampicillin (125 µg/ml) and tetracycline (12.5 µg/ml), to prevent the isolation of bacterial DNA in addition to the flies'. DNA was extracted from homogenized whole male flies using a combination of phenol/chloroform DNA extraction and Qiagen DNeasy spin-columns to achieve the required amount of DNA material. RNase A was used to reduce RNA contamination. Gel electrophoresis was run on each sample to check the quality of the extraction. Any samples with RNA contamination were run through a Qiagen QIAquick PCR Purification Kit spin column to filter contaminates. Extracted DNA was sent to the HudsonAlpha Institute for Biotechnology Genomic Services Lab (Huntsville, Alabama) for sequencing. One hundred base pair paired-end and mate pair sequencing was done on an Illumina HiSeq 2500 with one lane for each.

Genome assembly

Paired-end and mate pair Illumina reads were filtered and trimmed using step one of the A5 Pipeline [108]. This step uses SGA [109] and TagDust [110] with the quality scores from the Illumina FASTQ files to reduce the number of low quality reads. A5 was run on the Dense Memory Cluster of the Alabama Super Computer Center with four processing cores and 64 gigabytes of memory allocated for each run. With the reads cleaned they were assembled to the template genome. The reference genome of the Catalina Island population of *D. mojavensis* was assembled as part of the *Drosophila* 12 Genomes Consortium [17]. Version 1.04 of the reference genome was retrieved from FlyBase version FB2015_02 [111]. From the reference sequence, genome scaffolds [112] containing the protein-coding genes previously mapped to a chromosome, were extracted for use as a template for the assembly; these scaffolds are detailed in Additional file 2: Table S10. The reference templates as well as the Illumina reads were imported into Geneious 8.1. Assembly was done separately for paired-end and mate pair data. Using Geneious 8.1 and its Map to Reference feature the cleaned reads were assembled to each of the template scaffolds. BAM files were exported for each paired-end and mate pair assembly. SAMtools [113] was used to merge BAM files to create an assembly with both types of reads. This merged BAM file was imported into Geneious 8.1 where consensus sequences were determined for each scaffold using majority calling to limit the number of ambiguities. GTF files for each scaffold used were retrieved from FlyBase version FB2015_02 [111]. These annotations were transferred to each of the new genomes by aligning each assembled genome scaffold to the reference genome scaffold using Mauve Genome Alignment [114] with default settings except for selecting assume collinear genomes. After alignment, annotations were transferred from the reference to the new assembly. The resulting scaffolds were exported in GenBank format. Using the EMBOSS program, extractfeat [115], CDS sequences were extracted from the assembled scaffolds. Sequence files for each gene were concatenated and then aligned using the default settings of the aligner Muscle 3.8.31 [116]. Only the longest transcript for each gene was used as some genes have multiple splice variants.

Molecular evolution analysis

To generate substitution counts for filtering, the software KaKs Calculator 1.2 [45] was used. Files of aligned genes were converted to AXT format using the Perl script parseFastaIntoAXT.pl including in the package. After conversion each gene was run through the software using the NG method [117]. The output files for each loci were concatenated and then imported into JMP 10 for filtering.

Values for ω were calculated using codeml part of the PAML 4.9 package [46]. Aligned genes were converted to PHYLIP format using BioPerl [118]. As PAML requires a phylogenetic tree to be provided for its calculations a neighbor joining tree was constructed in MEGA 5 [119]. This was done by concatenating all exons from each population and then aligning them using Mauve Genome Alignment [114]. The alignment was converted to MEG format using MEGA and a neighbor joining tree was built using the default settings. The tree was exported in newick format for use by PAML. Genes were removed from analysis if they were not divisible by three, these genes were manually screened and if alignment errors appeared to be the cause, these were manually corrected. Screening was done for stop codons within the sequences by translating the DNA sequence to protein sequence with Transeq, part of the EMBOSS package [115] and any genes with internal stop codons were removed.

Using the BioPython PAML module [120], control files were built for each gene alignment with default values taken except codon frequency was set to F3x4. Site-class models 0 , 7, and 8 were used to calculate the ω values [121-123]. Model 0 is a single ratio based omega value for the entire gene. Model 7 is a null model with 10 classes, which does not allow for positive selection while model 8 adds an additional class that allows for positive selection. Both the ω values and log likelihood values were extracted from each output file and the data was organized in Microsoft Excel. If model 8 significantly better fits the data this is evidence of positive selection [46]. Significance values were found by taking the difference between the log likelihood values of the two outputs and multiplying them by two. This value was then compared a chi-square distribution to find *P* values for each gene. Genes with less than five total substitutions as determined by KaKs Calculator [45] were filtered out and not considered. This was done to help deal with the low power of these methods when there are very few changes between the populations. Genes with few changes are more likely to cause the software to either return an undefined result or to reach the maximum ω the software allows. In addition, genes with either no nonsynonymous or no synonymous changes were also removed. This yielded a total of 9,087 genes that were used in the analysis. Histograms of a $\log_2$ transformation of the ω values were produced using JMP 10. A comparison between the $\log_2$ transformations of the NG Ka/Ks and the omega value from model 0 of codeml was generated with JMP 10.

The length of each gene's coding sequence was extracted from the PHYLIP sequence headers. This was to determine if genes with longer length have significantly different omega values. Genes were binned based on length and an ANOVA with post-hoc Tukey test using JMP 10 was used to compare length bins for significance. Intron data was extracted from the reference genome annotation using Geneious 8.1. Based on this, genes were binned based on the number of exons. ANOVA with post-hoc Tukey test in JMP 10 compared the bin sets for significant difference in omega. To determine if there was a significant difference in omega between genes present on each Muller element ANOVA with post-hoc Tukey test was used in JMP 10 to compare omega value distribution on each element.

Expression analysis

Previous transcriptional studies provided differential expression data for cactus host shifts [43] and between populations [44]. Loci that were found to be significant with codeml model 7 and 8 were removed from this analysis. The model 0 omega for loci with a FDR significance greater than 0.001 for third-instar larva from the *D. mojavensis* Sonora population that were raised on agria cactus rot was compared to non-significant loci using ANOVA in JMP 10. Comparison of model 0 omega between FDR significant loci and non-significant loci was also done for differential expression between third-instar larva of the four host populations with ANOVA in JMP 10.

