Genome Biology and Evolution Advance Access published December 31, 2014 doi:10.1093/gbe/evu291

Genomics of ecological adaptation in cactophilic Drosophila

Yolanda Guillén[1], Núria Rius[1], Alejandra Delprat[1], Anna Williford[2], Francesc Muyas[1], Marta Puig[1], Sònia Casillas[1,3], Miquel Ràmia[1,3], Raquel Egea[1,3], Barbara Negre[4,5], Gisela Mir[6,7], Jordi Camps[8], Valentí Moncunill[9], Francisco J. Ruiz-Ruano[10], Josefa Cabrero[10], Leonardo G. de Lima[11], Guilherme B. Dias[11], Jeronimo C. Ruiz[12], Aurélie Kapusta[13], Jordi Garcia-Mas[6], Marta Gut[8], Ivo G. Gut[8], David Torrents[9], Juan Pedro M. Camacho[10], Gustavo C.S. Kuhn[11], Cédric Feschotte[13], Andrew G. Clark[14], Esther Betrán[2], Antonio Barbadilla[1,3] and Alfredo Ruiz[1*]

* Correspondig author

1 Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

2 Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA.

3 Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain.

4 EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.

5 Universitat Pompeu Fabra (UPF), Barcelona, Spain.

6 IRTA, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, 08193 Bellaterra (Barcelona), Spain.

7 The Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia.

8 Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Torre I, Baldiri Reixac 4, 08028 Barcelona, Spain.

9 Barcelona Supercomputing Center (BSC), Edifici TG (Torre Girona), Jordi Girona 31, 08034 Barcelona, Spain.

10 Departamento de Genética, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain.

11 Instituto de Ciências Biológicas, Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte (MG), Brazil.

12 Informática de Biossistemas, Centro de Pesquisas René Rachou - Fiocruz Minas, Belo Horizonte (MG), Brazil.

13 Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA.

14 Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA.

GENOME BIOLOGY AND EVOLUTION
SMBE

ABSTRACT

Cactophilic Drosophila species provide a valuable model to study gene-environment interactions and ecological adaptation. *D. buzzatii* and *D. mojavensis* are two cactophilic species that belong to the *repleta* group, but have very different geographical distributions and primary host plants. To investigate the genomic basis of ecological adaptation, we sequenced the genome and developmental transcriptome of *D. buzzatii* and compared its gene content to that of *D. mojavensis* and two other non-cactophilic Drosophila species in the same subgenus. The newly sequenced *D. buzzatii* genome (161.5 Mb) comprises 826 scaffolds (> 3 kb) and contains 13,657 annotated protein-coding genes. Using RNA-Seq data of five life-stages we found expression of 15,026 genes, 80% protein-coding genes and 20% ncRNA genes. In total, we detected 1,294 genes putatively under positive selection. Interestingly, among genes under positive selection in the *D. mojavensis* lineage, there is an excess of genes involved in metabolism of heterocyclic compounds that are abundant in Stenocereus cacti and toxic to nonresident Drosophila species. We found 117 orphan genes in the shared *D. buzzatii-D. mojavensis* lineage. In addition, gene duplication analysis identified lineage-specific expanded families with functional annotations associated with proteolysis, zinc ion binding, chitin binding, sensory perception, ethanol tolerance, immunity, physiology and reproduction. In summary we identified genetic signatures of adaptation in the shared *D. buzzatii-D. mojavensis* lineage, and in the two separate *D. buzzatii and D. mojavensis* lineages. Many of the novel lineage-specific genomic features are promising candidates for explaining the adaptation of these species to their distinct ecological niches.

Keywords:

Cactophilic Drosophila, genome sequence, ecological adaptation, positive selection, orphan genes, gene duplication.

INTRODUCTION

Drosophila species are saprophagous insects that feed and breed on a variety of fermenting plant materials, chiefly fruits, flowers, slime fluxes, decaying bark, leaves and stems, cactus necroses and fungi (Carson 1971). These substrates include bacteria and yeasts that decompose the plant tissues and contribute to the nutrition of larvae and adults (Starmer 1981; Begon 1982). Only two species groups use cacti as their primary breeding site: *repleta* (Oliveira et al. 2012) and *nannoptera* (Lang et al. 2014). Both species groups originated at the virilis-repleta radiation, 20-30 MYA (Throckmorton 1975; Morales-Hojas and Vieira 2012; Oliveira et al. 2012) but adapted independently to the cactus niche. The *cactus-yeast-Drosophila system* in arid zones provides a valuable model to investigate gene-environment interactions and ecological adaptation from genetic and evolutionary perspectives (Barker and Starmer 1982; Barker et al. 1990). Rotting cacti provide relatively abundant, predictable and long lasting resources that can sustain very large Drosophila populations. For instance, a single saguaro rot may weigh up to several tons, last for many months and sustain millions of Drosophila larvae and adults (Breitmeyer and Markow 1998). On the other hand, cacti are usually found in arid climates with middle to high temperatures that may impose desiccation and thermal stresses (Loeschcke et al. 1997; Hoffmann et al. 2003; Rajpurohit et al. 2013). Finally, some cacti may contain allelochemicals that can be toxic for Drosophila (see below). Thus, adaptation to use cacti as breeding sites must have entailed a fairly large number of changes in reproductive biology, behavior, physiology and biochemistry (Markow and O'Grady 2008).

We have sequenced the genome and developmental transcriptome of *D. buzzatii* to carry out a comparative analysis with those of *D. mojavensis*, *D. virilis* and *D. grimshawi* (Drosophila 12 Genomes Consortium et al. 2007). *D. buzzatii* and *D. mojavensis* are both cactophilic species that belong to the mulleri subgroup of the *repleta* group (Wasserman 1992; Oliveira et al. 2012), although they have very different geographical distributions and host plants (figure 1). *D. buzzatii* is a subcosmopolitan species which is found in four out of the six major biogeographic regions (David and Tsacas 1980). This species is originally from Argentina and Bolivia but now has a wide geographical distribution that includes other regions of South America, the Old World and Australia (Carson and Wasserman 1965; Fontdevila et al. 1981; Hasson et al. 1995; Manfrin and Sene 2006). It chiefly feeds and breeds in rotting tissues of several Opuntia cacti but can also occasionally use columnar cacti (Hasson et al. 1992; Ruiz et al. 2000; Oliveira et al. 2012). The geographical dispersal of Opuntia by humans in historical times is considered the main driver of the world-wide expansion of *D. buzzatii* (Fontdevila et al. 1981; Hasson et al. 1995).

On the other hand, *D. mojavensis* is endemic to the deserts of Southwestern USA and Northwestern Mexico. Its primary host plants are *Stenocereus gummosus* (pitaya agria) in Baja California and *Stenocereus thurberi* (organ pipe) in Arizona and Sonora, but uses also *Ferocactus cylindraceous* (California barrel) in Southern California and *Opuntia sp.* in Santa Catalina Island (Fellows and Heed 1972; Heed and Mangan 1986; Ruiz and Heed 1988; Etges et al. 1999). The ecological conditions of the Sonoran Desert are extreme (dry, arid and hot), as attested by the fact that only four Drosophila species are endemic (Heed and Mangan 1986). In

addition, *D. mojavensis* chief host plants, pitaya agria and organ pipe, are chemically complex and contain large quantities of triterpene glycosides, unusual medium-chain fatty acids and sterol diols (Kircher 1982; Fogleman and Danielson 2001). These allelochemicals are toxic to nonresident Drosophila species, decreasing significantly larval performance (Fogleman and Kircher 1986; Ruiz and Heed 1988; Fogleman and Armstrong 1989; Frank and Fogleman 1992). In addition, host plant chemistry and fermentation byproducts affect adult epicuticular hydrocarbons and mating behavior (Havens and Etges 2013) as well as expression of hundreds of genes (Matzkin et al. 2006; Etges et al. 2014; Matzkin 2014).

As a first step to understand the genetic bases of ecological adaptation, here we compare the genomes of the two cactophilic species with those of two non-cactophilic species of the Drosophila subgenus: *D. virilis* that belongs to the virilis species group and *D. grimshawi* that belongs to the picture wing group of Hawaiian Drosophila (figure 1). The lineage leading to the common ancestor of *D. buzzatii* and *D. mojavensis* after diverging from *D. virilis* (# 3 in figure 1) represents the lineage that adapted to the cactus niche (likely Opuntia; Oliveira et al. 2012), whereas the lineages leading to D. buzzatii (# 1) and D. mojavensis (# 2) adapted to the specific niche of each species. We carried out a genome-wide scan for (i) genes under positive selection, (ii) lineage-specific genes and (iii) gene-duplications in the three lineages (figure 1). Based on the results of our comparative analyses we provide a list of candidate genes that might play a meaningful role in the ecological adaptation of these fruit flies.

RESULTS

Features of the *D. buzzatii* genome

Genome sequencing and assembly

We sequenced and *de novo* assembled the genome of *D. buzzatii* line st-1 using shotgun and paired-end reads from 454/Roche, mate-pair and paired-end reads from Illumina, and Sanger BAC-end sequences (~22x total expected coverage; see Methods for details). We consider the resulting assembly (Freeze 1) as the reference *D. buzzatii* genome sequence (table 1). This assembly comprises 826 scaffolds >3 kb long with a total size of 161.5 Mb. Scaffold N50 and N90 indexes are 30 and 158, respectively, whereas scaffold N50 and N90 lengths are 1.38 and 0.16 Mb, respectively (table 1). Quality controls (see Methods) yielded a relatively low error rate of ~ 0.0005 (PHRED quality score Q = 33). For comparison, we also assembled the genome of the same line (st-1) using only four lanes of short (100 bp) Illumina paired-end reads (~76x expected coverage) and the SOAPdenovo software (Luo et al. 2012). This resulted in 10,949 scaffolds >3 kb long with a total size of 144.2 Mb (table 1). All scaffolds are available for download from the *Drosophila buzzatii* Genome Project web page (http://dbuz.uab.cat). This site also displays all the information generated in this project (see below).

Genome size and repeat content

The genome sizes of two *D. buzzatii* strains, st-1 and j-19, were estimated by Feulgen Image Analysis Densitometry on testis cells (Ruiz-Ruano et al. 2011) using *D. mojavensis* as reference. Integrative Optical Density (IOD) values were 21% (st-1)

and 25% (j-19) smaller than those for *D. mojavensis*. Thus, taking 194 Mb (total assembly size) as the genome size of *D. mojavensis* (Drosophila 12 Genomes Consortium et al. 2007) we estimated the genome sizes for *D. buzzatii* st-1 and j-19 lines as 153 and 146 Mb, respectively.

To assess the transposable element (TE) content of the *D. buzzatii* genome we masked the 826 scaffolds of Freeze 1 assembly using a library of TEs compiled from several sources (see Methods). We detected a total of 56,901 TE copies covering ~8.4 % of the genome (table 2). The most abundant TEs seem to be Helitrons, LINEs, LTR retrotransposons and TIR transposons that cover 3.4 %, 1.6 %, 1.5 % and 1.2 % of the genome, respectively (table 2). In addition, we identified tandemly repeated satellite DNAs (satDNA) with repeat units longer than 50 bp (Melters et al. 2013) (see Methods). The two most abundant tandem repeat families are the pBuM189 satellite (Kuhn et al. 2008) and the DbuTR198 satellite, a novel family with repeat units 198 bp long (table 3). The remaining tandem repeats had sequence similarity to integral parts of TEs, such as the internal tandem repeats of the transposon Galileo (de Lima et al. in preparation).

Chromosomal rearrangements

The basic karyotype of *D. buzzatii* is similar to that of the Drosophila genus ancestor and consists of six chromosome pairs: four pairs of equal-length acrocentric autosomes, one pair of "dot" autosomes, a long acrocentric X and a small acrocentric Y (Ruiz and Wasserman 1993). Because no interchromosomal reorganizations between *D. buzzatii* and *D. mojavensis* have previously been found (Ruiz et al. 1990;

Ruiz and Wasserman 1993) all 826 scaffolds were assigned to chromosomes by blastn against the *D. mojavensis* genome. In addition, the 158 scaffolds in the N90 index were mapped to chromosomes, ordered and oriented (supplementary figure S1, Supplementary Material online; Delprat et al. in preparation) using conserved linkage (Schaeffer et al. 2008) and additional information (González et al. 2005; Guillén and Ruiz 2012). A bioinformatic comparison of *D. buzzatii* and *D. mojavensis* chromosomes confirmed that chromosome 2 differs between these species by ten inversions (*2m, 2n, 2z*^{7}, *2c, 2f, 2g, 2h, 2q, 2r*, and *2s*), chromosomes X and 5 differ by one inversion each (*Xe* and *5g*, respectively) and chromosome 4 is homosequential as previously described (Ruiz et al. 1990; Ruiz and Wasserman 1993; Guillén and Ruiz 2012). In constrast, we find that chromosome 3 differs by five inversions instead of the expected two that were previously identified by cytological analyses (Ruiz et al. 1990). These three additional chromosome 3 inversions seem to be specific to the *D. mojavensis* lineage (Delprat et al. in preparation). One of these inversions, *3f*^{2}, is polymorphic in natural populations of *D. mojavensis*, but, conflicting with previous reports (Ruiz et al. 1990; Schaeffer et al. 2008), appears to be homozygous in the sequenced strain. This has been corroborated by the cytological reanalysis of its polytene chromosomes (Delprat et al. 2014).

Many developmental genes are arranged in gene complexes each comprising a small number of functionally related genes. We checked the organization of six of these gene complexes in the *D. buzzatii* genome: *Hox* gene complex (*HOM-C*), *Achaete-scute* complex (*AS-C*), *Iroquois* complex (*IRO-C*), *NK* homeobox gene cluster (*NK-C*), *Enhancer of split* complex (*E(spl)-C*) and *Bearded* complex (*Brd-C*) (Negre et al. in preparation). *Hox* genes were arranged in a single

complex in the Drosophila genus ancestor (Hughes and Kaufman 2002). However, this *HOM-C* suffered two splits (caused by chromosomal inversions) in the lineage leading to the repleta species group (Negre et al. 2005). In order to fully characterize *HOM-C* organization in *D. buzzatii*, we manually annotated all *Hox* genes and located them in three scaffolds (2, 5 and 229) of chromosome 2 (Negre et al. in preparation). The analysis of these scaffolds revealed that only two clusters of *Hox* genes are present. The distal cluster contains *pb*, *Dfd*, *Sex combs reduced* (*Scr*), *Antennapedia* (*Antp*) and *Ultrabithorax* (*Ubx*) whereas the proximal cluster contains *lab*, *abdA* and *Abdominal B* (*AbdB*). This is precisely the same *HOM-C* organization observed in *D. mojavensis* (Negre and Ruiz 2007). Therefore there seem to be no additional rearrangements of the HOM-C in *D. buzzatii* besides those already described in the genus Drosophila (Negre and Ruiz 2007). The other five developmental gene complexes contain four, three, six, 13 and six functionally related genes, respectively (Lai et al. 2000; Garcia-Fernàndez 2005; Irimia et al. 2008; Negre and Simpson 2009). All these complexes seem largely conserved in the *D. buzzatii* genome with few exceptions (Negre et al. in preparation). The gene *slouch* is separated from the rest of the NK-C in *D. buzzatii* and also in all other Drosophila species outside of the melanogaster species group; in addition, the gene *Bearded*, a member of the Brd-C, is seemingly absent from the *D. buzzatii* and *D. mojavensis* genomes, although it is present in *D. virilis* and *D. grimshawi*. On the other hand, genes flanking the complexes are often variable, presumably due to the fixation of chromosomal inversions with breakpoints in the boundaries of the complexes.