To explore the relationship between omega and gene expression level RNAseq data from [47] was retrieved for whole male and female *D. mojavensis* flies as aligned BAM files. Differential expression was calculated by using edgeR [124] to look for genes with significantly higher male or female expression. Box plots of omega model 0 for genes with significant male or female expressed genes as well as genes without sex based expression were compared using ANOVA with post-hoc Tukey test in JMP 10. Average adjusted (+0.25) $\log_2$ RPKM of non-sex biased genes was plotted against $\log_2$ omega model 0 and linear regression was performed on the data with JMP 10.

Gene ontology terms analysis

Network graphs were generated using Cytoscape 3.2.1 [125] with the add-on app ClueGO 2.2.5 [126]. GO term and KEGG pathway data used was from the June 2016 release. The custom *D. melanogaster* reference set was used for analysis. Both the TOP10 and PAML-FDR genes were run on, biological processes, molecular function and KEGG terms. Data for GO term summary tables was retrieved from FlyBase version FB2017_06 *D. melanogaster* release 6.19 [111]. For each *D. mojavensis* gene with a *D. melanogaster* ortholog, GO term summaries were phrased from the FlyBase GO Summary Ribbons for molecular function and biological process. Clustering done with JMP 10 using the Ward method and 15 groups allowed.

Abbreviations

X

Availability of data and materials

X

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CWA performed the assembly and analysis of the genomic data and was involved in the writing of the manuscript. LMM conceived of and designed the study, was involved in the analysis and the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors greatly acknowledge the work of Laurel Brandsmeier in this project. This work was supported by a Junior Faculty Distinguished Research award from the University of Alabama in Huntsville and partly supported by a grant from the National Science Foundation (DEB-1219387 and IOS-1557697 to LMM).

References

1. Feder ME, Mitchell-Olds T. Evolutionary and ecological functional genomics. Nature Review Genetics 2003, 4:649-655.
2. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J. Adaptation genomics: the next generation. Trends in Ecology & Evolution 2010, 25(12):705-712.
3. Barrett RDH, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nature Reviews Genetics 2011, 12(11):767-780.
4. Storz JF, Wheat CW. Integrating Evolutionary and Functional Approaches to Infer Adaptation at Specific Loci. Evolution 2010, 64(9):2489-2509.
5. Ungerer MC, Johnson LC, Herman MA. Ecological genomics: understanding gene and genome function in the natural environment. Heredity 2008, 100(2):178-183.
6. Nosil P. Ecological Speciation. Oxford: Oxford University Press; 2012.
7. Rundle HD, Nosil P. Ecological speciation. Ecology Letters 2005, 8(3):336-352.
8. Funk DJ. Isolating a role for natural selection in speciation: Host adaptation and sexual isolation in *Neochlamisus bebbianae* leaf beetles. Evolution 1998, 52(6):1744-1759.
9. Wu CI, Ting CT. Genes and speciation. Nat Rev Genet 2004, 5(2):114-122.
10. Feder JL, Opp SB, Wlazlo B, Reynolds K, Go W, Spisak S. Host Fidelity Is an Effective Premating Barrier between Sympatric Races of the Apple Maggot Fly. Proc Natl Acad Sci 1994, 91(17):7990-7994.

11. Funk DJ, Egan SP, Nosil P. Isolation by adaptation in Neochlamisus leaf beetles: host-related selection promotes neutral genomic divergence. Mol Ecol 2011, 20(22):4671-4682.
12. Egan SP, Janson EM, Brown CG, Funk DJ. Postmating isolation and genetically variable host use in ecologically divergent host forms of Neochlamisus bebbianae leaf beetles. J Evol Biol 2011, 24(10):2217-2229.
13. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ *et al*. A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biology 2005, 3(6):976-985.
14. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B *et al*. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 2003, 302(5652):1960-1963.
15. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD *et al*. Natural selection on protein-coding genes in the human genome. Nature 2005, 437(7062):1153-1157.
16. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. Patterns of Positive Selection in Six Mammalian Genomes. Plos Genet 2008, 4(8).
17. Consortium DG. Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007, 450(7167):203-218.
18. Yang Z. The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci 2005, 102(9):3179-3180.

19. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ *et al*. Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. Plos Genet 2012, 8(12):e1003080.

20. Langley CH, Stevens K, Cardeno C, Lee YC, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B *et al*. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 2012, 192(2):533-598.

21. Bergland AO, Tobler R, Gonzalez J, Schmidt P, Petrov D. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. Mol Ecol 2016, 25(5):1157-1174.

22. Campo D, Lehmann K, Fjeldsted C, Souaiaia T, Kao J, Nuzhdin SV. Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. Mol Ecol 2013, 22(20):5084-5097.

23. Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG. Global Diversity Lines-A Five-Continent Reference Panel of Sequenced *Drosophila melanogaster* Strains. G3-Genes Genom Genet 2015, 5(4):593-603.

24. Pool JE. The Mosaic Ancestry of the *Drosophila* Genetic Reference Panel and the *D. melanogaster* Reference Genome Reveals a Network of Epistatic Fitness Interactions. Mol Biol Evol 2015, 32(12):3236-3251.

25. Shiao MS, Chang JM, Fan WL, Lu MY, Notredame C, Fang S, Kondo R, Li WH. Expression Divergence of Chemosensory Genes between *Drosophila sechellia* and

Its Sibling Species and Its Implications for Host Shift. Genome Biol Evol 2015, 7(10):2843-2858.

26. Chiu JC, Jiang XT, Zhao L, Hamm CA, Cridland JM, Saelao P, Hamby KA, Lee EK, Kwok RS, Zhang GJ *et al*. Genome of Drosophila suzukii, the Spotted Wing Drosophila. G3-Genes Genom Genet 2013, 3(12):2257-2271.