Protein-coding gene content

We used a combination of *ab initio* and similarity-based algorithms in order to reduce the high false-positive rate associated with *de novo* gene prediction (Wang et al. 2003; Misawa and Kikuno 2010) as well as to avoid the propagation of false-positive predicted gene models when closely related species are used as references (Poptsova and Gogarten 2010). A total of 13,657 protein-coding genes (PCGs) were annotated in the *D. buzzatii* genome (Annotation Release 1). These PCG models contain a total of 52,250 exons with an average of 3.8 exons per gene. Gene expression analyses provided transcriptional evidence for 88.4% of these gene models (see below). The number of PCGs annotated in *D. buzzatii* is lower than the number annotated in *D. mojavensis* (14,595, Release 1.3), but quite close to the number annotated in *D. melanogaster* (13,955, Release 5.56), one of the best-known eukaryotic genomes (St Pierre et al. 2014). However PCGs in both *D. buzzatii* and *D. mojavensis* genomes tend to be smaller and contain fewer exons than those in the *D. melanogaster* genome (supplementary table S1, Supplementary Material online), which suggests that the annotation in the two cactophilic species might be incomplete. After applying several quality filters, a total of 12,977 high confidence protein-coding sequences (CDS) were selected for further analysis (see Methods).

Developmental transcriptome

To characterize the expression profile throughout *D. buzzatii* development we performed RNA-Seq experiments using samples from five different stages: embryos, larvae, pupae, adult females and adult males. Gene expression levels were calculated based on FPKM values. PCG models that did not show evidence of transcription (FPKM < 1) were classified as non-expressed PCGs whereas

transcribed regions that did not overlap with any annotated PCG model were tentatively considered non-coding RNA (ncRNA) genes (figure 2a). We detected expression (FPKM > 1) of 26,455 transcripts and 15,026 genes, 12,066 (80%) are PCGs and 2,960 (20%) are ncRNA genes. The number of expressed genes (PCGs + ncRNA) increases through the life cycle with a maximum of 12,171 in adult males (figure 2a and supplementary table S2, Supplementary Material online), a pattern similar to that found in *D. melanogaster* (Graveley et al. 2011). In addition, we observed a clear sex-biased expression in adults: males express 1,824 more genes than females. Previous studies have attributed this sex-biased gene expression mainly to the germ cells, indicating that the differences between ovary and testis are comparable to those between germ and somatic cells (Parisi et al. 2004; Graveley et al. 2011).

We assessed expression breadth for each gene simply as the number of developmental stages with evidence of expression (figure 2b and supplementary table S2, Supplementary Material online). Expression breadth is significantly different ($P<0.001$) for PCGs and ncRNA genes. A total of 6,546 expressed PCGs (54.2%) are constitutively expressed (i.e. we observed expression in the five stages), but only 260 of ncRNA genes (8.8%) are constitutively expressed (supplementary table S2, Supplementary Material online). In contrast, 925 expressed PCGs (7.7%) and 1,292 ncRNA genes (43.6%) are expressed only in one stage. Mean expression breadth was 3.9 for PCGs and 2.2 for ncRNA genes. Adult males show more stage-specific genes (844 genes) compared to adult females (137 genes).

PCGs with no expression in this study (FPKM < 1) might be expressed at a higher level in other tissues or times, or they might be inducible under specific

conditions that we did not test (Weake and Workman 2010; Etges et al. 2014; Matzkin 2014). We also must expect that some remaining fraction of gene models will be false positives (Wang et al. 2003). However, because we used a combination of different annotation methods to reduce the proportion of false-positives, we expect this proportion to be very small. On the other hand, transcribed regions that do not overlap with any annotated PCG models, are likely ncRNA genes although we cannot discard that some of them might be false negatives, i.e. genes that went undetected by our annotation methods perhaps because they contain small open reading frames (Ladoukakis et al. 2011). One observation supporting that most of them are in fact ncRNA genes is that their expression breadth is quite different from that of PCGs and a high fraction of them are stage-specific genes. In most Drosophila species, with limited analyses of the transcriptome (Celniker et al. 2009), few ncRNA genes have been annotated. By contrast, in *D. melanogaster* with a very well annotated genome, 2,096 ncRNA genes have been found (Release 5.56, FlyBase). Thus, the number of ncRNA found in *D. buzzatii* is comparable to that of *D. melanogaster*.

Website

A website (http://dbuz.uab.cat) has been created to provide free access to all information and resources generated in this work. It includes a customized browser (GBrowse; Stein et al. 2002) for the *D. buzzatii* genome incorporating multiple tracks for gene annotations with different gene predictors, for expression levels and transcript annotations for each developmental stage, and for repeat annotations. It

contains also utilities to download contigs, scaffolds and data files and to carry out Blast searches against all *D. buzzatii* contigs and scaffolds.

Lineage-specific analyses

We set up to analyze three lineages for several aspects that could reveal genes involved in adaptation to the cactophilic niche. These lineages are denoted as #1, #2 and #3, respectively, in figure 1: *D. buzzatii* lineage, *D. mojavensis* lineage and cactophilic lineage (i.e. lineage shared by *D. buzzatii* and *D. mojavensis*) . We searched for genes under positive selection, duplicated genes and orphan genes in those lineages.

Genes under positive selection

We first searched for genes evolving under positive selection during the divergence between *D. buzzatii* and *D. mojavensis,* using codon substitution models implemented in the PAML 4 package (Yang 2007). Two pairs of different site models (SM) were compared by the likelihood ratio test (LRT), M1a vs. M2a and M7 vs. M8 (see Methods). In each case, a model that allows for sites with $\omega \geq 1$ (positive selection) is compared with a null model that considers only sites with $\omega < 1$ (purifying selection) and $\omega = 1$ (neutrality). At P<0.001, the first comparison (M1a vs. M2a) detected 915 genes while the second comparison (M7 vs. M8) detected 802 genes. Comparison of the two gene sets allowed us to detect 772 genes present in both,

and this was taken as the final list of genes putatively under positive selection using SM (supplementary table S3, Supplementary Material online).

Next, we used branch-site models (BSM) from PAML 4 package (Yang 2007) to search for genes under positive selection in the phylogeny of the four Drosophila subgenus species, *D. buzzatii*, *D. mojavensis*, *D. virilis* and *D. grimshawi* (figure 1). Orthologous relationships among the four species were inferred from *D. buzzatii*-*D. mojavensis* list of orthologs and the OrthoDB catalog (see Methods). A total of 8,328 unequivocal 1:1:1:1 orthologs were included in the comparison of a branch-site model allowing sites with $\omega > 1$ (positive selection) and a null model that does not. We selected three branches to test for positive selection (the foreground branches): *D. buzzatii* lineage, *D. mojavensis* lineage and cactophilic lineage (denoted as #1, #2 and #3 in figure 1). The number of genes putatively under positive selection detected at P<0.001 in the three branches was 350, 172 and 458, respectively (supplementary table S3, Supplementary Material online). These genes only partially overlap those previously detected in the *D. buzzatii*-*D. mojavensis* comparison using SM (figure 3). While 69.4% and 55.8% of the genes putatively under positive selection in the *D. buzzatii* and *D. mojavensis* lineages were also detected in the *D. buzzatii*-*D. mojavensis* comparison, only 22.3% of the genes detected in the cactophilic lineage were present in the previous list (figure 3). Thus the total number of genes putatively under positive selection is 1,294.

We looked for functional categories overrepresented among the candidates genes reported by both site and branch-site models (table 4). We first performed a Gene Ontology (GO) enrichment analysis with the 772 candidate genes uncovered by site models comparing *D. mojavensis* and *D. buzzatii* orthologs using DAVID tools

(Huang et al. 2007). Two molecular functions show higher proportion than expected by chance (relative to *D. mojavensis* genome) within the list of candidate genes: antiporter activity and transcription factor activity. With respect to the biological process, regulation of transcription is the only overrepresented category. A significant enrichment in Src Homology-3 domain was observed. This domain is commonly found within proteins with enzymatic activity and it is associated with protein binding function.

A similar GO enrichment analysis was carried out with candidate genes found using branch-site models in each of the three targeted branches. The 350 candidate genes in *D. buzzatii* lineage show a significant enrichment in DNA-binding function. DNA-dependent regulation of transcription and phosphate metabolic processes were also overrepresented. We also found a significant enrichment in the Ig-like domain, involved in functions related to cell-cell recognition and immune system. The 172 candidate genes in *D. mojavensis* lineage show a significant excess of genes related to the heterocycle catabolic process (P = 5.9e-04). Interestingly, the main hosts of *D. mojavensis* (columnar cacti), contain large quantities of triterpene glycosides, which are heterocyclic compounds. Among the candidate genes in the branch leading to the two cactophilic species, there are three overrepresented molecular functions related to both metal and DNA binding. The GO terms with the highest significance in the biological process category are cytoskeleton organization and, once again, regulation of transcription.

Using the RNA-Seq data we determined the expression profiles of all 1,294 genes putatively under positive selection. A total of 1,213 (93.7%) of these genes are expressed in at least one developmental stage (supplementary table S2,

Supplementary Material online). A comparison of expression level and breadth between candidate and non-candidate genes revealed that genes putatively under positive selection are expressed at a lower level (X^2 = 84.96, P < 2e-16) and in fewer developmental stages (X^2 = 26.99, P < 2e-6) than the rest.

Orphan genes in the cactophilic lineage

To detect orphan genes in the cactophilic lineage we blasted the amino acid sequences encoded by 9,114 *D. buzzatii* genes with *D. mojavensis* 1:1 orthologs against all proteins from the 12 Drosophila genomes except *D. mojavensis* available in FlyBase (St Pierre et al. 2014). We found 117 proteins with no similarity to any predicted Drosophila protein (cutoff value of 1e-05) and were considered to be encoded by putative orphan genes. We focused on the evolutionary dynamics of these orphan genes by studying their properties in comparison to the remaining 8,997 1:1 orthologs (figure 4). We observed that median dn of orphan genes was significantly higher than that of non-orphan genes (dn_{orphan} = 0.1291; $dn_{non\text{-}orphan}$ = 0.0341; W=846254, P < 2.2e-16) and the same pattern was observed for ω (ω_{orphan} = 0.4253, $\omega_{non\text{-}orphan}$ = 0.0887, W=951117, P < 2.2e-16). However median ds of orphan genes is somewhat lower than that for the rest of genes (ds_{orphan}=0.3000, $ds_{non\text{-}orphan}$ = 0.4056, W=406799, P=2.4e-05).

We found 19 out of the 117 orphan genes in the list of candidate genes detected in the *D. buzzatii-D. mojavensis* comparison (see above). This proportion (16.3%) was significantly higher than that found in non-orphan 1:1 orthologs (753/8997 = 8.4%), which indicates an association between gene lineage-specificity

and positive selection (Fisher exact test, two tailed, P < 0.0001). The 19 orphan genes included in the candidate gene group are not associated with any GO category. As a matter of fact, information about protein domains was found for only two of these genes (GYR and YLP motifs in both cases: GI20994 and GI20995). These results should be viewed cautiously as newer genes are functionally undercharacterized and GO databases are biased against them (Zhang et al. 2013). We also compared the protein length between orphan and non-orphan gene products. Our results showed that orphan genes are shorter (W=68825.5, P<2.2e-16) and have fewer exons than non-lineage-specific genes (W=201068, P<2.2e-16).

RNA-Seq data allowed us to test for expression of orphan genes. From the 117 gene candidates, 82 (70%) are expressed at least in one of the five analyzed developmental stages. A comparison of the expression profiles between orphan and the rest of 1:1 orthologous genes showed that the expression breadth of orphans is different from that of non-orphans (X^2=101.4, P< 0.001): most orphan genes are expressed exclusively in one developmental stage with mean expression breadth of 2.56 (versus 3.94 for non-orphans).

Gene duplications

The annotated PCGs from four species of the Drosophila subgenus were used to study gene family expansions in the *D. buzzatii*, *D. moja*vensis and cactophilic lineages (figure 1). Proteins that share 50% identity over 50% of their length were clustered into gene families using Markov Cluster Algorithm. After additional quality filters (see Methods), the final dataset consisted of a total of 56,587

proteins from four species clustered into 19,567 families, including single-gene families (supplementary tables S4 to S7, Supplementary Material online).

Considering the *D. buzzatii* genome alone (supplementary table S4, Supplementary Material online), we find 11,251 single-copy genes and 1,851 duplicate genes (14%) clustered in 691 gene families. Among *D. buzzatii* gene families, about 70% of families have 2 members and the largest family includes 16 members (supplementary table S4, Supplementary Material online). Among single copy genes, 1,786 genes are only present in the *D. buzzatii* lineage. This number decreases only to 1,624 when proteins are clustered into families with a less stringent cutoff of 35% identity and 50% coverage. Such lineage-specific single-copy genes have been found in all the 12 Drosophila genomes that have been analyzed, including *D. mojavensis* (Hahn et al. 2007), and although traditionally they have been viewed as annotation artifacts, many of these genes may be either *de novo* or fast-evolving genes (Reinhardt et al. 2013; Palmieri et al. 2014).

Lineage-specific expansions were identified by analyzing the gene count for each family from the four species using CAFE3.1 (see Methods). This analysis detected expansions of 86 families along the *D. buzzatii* lineage. However, 15 families increased in size as the result of extra copies added to the dataset after taking into account high sequence coverage. The expansions of these families cannot be confirmed with the current genome assembly. The remaining families were analyzed further in order to confirm *D. buzzatii*-specific duplications. To do that, we first selected gene families with members that have ds < 0.4 (median ds for *D. mojavensis*-*D. buzzatii* orthologs) and then manually examined syntenic regions in *D. mojavensis* genome. Although this approach might miss some true lineage-specific

expansions, it reduces the possibility of including old families into the expansion category that might have been misclassified as a result of incomplete gene annotation in the genomes under study or independent loss of family members in different lineages. Of the 30 gene families whose members had ds < 0.4, we confirmed the expansion of 20 families (supplementary table S8, Supplementary Material online). In 12 of the 20 families, new family members are found on the same scaffold in close proximity suggesting unequal crossing over or proximate segmental duplication as the mechanisms for duplicate formation. The remaining 8 families contain dispersed duplicates found in different scaffolds. Six of these families expanded through retroposition, the RNA-mediated duplication mechanism that allows insertion of reverse-transcribed mRNA nearly anywhere in the genome. In most cases, family expansions are due to addition of a new single copy in the *D. buzzatii* lineage (in 25 of total 35 families). Two families that expanded the most, with up to 5 (Family 95) and 9 (Family 126) new members, encode various peptidases involved in protein degradation. Other expanded families are associated with a broad range of functions, including structural proteins of insect cuticle and chorion, enzymes involved in carbohydrate and lipid metabolism, proteins that function in immune response and olfactory receptors. In addition, Family 128 encodes female reproductive peptidases (Kelleher and Markow 2009) and it appears that new family members have been acquired independently in *D. buzzatii* and *D. mojavensis* lineages (supplementary table S11, Supplementary Material online).