27. Matzkin LM. Ecological Genomics of Host Shifts in *Drosophila mojavensis*. Adv Exp Med Biol 2014, 781(781):233-247.

28. Heed WB. Ecology and genetics of Sonoran desert *Drosophila*. In: Ecological genetics: The interface. Edited by Brussard PF: Springer-Verlag; 1978: 109-126.

29. Reed LK, Nyboer M, Markow TA. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol Ecol 2007, 16(5):1007-1022.

30. Matzkin LM, Eanes WF. Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. Genetics 2003, 163:181-194.

31. Matzkin LM. The Molecular Basis of Host Adaptation in Cactophilic Drosophila: Molecular Evolution of a Glutathione S-Transferase Gene (*GstD1*) in *Drosophila mojavensis*. Genetics 2008, 178(2):1073-1083.

32. Matzkin LM. Population genetics and geographic variation of alcohol dehydrogenase (*Adh*) paralogs and glucose-6-phosphate dehydrogenase (*G6pd*) in *Drosophila mojavensis*. Mol Biol Evol 2004, 21(2):276-285.

33. Smith G, Lohse K, Etges WJ, Ritchie MG. Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*. Mol Ecol 2012, 21(13):3293-3307.

34. Starmer WT. Analysis of the Community Structure of Yeasts Associated with the Decaying Stems of Cactus. I. *Stenocereus gummosus*. Microb Ecol 1982, 8(1):71-81.
35. Starmer WT. Associations and Interactions Among Yeasts, Drosophila and Their Habitats. In: Ecological genetics and evolution: The cactus-yeast-Drosophila model system. Edited by Barker JSF, Starmer WT. New York: Academic Press; 1982: 159-174.
36. Fogleman JC, Starmer WT. Analysis of the community structure of yeasts associated with the decaying stems of cactus. III. *Stenocereus thurberi*. Microb Ecol 1985, 11(2):165-173.
37. Starmer WT, Lachance MA, Phaff HJ, Heed WB. The biogeography of yeasts Associated with decaying cactus tissue in North America, the Caribbean, and Northern Venezuela. In: Evolutionary Biology. Edited by Hecht MK, Wallace B, Macintyre RJ, vol. 24: Plenum Publishing Corporation; 1990: 253-296.
38. Fellows DF, Heed WB. Factors affecting host plant selection in desert-adapted cactiphilic *Drosophila*. Ecology 1972, 53:850-858.
39. Kircher HW. Chemical composition of cacti and its relationship to Sonoran Desert Drosophila. In: Ecological genetics and evolution: The cactus-yeast-Drosophila model system. Edited by Barker JSF, Starmer WT. New York: Academic Press; 1982: 143-158.
40. Fogleman JC, Abril JR. Ecological and evolutionary importance of host plant chemistry. In: Ecological and evolutionary genetics of Drosophila. Edited by Barker JSF, Starmer WT, MacIntyre RJ. New York: Plenum Press; 1990: 121-143.

41. Fogleman JC, Danielson PB. Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. American Zoologist 2001, 41(4):877-889.
42. Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. Functional genomics of cactus host shifts in *Drosophila mojavensis*. Mol Ecol 2006, 15:4635-4643.
43. Matzkin LM. Population transcriptomics of cactus host shifts in *Drosophila mojavensis*. Mol Ecol 2012, 21(10):2428-2439.
44. Matzkin LM, Markow TA. Transcriptional differentiation across the four cactus host races of *Drosophila mojavensis*. In: Speciation: Natural Processes, Genetics and Biodiversity. Edited by Michalak P. Hauppauge: Nova Science Publishers Inc.; 2013: 119-136.
45. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. Genomics, proteomics & bioinformatics 2006, 4(4):259-263.
46. Yang ZH. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007, 24(8):1586-1591.
47. Graveley BR, Brooks AN, Carlson J, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW *et al*. The developmental transcriptome of *Drosophila melanogaster*. Nature 2011, 471(7339):473-479.
48. Wasserman M. Cytological and Phylogenetic Relationships in the Repleta Group of the Genus Drosophila. Proc Natl Acad Sci 1960, 46(6):842-859.
49. Riddle NC, Elgin SCR. The Drosophila Dot Chromosome: Where Genes Flourish Amidst Repeats. Genetics 2018, 210(3):757-772.

50. Bridges CB. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. J Hered 1935, 26(2):60-64.

51. Singh ND, Petrov DA. Evolution of Gene Function on the X Chromosome Versus the Autosomes. Gene and Protein Evolution 2007, 3:101-118.

52. Thornton K, Bachtrog D, Andolfatto P. X chromosomes and autosomes evolve at similar rates in Drosophila: No evidence for faster-X protein evolution. Genome Research 2006, 16(4):498-504.

53. Leung W, Shaffer CD, Reed LK, Smith ST, Barshop W, Dirkes W, Dothager M, Lee P, Wong J, Xiong D *et al*. *Drosophila* Muller F elements maintain a distinct set of genomic properties over 40 million years of evolution. G3 2015, 5(5):719-740.

54. Ruiz A, Heed WB, Wasserman M. Evolution of the Mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. J Hered 1990, 81:30-42.

55. Hasson E, Eanes WF. Contrasting histories of three gene regions associated with In(3L)Payne of *Drosophila melanogaster*. Genetics 1996, 144(4):1565-1575.

56. Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. Evolution of protein-coding genes in Drosophila. Trends in Genetics 2008, 24(3):114-123.

57. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci 2000, 25(3):106-110.

58. Hawkins JD. A Survey on Intron and Exon Lengths. Nucleic Acids Res 1988, 16(21):9893-9908.

59. Comeron JM, Guthrie TB. Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in Drosophila. Mol Biol Evol 2005, 22(12):2519-2530.

60. Hill WG, Robertson A. Effect of Linkage on Limits to Artificial Selection. Genet Res 1966, 8(3):269-294.

61. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci 2005, 102(40):14338-14343.

62. Wilke CO, Drummond DA. Population genetics of translational robustness. Genetics 2006, 173(1):473-481.

63. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics 2001, 158(2):927-931.

64. Akashi H. Gene expression and molecular evolution. Curr Opin Genet Dev 2001, 11(6):660-666.

65. Nuzhdin S, Wayne M, Harmon K, McIntyre L. Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol 2004, 21(7):1308-1317.