We find 6 families in *D. buzzatii* that expanded through retroposition in the 11 MY since the split between *D. buzzatii* and *D. mojavensis* (supplementary table S9, Supplementary Material online). This gives a rate of 0.55 retrogenes/MY, which is

consistent with previous estimates of functional retrogene formation in *Drosophila* of 0.5 retrogenes/MY (Bai et al. 2007). The expression of all but one retrogene is supported by RNA-Seq data, with no strong biases in expression between the sexes. Four retrogenes are duplicates of ribosomal proteins, and the parental genes from two of these families (*RpL37a* and *RpL30*) have been previously shown to generate retrogenes in other *Drosophila* lineages (Bai et al. 2007; Han and Hahn 2012). Frequent retroposition of ribosomal proteins could be explained by the high levels of transcription of ribosomal genes although other *Drosophila* lineages do not show a bias in favor of retroduplication of ribosomal proteins (Bai et al. 2007; Han and Hahn 2012). The remaining two retrogenes include the duplicate of Caf1, protein that is involved in histone modification, and the duplicate of VhaM9.7-b, a subunit of ATPase complex.

CAFE analysis identified 127 families that expanded along the *D. mojavensis* lineage. Of these families, 86 contain members with ds < 0.4. Further examination of syntenic regions confirmed expansion of only 17 families (supplementary table S8, Supplementary Material online). New members in two families (Families 1121 and 1330) are found in different scaffolds and originated through RNA-mediated duplications. These instances have been previously identified as *D. mojavensis*-specific retropositions (Han and Hahn 2012). Members of expanded families encode proteins that function in proteolysis, peptide and ion transport, aldehyde and carbohydrate metabolism, as well as sensory perception (supplementary table S11, Supplementary Material online). At least four of the 17 expanded families play a role in reproductive biology: proteases of Family 128 with three new members have been shown to encode female reproductive peptidases (Kelleher and Markow 2009), and

members of 3 additional families (Families 187, 277 and 1234) encode proteins that are found in *D. mojavensis* accessory gland proteome (Kelleher et al. 2009).

There are 20 gene families that expanded along the cactophilic branch, i.e., before the split between *D. buzzatii* and *D. mojavensis* (see Methods; supplementary table S10, Supplementary Material online). Most families (16 of 20) have expanded through tandem or nearby segmental duplication and are still found within the same scaffold. The remaining families with dispersed duplicates included one retrogene, the duplicate of T-cp1, identified previously in *D. mojavensis* lineage (Han and Hahn 2012). The extent of per-family expansions in the cactophilic lineage is modest, with two new additional members found in four families and a single new copy in the remaining families. Members of the most expanded families encode guanylate cyclases that are involved in intracellular signal transduction, peptidases and carbon-nitrogen hydrolases. Members of other families include various proteins with metal-binding properties as well as proteins with a role in vesicle and transmembrane transport (supplementary table S11, Supplementary Material online). We also see expansion of three families (Family 775, Family 776 and Family 800) with functions related to regulation of juvenile hormone levels (see Discussion).

DISCUSSION

The *D. buzzatii* genome

Drosophila is a leading model for comparative genomics, with 24 genomes of different species already sequenced (Adams et al. 2000; Drosophila 12 Genomes Consortium et al. 2007; Zhou et al. 2012; Zhou and Bachtrog 2012; Fonseca et al.

2013; Ometto et al. 2013; Chen et al. 2014). However only five of these species belong to the species-rich Drosophila subgenus, and only one of these species, *D. mojavensis*, is a cactophilic species from the large repleta species group. Here we sequenced the genome and transcriptome of *D. buzzatii*, another cactophilic member of the repleta group, to investigate the genomic basis of adaptation to this distinct ecological niche. Using different sequencing platforms and a three-stage *de novo* assembly strategy, we generated a high quality genome sequence that consists of 826 scaffolds >3 kb (Freeze 1). A large portion (>90%) of the genome is represented by 158 scaffolds with a minimum size of 160 kb that have been assigned, ordered and oriented in the six chromosomes of the *D. buzzatii* karyotype. As expected, the assembly is best for chromosome 2 (because of the use of Sanger generated BAC-end sequences) and worst for chromosome X (because of the ¾ representation of this chromosome in adults of both sexes). The quality of our Freeze 1 assembly compares favorably with the assembly generated using only Illumina reads and the SOAPdenovo assembler, and with those of other Drosophila genomes generated using second-generation sequencing platforms (Zhou et al. 2012; Zhou and Bachtrog 2012; Fonseca et al. 2013; Ometto et al. 2013; Chen et al. 2014), although our Freeze 1 does not attain the quality of the 12 Drosophila genomes generated using Sanger only (Drosophila 12 Genomes Consortium et al. 2007).

D. buzzatii is a subcosmopolitan species that has been able to colonize four of the six major biogeographical regions (David and Tsacas 1980). Only two other repleta group species (*D. repleta* and *D. hydei*) have reached such widespread distribution. Invasive species are likely to share special genetic traits that enhance their colonizing ability (Parsons 1983; Lee 2002). From an ecological point of view we

would expect colonizing species to be r-strategists with a short developmental time (Lewontin 1965). Because there is a correlation between developmental time and genome size (Gregory and Johnston 2008), colonizing species are also expected to have a small genome size (Lavergne et al. 2010). The genome size of *D. buzzatii* was estimated in our assembly as 161 Mb and by cytological techniques as 153 Mb, ~20% smaller than the *D. mojavensis* genome. The genome size of a second *D. buzzatii* strain, estimated by cytological techniques, is even smaller, 146 Mb. However, the relationship between genome size and colonizing ability does not hold in the Drosophila genus at large. Although colonizing species such as *D. melanogaster* and *D. simulans* have relatively small genomes, specialist species with a narrow distribution such as *D. sechelia* and *D. erecta* also have small genomes. On the other hand, *D. ananassae, D. malerkotliana, D. suzuki, D. virilis*, and *Zaprionus indianus* are also colonizing Drosophila species but have relatively large genomes (Nardon et al. 2005; Bosco et al. 2007; Drosophila 12 Genomes Consortium et al. 2007; Gregory and Johnston 2008). Further, there seem to be little difference in genome size between original and colonized populations within species (Nardon et al. 2005). Seemingly, other factors such as historical or chance events, niche dispersion, genetic variability or behavioral shifts are more significant than genome size in determining the current distribution of colonizing species (Markow and O'Grady 2008).

TE content in the *D. buzzatii* genome was estimated as 8.4 % (table 2), a relatively low value compared with that of *D. mojavensis*, 10-14% (Ometto et al. 2013; Rius et al. in preparation). These data agree well with the smaller genome size of *D. buzzatii* because genome size is positively correlated with the contribution of

TEs (Kidwell 2002; Feschotte and Pritham 2007). However, TE copy number and coverage estimated in *D. buzzatii* (table 2) must be taken cautiously. Coverage is surely underestimated due to the difficulties in assembling repeats, in particular with short sequence reads, whereas the number of copies may be overestimated due to copy fragmentation (Rius et al. in preparation). The contribution of satDNAs (table 3) is also an underestimate and further experiments are required for a correct assessment of this component (de Lima et al. in preparation). However, we identified the pBuM189 satDNA as the most abundant tandem repeat of *D. buzzatii*. Previous *in situ* hybridization experiments revealed that pBuM189 copies are located in the centromeric region of all chromosomes, except chromosome X (Kuhn et al. 2008). Thus pBuM189 satellite is likely the main component of the *D. buzzatii* centromere. Interestingly, a pBuM189 homologous sequence has recently been identified as the most abundant tandem repeat of *D. mojavensis* (Melters et al. 2013). Although the chromosome location in *D. mojavensis* has not been determined, the persistence of pBuM189 as the major satellite DNA in *D. buzzatii* and *D. mojavensis* may reflect a possible role for these sequences in centromere function (Ugarković 2009).

Chromosome evolution

The chromosomal evolution of *D. buzzatii* and *D. mojavensis* has been previously studied by comparing the banding pattern of the salivary gland chromosomes (Ruiz et al. 1990; Ruiz and Wasserman 1993). *D. buzzatii* has few fixed inversions (*2m, 2n, 2z*[7], *5g*) when compared with the ancestor of the repleta group. In contrast, *D. mojavensis* showed ten fixed inversions (*Xe, 2c, 2f, 2g, 2h, 2q, 2r, 2s, 3a, 3d*), five of them (*Xe, 2q, 2r, 2s* and *3d*) exclusive to *D. mojavensis* and

the rest shared with other cactophilic Drosophila (Guillén and Ruiz 2012). Thus, the *D. mojavensis* lineage appears to be a derived lineage with a relatively high rate of rearrangement fixation. Here we compared the organization of both genomes corroborating all known inversions in chromosomes X, 2, 4 and 5. In *D. mojavensis* chromosome 3, however, we found five inversions instead of the two expected (Delprat et al. in preparation). One of the three additional inversions is the polymorphic inversions $3f^2$ (Ruiz et al. 1990). This inversion has previously been found segregating in Baja California and Sonora (Mexico) and is homozygous in the strain of Santa Catalina Island (California) that was used to generate the *D. mojavensis* genome sequence (Drosophila 12 Genomes Consortium et al. 2007). Previously, the Santa Catalina Island population was thought to have the standard (ancestral) arrangements in all chromosomes, like the populations in Southern California and Arizona (Ruiz et al. 1990; Etges et al. 1999). The presence of inversion $3f^2$ in Santa Catalina Island is remarkable because it indicates that the flies that colonized this island came from Baja California and are derived instead of ancestral with regard to the rest of *D. mojavensis* populations (Delprat et al. 2014). The other two additional chromosome 3 inversions are fixed in the *D. mojavensis* lineage and emphasize its rapid chromosomal evolution. Guillén and Ruiz (2012) analyzed the breakpoint of all chromosome 2 inversions fixed in *D. mojavensis* and concluded that the numerous gene alterations at the breakpoints with putative adaptive consequences point directly to natural selection as the cause of *D. mojavensis* rapid chromosomal evolution. The four fixed chromosome 3 inversions provide an opportunity for further testing this hypothesis (Delprat et al. in preparation).

Candidate genes under positive selection and orphan genes

Several methods have been developed to carry out genome-wide scans for genes evolving under positive selection (Nielsen 2005; Anisimova and Liberles 2007; Vitti et al. 2013). We used here a rather simple approach based on the comparison of the nonsynonymous substitution rate (dn) with the synonymous substitution rate (ds) at the codon level (Yang et al. 2000; Wong et al. 2004; Zhang et al. 2005; Yang 2007). Genes putatively under positive selection were detected on the basis of statistical evidence for a subset of codons where replacement mutations were fixed faster than mutation at silent sites. Four species of the Drosophila subgenus (figure 1) were employed to search for genes under positive selection using site models (SM) and branch-site models (BSM). We restricted the analysis to this subset of the Drosophila phylogeny to avoid the saturation of synonymous substitutions expected with phylogenetically very distant species (Bergman et al. 2002; Larracuente et al. 2008), and also because these are the genomes with the highest quality available (Schneider et al. 2009). A total of 1,294 candidate genes were detected with both SM and BSM, which represents ~14% of the total set of 1:1 orthologs between *D. mojavensis* and *D. buzzatii*. Positive selection seems pervasive in Drosophila (Sawyer et al. 2007; Singh et al. 2009; Sella et al. 2009; Mackay et al. 2012) and, using methods similar to ours, it has been estimated that 33% of single-copy orthologs in the *melanogaster* group have experienced positive selection (Drosophila 12 Genomes Consortium et al. 2007). The smaller fraction of genes putatively under positive selection in our analyses may be due to the fewer lineages considered in our study. In addition, both studies may be underestimating the true proportion of

positively selected genes because only 1:1 orthologs were included in the analyses and genes that evolve too fast may be missed by the methods used to establish orthology relationships (Bierne and Eyre-Walker 2004). At any rate, the 1,294 candidate genes found here should be evaluated using other genomic methods for detecting positive selection, e.g. those comparing levels of divergence and polymorphism (Vitti et al. 2013). Furthermore, functional follow-up tests will be necessary for a full validation of their adaptive significance (Lang et al. 2012).

Branch-site models allowed us to search for positively selected genes in the three-targeted lineages (*D. buzzatii*, *D. mojavensis* and cactophilic branch). We then performed GO enrichment analyses in order to identify potential candidates for environmental adaptation given the ecological properties of both cactophilic species (table 4). The most interesting result of this analysis is that genes putatively under positive selection in *D. mojavensis* branch are enriched in genes involved in heterocyclic catabolic processes. Four candidate *D. mojavensis* genes, *GI19101, GI20678, GI21543* and *GI22389,* that are orthologous to *D. melanogaster* genes *nahoda*, *CG5235*, *slgA* and *knk*, respectively, participate in these processes and might be involved in adaptation of *D. mojavensis* to the Stenocereus cacti, plants with particularly large quantities of heterocyclic compounds (see Introduction). A difficulty with this interpretation is the fact that the *D. mojavensis* genome sequence was generated using a strain from Santa Catalina Island where *D. mojavensis* inhabits Opuntia cactus (Drosophila 12 Genomes Consortium et al. 2007). However, the evidence indicates that the ancestral *D. mojavensis* population is the agria-inhabiting Baja California population and that the Mainland Sonora population split from Baja California ~0.25 MYA while the Mojave Desert and Mainland Sonora populations

diverged more recently, ~0.125 MYA (Smith et al. 2012). Moreover, the presence of inversion $3f^2$ in the Santa Catalina Island population suggests that the flies that colonized this island came from Baja California populations, where this inversion is currently segregating, and not from the Mojave Desert, where this inversion is not present (Delprat et al. 2014). This is compatible with mtDNA sequence data (Reed et al. 2007) although in contrast to other data (Machado et al. 2007). Finally, the transcriptional profiles of the four *D. mojavensis* subpopulations reveal only minor gene expression differences between individuals from Santa Catalina Island and Baja California (Matzkin and Markow 2013).

Orphan genes are genes with restricted taxonomic distribution. Such genes have been suggested to play an important role in phenotypic and adaptive evolution in multiple species (Domazet-Lošo and Tautz 2003; Khalturin et al. 2009; Chen et al. 2013). The detection of orphan genes is highly dependent on the availability of sequenced and well-annotated genomes of closely related species, and the total number of lineage-specific genes tend to be overestimated (Khalturin et al. 2009). We were as conservative as possible by considering only high-confidence 1:1 orthologs in two species, *D. buzzatii* and *D. mojavensis*. The result is a set of 117 orphans in the cactophilic lineage.

We observe that orphan genes clearly show a different pattern of molecular evolution compared to that of older conserved genes. Orphans exhibit a higher dn that can be attributed to more beneficial mutations fixed by positive selection or to lower constraint, or both (Cai and Petrov 2010; Chen et al. 2010). However, since the number of genes putatively under positive selection within the set of orphan genes is

higher than expected by chance, we suggest that the elevated dn likely reflects adaptive evolution.

Orphans also have fewer exons and encode shorter proteins than non-orphans. This observation has been reported in multiple eukaryotic organisms like yeasts (Carvunis et al. 2012), fruitflies (Domazet-Lošo and Tautz 2003) and primates (Cai and Petrov 2010), and it is further supported by a positive correlation between protein length and sequence conservation (Lipman et al. 2002) (see above). We did not find expression support for all the orphan genes detected. This suggests to us that either orphans are more tissue- or stage-specific than non-orphans (Zhang et al. 2012) or we are actually detecting artifactual CDSs that are not expressed. However, given the patterns of sequence evolution of orphan genes, we favor the first explanation for the majority of them. Collectively, all these results support the conclusion that orphan genes evolve faster than older genes, and that they experience lower levels of purifying selection and higher rates of adaptive evolution (Chen et al. 2010).