66. Zhang Z, Hambuch TM, Parsch J. Molecular evolution of sex-biased genes in *Drosophila*. Mol Biol Evol 2004, 21(11):2130-2139.

67. Grath S, Parsch J. Sex-Biased Gene Expression. Annu Rev Genet 2016, 50:29-44.

68. Meisel RP. Towards a More Nuanced Understanding of the Relationship between Sex-Biased Gene Expression and Rates of Protein-Coding Sequence Evolution. Mol Biol Evol 2011, 28(6):1893-1900.

69. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nature Reviews Genetics 2002, 3(2):137-144.

70. Bono JM, Markow TA. Post-zygotic isolation in cactophilic *Drosophila*: larval viability and adult life-history traits of *D. mojavensis*/*D. arizonae* hybrids. J Evol Biol 2009, 22(7):1387-1395.

71. Bono JM, Matzkin LM, Hoang K, Brandsmeier L. Molecular evolution of candidate genes involved in post-mating-prezygotic reproductive isolation. J Evol Biol 2015, 28(2):403-414.

72. Bono JM, Matzkin LM, Kelleher ES, Markow TA. Postmating transcriptional changes in reproductive tracts of con- and heterospecifically mated *Drosophila mojavensis* females. Proc Natl Acad Sci 2011, 108(19):7878-7883.

73. Kelleher ES, Markow TA. Reproductive tract interactions contribute to isolation in Drosophila. Fly 2007, 1(1):33-37.

74. Knowles LL, Markow TA. Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert Drosophila. Proc Natl Acad Sci 2001, 98(15):8692-8696.

75. Krebs RA, Markow TA. Courtship behavior and control of reproductive isolation in *Drosophila mojavensis*. Evolution 1989, 43:908-913.

76. Markow TA. Sexual isolation among populations of *Drosophila mojavensis*. Evolution 1991, 45:1525-1529.

77. Pitnick S, Miller GT, Schneider K, Markow TA. Ejaculate-female coevolution in Drosophila mojavensis. P Roy Soc B-Biol Sci 2003, 270(1523):1507-1512.

78. Etges WJ, Heed WB. Sensitivity to larval density in populations of *Drosophila mojavensis*: Influences of host plant variation on components fitness. Oecologia 1987, 71:375-381.

79. Etges WJ. Direction of life history evolution in *Drosophila mojavensis*. In: Ecological and evolutionary genetics of Drosophila. Edited by Barker JSF, Starmer WT, MacIntyre RJ. New York: Plenum Press; 1990: 37-56.

80. Casida JE, Quistad GB. Serine hydrolase targets of organophosphorus toxicants. Chem-Biol Interact 2005, 157:277-283.

81. Luque T, O'Reilly DR. Functional and phylogenetic analyses of a putative *Drosophila melanogaster* UDP-glycosyltransferase gene. Insect Biochem Mol Biol 2002, 32(12):1597-1604.

82. Ranson H, Rossiter L, Ortelli F, Jensen B, Wang XL, Roth CW, Collins FH, Hemingway J. Identification of a novel class of insect Glutathione S-transferases involved in resistance to DDT in the malaria vector *Anopheles gambiae*. Biochem J 2001, 359:295-304.

83. Ranson H, Hemingway J. Glutathione Transferases. In: Comprehensive Molecular Insect Science. Edited by Gilbert LI, Iatrou K, Gill SS, vol. 5. Amsterdam: Elsevier; 2005: 383-402.

84. Feyereisen R. Insect Cytochrome P450. In: Comprehensive Molecular Insect Science. Edited by Gilbert LI, Iatrou K, Gill SS, vol. 4. Amsterdam: Elsevier; 2005: 1-77.

85. Li XC, Schuler MA, Berenbaum MR. Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol 2007, 52:231-253.

86. Guillen Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, Casillas S, Ramia M, Egea R, Negre B *et al*. Genomics of Ecological Adaptation in Cactophilic Drosophila. Genome Biology and Evolution 2015, 7(1):349-366.

87. Rane RV, Pearce SL, Li F, Coppin C, Schiffer M, Shirriffs J, Sgro CM, Griffin PC, Zhang G, Lee SF *et al*. Genomic changes associated with adaptation to arid environments in cactophilic *Drosophila* species. BMC Genomics 2019, 20(1):52.

88. Foster JLM, Fogleman JC. Identification and Ecology of Bacterial Communities Associated with Necroses of 3 Cactus Species. Appl Environ Microb 1993, 59(1):1-6.

89. Foster J, Fogleman J. Bacterial succession in necrotic tissue of Agria cactus (*Stenocereu gummosus*). Appl Environ Microb 1994, 60(2):619-625.

90. Schlenke T, Begun D. Natural selection drives *Drosophila* immune system evolution. Genetics 2003, 164(4):1471-1480.

91. Markow TA. Assortative fertilization in *Drosophila*. Proc Natl Acad Sci 1997, 94(15):7756-7760.

92. Fogleman JC, Starmer WT, Heed WB. Larval Selectivity for Yeast Species by *Drosophila mojavensis* in Natural Substrates. Proc Natl Acad Sci 1981, 78(7):4435-4439.

93. Coleman JM, Benowitz KM, Jost AG, Matzkin LM. Behavioral evolution accompanying host shifts in cactophilic *Drosophila* larvae. Ecology and Evolution 2018, 8(14):6921-6931.

94. Vosshall LB, Stocker RE. Molecular architecture of smell and taste in *Drosophila*. Annu Rev Neurosci 2007, 30:505-533.

95. McBride CS, Arguello JR. Five *Drosophila* genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics 2007, 177(3):1395-1416.

96. Arguello JR, Cardoso-Moreira M, Grenier JK, Gottipati S, Clark AG, Benton R. Extensive local adaptation within the chemosensory system following *Drosophila melanogaster*'s global expansion. Nature Communications 2016, 7.

97. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proc Natl Acad Sci 2007, 104(12):4996-5001.

98. Newby BD, Etges WJ. Host preference among populations of *Drosophila mojavensis* (Diptera: Drosophilidae) that use different host cacti. Journal of Insect Behavior 1998, 11(5):691-712.