It has been widely reported that younger genes have lower expression levels than older genes on average (Cai and Petrov 2010; Tautz and Domazet-Lošo 2011; Zhang et al. 2012). Here we observe that orphan genes that are being transcribed are less expressed than non-orphans (Kruskal test, $X^2 = 9.37$, P=0.002). One of the proposed hypotheses to explain these observations is that genes that are more conserved are indeed involved in more functions (Pál et al. 2006; Tautz and Domazet-Lošo 2011).

Different studies have demonstrated that newer genes are more likely to have stage-specific expression than older genes (Zhang et al. 2012). Here we show

that the number of stage-specific expressed orphans is significantly higher than that of older genes. It has been proposed that newer genes tend to be more developmentally regulated than older genes (Tautz and Domazet-Lošo 2011). This means that they contribute most to the ontogenic differentiation between taxa (Chen et al. 2010). In *D. buzzatii* the vast majority of stage-specific orphan genes are expressed in larvae (15/29), indicating that expression of younger genes is mostly related to stages in which *D. buzzatii* and *D. mojavensis* lineages most diverge from each other.

Gene duplication

The study of gene duplications in the *D. buzzatii* and *D. mojavensis* lineages aims at understanding the genetic bases of the ecological specialization associated with colonization of novel cactus habitats. Although we only considered expanded families, it is known that specialization sometimes involves gene losses. For example, *D. sechellia* and *D. erecta,* which are specialized to grow on particular substrates, have lost gustatory receptors and detoxification genes (Drosophila 12 Genomes Consortium et al. 2007; Dworkin and Jones 2009). Sometimes the losses are driven by positive selection, as has been suggested in the case of the *neverland* gene in *D. pachea* (Lang et al. 2012) where positive selection appears to have favored a novel *neverland* allele that has lost the ability to metabolize cholesterol. In our study of gene families, the incompleteness of the annotation of *D. buzzatii* protein-coding genes precludes us from being able to reliably identify gene families that lost family members.

To minimize the possibility of missing gene copies that were potentially collapsed into single genes during *D. buzzatii* genome assembly, we used sequence coverage to adjust the size of gene families. Two of the families that expanded as a result of this correction encoded chorion genes. However, chorion genes are known to undergo somatic amplifications in ovarian follicle cells (Claycomb and Orr-Weaver 2005), and the use of sequence coverage to correct for "missing" copies can be misleading in these cases. As there is no easy way to verify families that were placed into the expanded category due to high sequence coverage alone, our discussion below is limited to gene duplicates that were annotated in the *D. buzzatii* genome.

A recent survey of the functional roles of new genes across various taxa offers evidence for the rapid recruitment of new genes into gene networks underlying a wide range of phenotypes including reproduction, behavior and development (Chen et al. 2013). A number of lineage-specific duplicates identified in our study fit this description, but further experimental confirmation of their functions through loss-of-function studies and characterization of molecular interactions are necessary. Among families that expanded in the *D. buzzatii*, the *D. mojavensis* and the cactophilic lineages, 35% have functional annotations that are similar to those of rapidly evolving families identified in the analysis of the 12 *Drosophila* genomes (Hahn et al. 2007). These families include genes that are involved in proteolysis, zinc ion binding, chitin binding, sensory perception, immunity and reproduction. A fraction of these expanded families may reflect physiological adaptations to a novel habitat. For example, given the importance of olfactory perception in recognition of the host cactus plants (Date et al. 2013), the duplication of an olfactory receptor in *D. buzzatii* may represent an adaptation to cactophilic substrates. Another *D. buzzatii* family includes *ninjurin*, a gene involved in tissue regeneration that is one of the

components of the innate immune response (Boutros et al. 2002). In *D. mojavensis*, we also observe the duplication of an odorant receptor and, coinciding with a previous report (Croset et al. 2010), of an ionotropic glutamate receptor that belongs to a novel family of diversified chemosensory receptors (Benton et al. 2009; Croset et al. 2010). An aldehyde dehydrogenase is also duplicated in the *D. mojavensis* lineage and might reveal a role in detoxification of particular aldehydes and ethanol (Fry and Saweikis 2006). In the *D. buzzatii*-*D. mojavensis* lineage, one family contains proteins with the MD-2 related lipid recognition domain involved in pathogen recognition and in *D. mojavensis* we find a duplicate of a phagosome-associated peptide transporter that is involved in bacterial response in *D. melanogaster* (Charrière et al. 2010).

Several of the *D. mojavensis*-specific gene duplicates have been described as male and female reproductive proteins. Unlike the accessory gland proteins of *D. melanogaster*, the proteome of *D. mojavensis* accessory glands is rich in metabolic enzymes and nutrient transport proteins (Kelleher et al. 2009). Three of the expanded families include metabolic proteins previously identified as candidate seminal fluid proteins specific to *D. mojavensis* lineage (Kelleher et al. 2009). We also detect an increase of female reproductive tract proteases as a possible counter adaptation to fast-evolving male ejaculate (Kelleher and Markow 2009).

Three gene families are of particular interest among those that were expanded in the lineages leading to *D. buzzatii* and *D. mojavensis*, as they contain duplicates of genes with functions related to the regulation of juvenile hormone (JH) levels. One family includes a new duplicate of juvenile hormone esterase duplication gene (*Jhedup* in *D. melanogaster*). Juvenile hormone esterases are involved in juvenile hormone degradation (Bloch et al. 2013), although *Jhedup* has much lower

level of JH esterase activity than *Jhe* (Crone et al. 2007). Another family includes new duplicate that encodes protein with sequence similarity to hemolymph juvenile hormone binding protein (*CG5945* in *D. melanogaster*). Juvenile hormone binding proteins belong to a large gene family regulated by circadian genes and affect circadian behavior, courtship behavior, metabolism and aging (Vanaphan et al. 2012). This family includes juvenile hormone binding proteins that function as carriers of juvenile hormone through the hemolymph to its target tissues (Bloch et al. 2013). The third family includes a new duplicate of a dopamine synthase gene (*ebony* in *D. melanogaster*). *ebony* is involved in the synthesis of dopamine, and it is known that dopamine levels affect behavior and circadian rhythms through regulation of hormone levels including JH (Rauschenbach et al. 2012). All three duplicates are expressed in *D. buzzatii* adults. At insect adult stage, JHs play a role in physiology and behavior, and their levels oscillate daily (Bloch et al. 2013). Gene duplications of JHBP, JHE and *ebony* may change the timing and levels of active JHs which, in turn, alter the behavior and physiology regulated by JHs. One interesting effect of mutations in circadian rhythm genes, or of direct perturbations of the circadian rhythm, is a reduced ethanol tolerance in *D. melanogaster* (Pohl et al. 2013). Intriguingly, *Jhedup* and another gene duplicated in the cactophilic lineage, *Sirt2* (a protein deacetylase), have been also shown to affect ethanol tolerance and sensitivity when mutated (Kong et al. 2010). Given that both *D. mojavensis* and *D. buzzatii* breed and feed on rotting fruit, a shift in tolerance to ethanol and other cactus-specific compounds is one of the expected adaptations associated with a switch to a cactus host. Future functional studies of these new duplicates are required to understand their role in physiological and behavioral changes associated with a change to a new habitat.