99. Date P, Dweck HKM, Stensmyr MC, Shann J, Hansson BS, Rollmann SM. Divergence in Olfactory Host Plant Preference in *D. mojavensis* in Response to Cactus Host Use. Plos One 2013, 8(7).

100. Date P, Crowley-Gall A, Diefendorf AF, Rollmann SM. Population differences in host plant preference and the importance of yeast and plant substrate to volatile composition. Ecology and Evolution 2017, 7(11):3815-3825.

101. Diaz F, Allan CW, Matzkin LM. Positive selection at sites of chemosensory genes is associated with the recent divergence and local ecological adaptation in cactophilic *Drosophila*. BMC Evol Biol 2018, 18.

102. Wolfner MF. The gifts that keep on giving: Physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity 2002, 88:85-93.

103. Avila FW, Sirot LK, LaFlamme BA, Rubinstein CD, Wolfner MF. Insect seminal fluid proteins: identification and function. Annu Rev Entomol 2011, 56:21-40.

104. Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. Plos Genet 2014, 10(1):e1004108.

105. Kelleher ES, Pennington JE. Protease gene duplication and proteolytic activity in *Drosophila* female reproductive tracts. Mol Biol Evol 2009, 26(9):2125-2134.

106. Kelleher ES, Swanson WJ, Markow TA. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. Plos Genet 2007, 3(8):1541-1549.

107. Kelleher ES, Watts TD, LaFlamme BA, Haynes PA, Markow TA. Proteomic analysis of *Drosophila mojavensis* male accessory glands suggests novel classes of seminal fluid proteins. Insect Biochem Mol Biol 2009, 39(5-6):366-371.

108. Tritt A, Eisen JA, Facciotti MT, Darling AE. An Integrated Pipeline for de Novo Assembly of Microbial Genomes. Plos One 2012, 7(9).

109. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Research 2012, 22(3):549-556.

110. Lassmann T, Hayashizaki Y, Daub CO. TagDust-A program to eliminate artifacts from next generation sequencing data. Bioinformatics 2009, 25(21):2839-2840.

111. Gramates LS, Marygold SJ, dos Santos G, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB *et al*. FlyBase at 25: Looking to the future. Nucleic Acids Res 2017, 45(D1):D663-D671.

112. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW *et al*. Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics 2008, 179(3):1601-1655.

113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.

114. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004, 14(7):1394-1403.

115. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. Trends in Genetics 2000, 16(6):276-277.

116. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32(5):1792-1797.

117. Nei M, Gojobori T. Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions. Mol Biol Evol 1986, 3(5):418-426.

118. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H *et al*. The bioperl toolkit: Perl modules for the life sciences. Genome Research 2002, 12(10):1611-1618.

119. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 2011, 28(10):2731-2739.

120. Talevich E, Invergo BM, Cock PJA, Chapman BA. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics 2012, 13.

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

122. Goldman N, Yang ZH. Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA-Sequences. Mol Biol Evol 1994, 11(5):725-736.

123. Yang ZH, Nielsen R, Goldman N, Pedersen AMK. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000, 155(1):431-449.

124. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26(1):139-140.

125. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 2003, 13(11):2498-2504.

126. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009, 25(8):1091-1093.

Table 1 Number of cleaned reads and assembled reads for each population.

Population	Reads Mapped	Total Reads	Proportion Mapped
Baja California			
ME	12,052,662	44,912,130	0.27
PE	88,976,029	100,263,663	0.89
Total	101,028,691	145,175,793	0.70
Mojave			
ME	26,638,794	52,910,406	0.50
PE	73,196,313	83,000,942	0.88
Total	99,835,107	135,911,348	0.73
Sonora			
ME	39,962,094	63,240,890	0.63
PE	93,857,309	105,723,406	0.89
Total	133,819,403	168,964,296	0.79

ME mate pair end reads; PE paired end reads

Figure legends

Fig. 1 Distribution of the four cactus host populations of *D. mojavensis.*

Fig. 2 Boxplot of $\log_2 \omega$ values for loci located in each of the *D. mojavensis* Muller elements. Elements with different letters are significantly different using a Tukey HSD test (see Table S2).

Fig. 3 Boxplot of $\log_2 \omega$ values for loci in five different coding length bins. Bins with different letters are significantly different using a Tukey HSD test (see Table S3).

Fig. 4 Proportion of TOP10 loci that show female-bias, male-bias or unbiased gene expression. Dashed line indicates the genome wide proportion of TOP10 loci (0.10). Gene expression data is from (Gravely et al 2011). Asterisk indicate significance via Fisher's Exact test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

Fig. 5 Functional clustering of Biological Process GO terms of the TOP10 loci. Details of gene composition of each cluster is in Additional file 3: Table S11.

Fig. 6 Network clustering of Biological Process GO terms of the TOP10 loci. Network clustering was performed using ClueGo using the following parameters: Min GO Level = 3, Max GO Level = 8, All GO Levels = false, Number of Genes = 3, Get All Genes = false, Min Percentage = 5.0, Get All Percentage = false, GO Fusion = true, GO Group = true, Kappa Score Threshold = 0.3, Over View Term = Smallest PValue, Group By Kappa Statistics = true, Initial Group Size = 1, Sharing Group Percentage = 50.0.

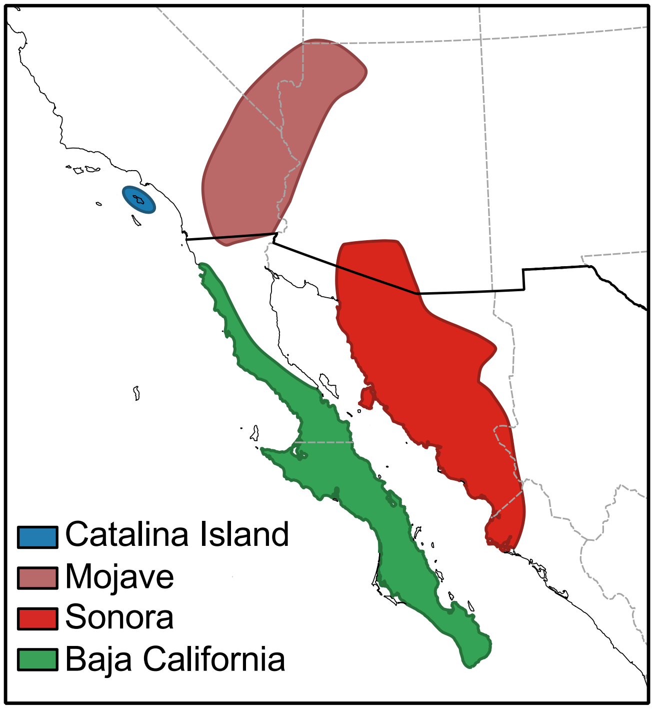

Fig. 1 Distribution of the four cactus host populations of *D. mojavensis.*

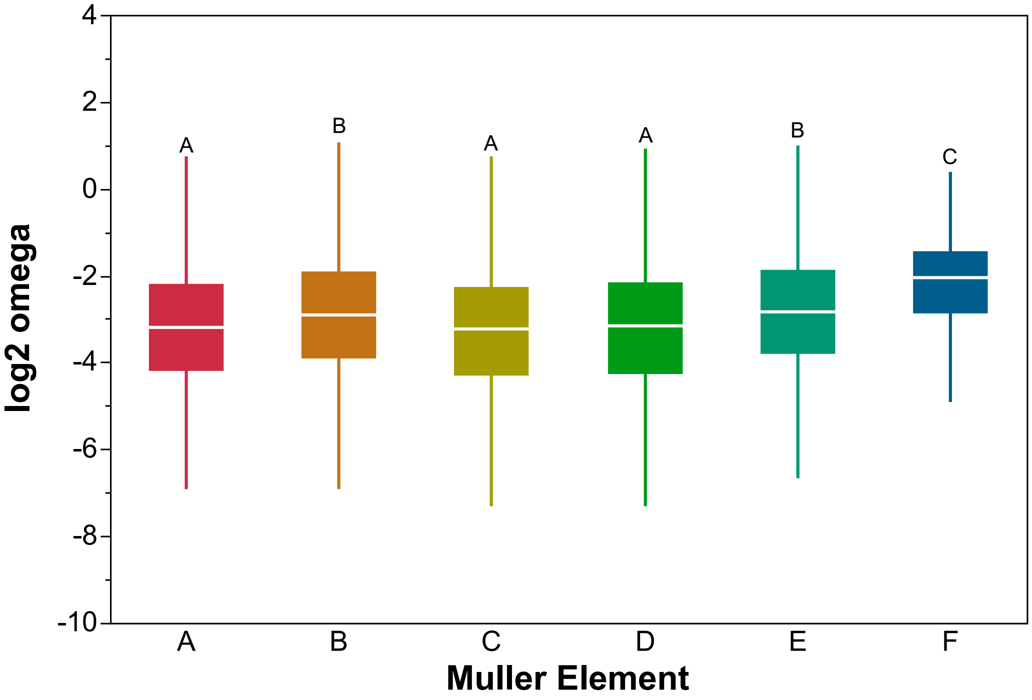

Fig. 2 Boxplot of $\log_2$ ω values for loci located in each of the *D. mojavensis* Muller elements. Elements with different letters are significantly different using a Tukey HSD test (see Table S2).

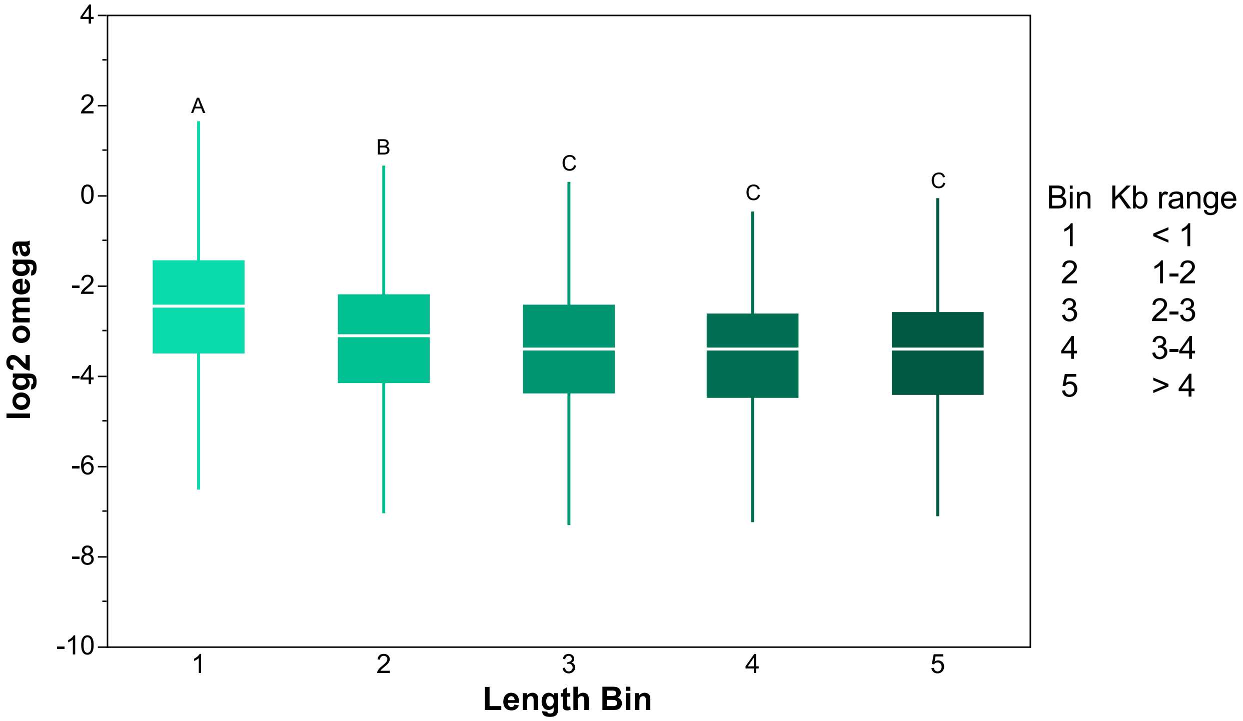

Fig. 3 Boxplot of $\log_2 \omega$ values for loci in five different coding length bins. Bins with different letters are significantly different using a Tukey HSD test (see Table S3).

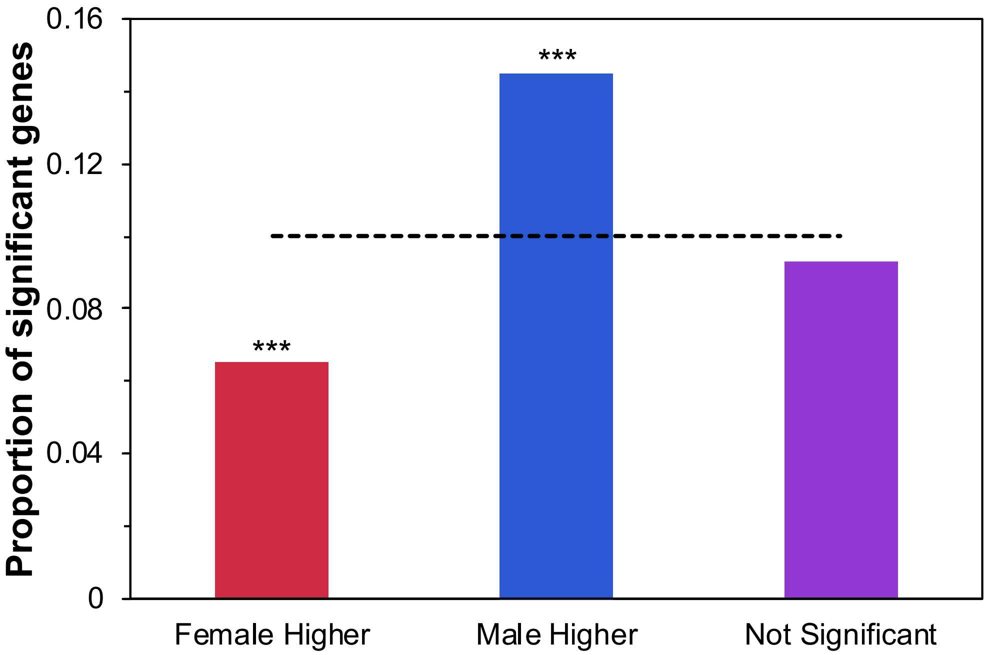

Fig. 4 Proportion of TOP10 loci that show female-bias, male-bias or unbiased gene expression. Dashed line indicates the genome wide proportion of TOP10 loci (0.10). Gene expression data is from (Gravely et al 2011). Asterisk indicate significance via Fisher's Exact test (* P < 0.05, ** P < 0.01, *** P < 0.001).

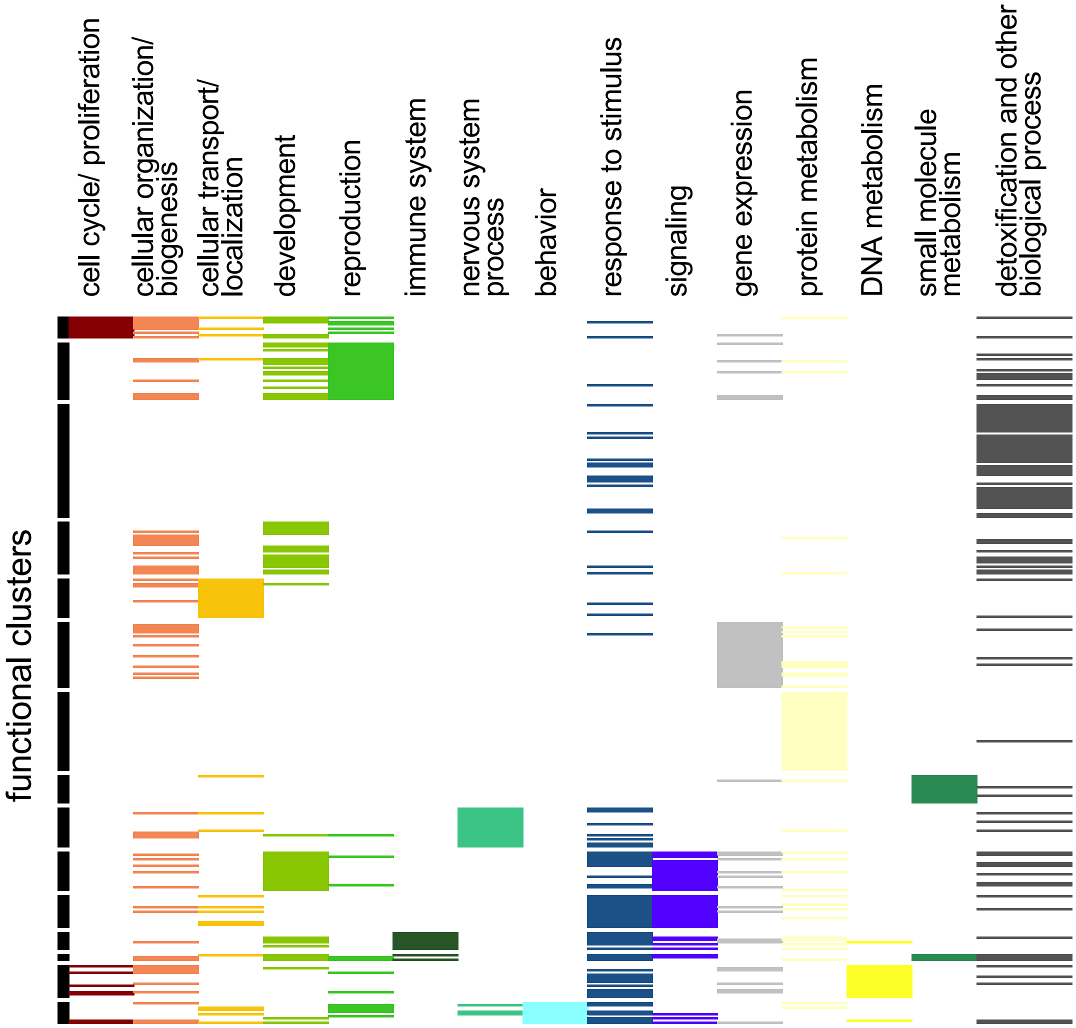

Fig. 5 Functional clustering of Biological Process GO terms of the TOP10 loci. Details of gene composition of each cluster is in Additional file 3: Table S11.

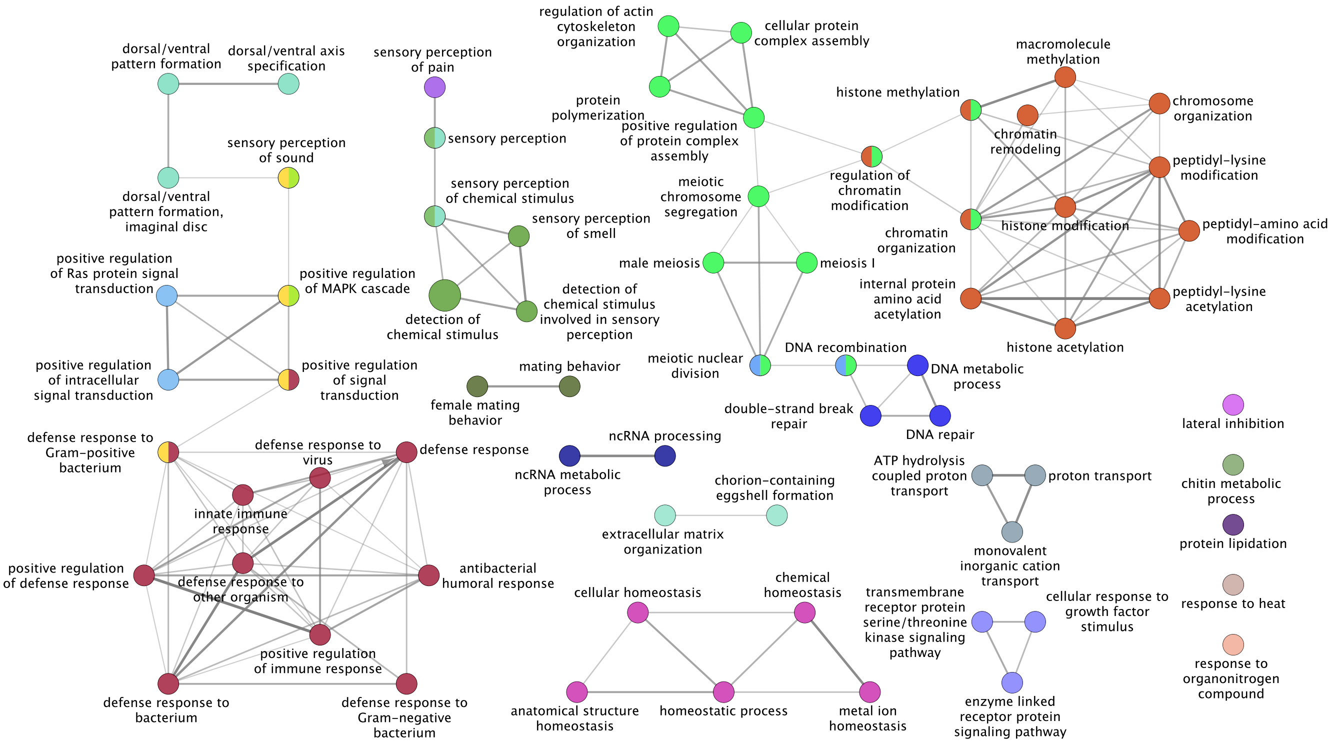
regulation of actin cytoskeleton organization
cellular protein complex assembly
dorsal/ventral pattern formation
dorsal/ventral axis specification
sensory perception of pain
macromolecule methylation
protein polymerization
histone methylation
positive regulation of protein complex assembly
chromosome organization
sensory perception
chromatin remodeling
sensory perception of sound
regulation of chromatin modification
peptidyl-lysine modification
dorsal/ventral pattern formation, imaginal disc
sensory perception of chemical stimulus
meiotic chromosome segregation
sensory perception of smell
histone modification
peptidyl-amino acid modification
chromatin organization
positive regulation of Ras protein signal transduction
positive regulation of MAPK cascade
male meiosis
meiosis I
detection of chemical stimulus involved in sensory perception
internal protein amino acid acetylation
peptidyl-lysine acetylation
detection of chemical stimulus
histone acetylation
DNA recombination
meiotic nuclear division
DNA metabolic process
positive regulation of intracellular signal transduction
positive regulation of signal transduction
mating behavior
female mating behavior
double-strand break repair
DNA repair
lateral inhibition
ncRNA processing
defense response to Gram-positive bacterium
defense response to virus
defense response
ncRNA metabolic process
ATP hydrolysis coupled proton transport
proton transport
chorion-containing eggshell formation
chitin metabolic process
innate immune response
extracellular matrix organization
protein lipidation
monovalent inorganic cation transport
positive regulation of defense response
defense response to other organism
antibacterial humoral response
chemical homeostasis
cellular homeostasis
transmembrane receptor protein serine/threonine kinase signaling pathway
cellular response to growth factor stimulus
response to heat
positive regulation of immune response
response to organonitrogen compound
defense response to bacterium
defense response to Gram-negative bacterium
anatomical structure homeostasis
homeostatic process
metal ion homeostasis
enzyme linked receptor protein signaling pathway

Fig. 6 Network clustering of Biological Process GO terms of the TOP10 loci. Network clustering was performed using ClueGo using the following parameters: Min GO Level = 3, Max GO Level = 8, All GO Levels = false, Number of Genes = 3, Get All Genes = false, Min Percentage = 5.0, Get All Percentage = false, GO Fusion = true, GO Group = true, Kappa Score Threshold = 0.3, Over View Term = Smallest PValue, Group By Kappa Statistics = true, Initial Group Size = 1, Sharing Group Percentage = 50.0.