MATERIALS AND METHODS

We sequenced the genome of a highly inbreed *D. buzzatii* strain, st-1 (Betran et al. 1998). DNA was extracted from male and female adults (Piñol et al. 1988; Milligan 1998). Reads were generated with three different sequencing platforms (supplementary figure S2 and table S12, Supplementary Material online). The assembly of the genome was performed in three stages (supplementary table S13, Supplementary Material online): preassembly (Margulies et al. 2005), scaffolding (Boetzer et al. 2011) and gapfilling (Nadalin et al. 2012). In each step, a few chimeric scaffolds were identified and split. The final assembly, named Freeze 1, contains 826 scaffolds >3 kb and N50 and N90 index are 30 and 158, respectively. The distribution of read depth in the preassembly showed a Gaussian distribution with a prominent mode centered at ~22x (supplementary figure S3, Supplementary Material online). CG content is ~35% overall, ~42% in gene regions (including introns) and reaches ~52% in exons (supplementary table S14, Supplementary Material online). Unidentified nucleotides (N's) represent ~9% overall, ~4% in gene regions and 0.004% in exons. Sequence quality was assessed by comparing Freeze 1 with five Sanger sequenced BACs (Negre et al. 2005; Prada et al. 2010; Calvete et al. 2012) and with Illumina genomic and RNA-Seq reads (supplementary figure S4, Supplementary Material online). Quality assessments gave an overall error rate of ~0.0005 and a PHRED quality score of ~Q33 (supplementary tables S15 and S16, Supplementary Material online). An overall proportion of segregating sites of ~0.1% was estimated (supplementary table S17, Supplementary Material online).

The genome size of two *D. buzzatii* strains, st-1 and j-19, was estimated by Feulgen Image Analysis Densitometry. The genome size of *D. mojavensis* 15081-1352.22 strain (193,826,310 bp) was used as reference (Drosophila 12 Genomes Consortium et al. 2007).

Testicles from anesthetized males were dissected in saline solution and fixed in acetic-alcohol 3:1. Double preparations of *D. mojavensis* and *D. buzzatii* were prepared by crushing the fixed testicles in 50% acetic acid. Following Ruiz-Ruano et al. (2011), the samples were stained by Feulgen reaction and images obtained by optical microscopy were analyzed with the pyFIA software (supplementary figure S5 and table S18, Supplementary Material online).

The 826 scaffolds in Freeze 1 were assigned to chromosomes by aligning their sequences with the *D. mojavensis* genome using MUMmer (Delcher et al. 2003). In addition, the 158 scaffolds in the N90 index were mapped, ordered and oriented (supplementary figure S1, Supplementary Material online) using conserved linkage (Schaeffer et al. 2008), in situ hybridization and additional information (González et al. 2005; Guillén and Ruiz 2012). To estimate the number of rearrangements between *D. buzzatii* and *D. mojavensis*, their chromosomes were compared using GRIMM (Tesler 2002; Delprat et al. in preparation). Genes in the HOX gene complex (HOM-C) and five other gene complexes were searched *in silico* in the *D. buzzatii* genome and manually annotated using available information (Negre et al. 2005), the annotated *D. mojavensis* and *D. melanogaster* genomes, and the RNA-seq data generated for *D. buzzatii* (Negre et al. in preparation). TEs were annotated with RepeatMasker using a comprehensive TE library compiled from FlyBase (St Pierre et al. 2014), Repbase (Jurka et al. 2005) and RepeatModeler. Tandem Repeats Finder version 4.04 (Benson 1999) was used to identify satDNAs.

For the RNA-Seq experiments, RNA from frozen samples (embryos, larvae, pupae, adult males and adult females) was processed using the TruSeq RNA sample preparation kit provided by Illumina. We used a Hi-Seq2000 Illumina Sequencer to generate non-strand-specific paired-end ~100 bp reads from poly(A)+ RNA. Between 60 and 89 million reads were generated per sample. A total of ~286 million filtered reads were mapped to Freeze 1 with TopHat (Trapnell et al. 2009) representing ~180 x coverage of the total genome size

(supplementary table S19, Supplementary Material online). Transcripts were assembled with Cufflinks (Trapnell et al. 2010) using Annotation Release 1 as reference (see Methods).

Protein-coding genes were annotated combining with EVidence Modeler (EVM; Haas et al. 2008) the results of different predictors: Augustus (Stanke and Waack 2003), SNAP (Korf 2004), N-SCAN (Korf et al. 2001) and Exonerate (Slater and Birney 2005). The EVM set contained 12,102 gene models. We noticed that orthologs for a considerable number of *D. mojavensis* PCGs were absent from this data set. Thus, we used the Exonerate predictions to detect another 1,555 PCGs not reported by EVM (Poptsova and Gogarten 2010). Altogether, we predicted a total of 13,657 PCG models in the *D. buzzatii* reference genome (Annotation Release 1). Features of these models are given in supplementary table S20, Supplementary Material online. The RSD (Reciprocal Smallest Distance) algorithm (Wall and Deluca 2007) was used to identify 9,114 1:1 orthologs between *D. mojavensis* and *D. buzzatii*. Orthology relationships among the four species in the Drosophila subgenus (figure 1) were inferred from *D. buzzatii*-*D. mojavensis* list of orthologs and the OrthoDB catalog (version 6; Kriventseva et al. 2008). To test for positive selection we compared different codon substitution models using the likelihood ratio test. We run two pairs if site models (SM) on the orthologs set between *D. buzzatii* and *D. mojavensis*: M7 versus M8 and M1a versus M2a (Yang 2007). Then we used branch-site models (BSM) to test for positive selection in three lineages (figure 1): *D. mojavensis* lineage, *D. buzzatii* lineage, and the lineage that led to the two cactophilic species (*D. buzzatii* and *D. mojavensis*). We run Venny software (Oliveros 2007) to create a Venn diagram showing shared selected genes among the different models. We identified genes that are only present in the two cactophilic species, *D. mojavensis* and *D. buzzatii*, by blasting the amino acid sequences from the 9,114 1:1 orthologs between *D. mojavensis* and *D. buzzatii* (excluding missannotated genes) against all the proteins from the remaining 11 Drosophila species available in FlyBase protein database, excluding *D. mojavensis* (St Pierre et al. 2014).

For gene duplication analysis (DNA- and RNA-mediated duplications), we used annotated PCGs from the four species of the Drosophila subgenus (see Supplementary Methods). Briefly, we ran all-against-all blastp and selected hits with alignment length extending over at least 50% of both proteins and with amino acid identity of at least 50%. Markov Cluster Algorithm (Enright et al. 2002) was used to cluster retained proteins into gene families. The dataset was further modified to include additional family members based on sequence coverage and to exclude family members with internal stop codons and matches to transposable elements. Gene counts for each family from the 4 species were analyzed with an updated version of CAFE (CAFE 3.1 provided by the authors; Han et al. 2013) to identify lineage-specific expansions. The sets of CAFE-identified expanded families in the *D. buzzatii* and *D. mojavensis* genomes were examined for the presence of lineage-specific duplications. Families that included members with $d_S < 0.4$ were examined manually and lineage-specific duplications were inferred when no hits were found in the syntenic region of the genome with a missing copy. *D. buzzatii*-specific RNA-mediated duplications were identified by examining intron-less and intron-containing gene family members. A duplicate was considered a retrocopy if its sequence spanned all introns of the parental gene. The number of families identified by CAFE as expanded along the internal cactophilic branch was reduced by considering only those families that were also found in expanded category after rerunning the analysis with a less stringent cutoff (35% amino acid identity, 50% coverage). The overlapping set of expanded families was manually examined to verify the absence of *D. buzzatii* and *D. mojavensis* new family members in the *D. virilis* genome. Functional annotation (i.e., GO term) for all expanded families was obtained using the DAVID annotation tool (Huang et al. 2009a; Huang et al. 2009b). For genes without

functional annotation in DAVID, annotations of *D. melanogaster* orthologs were used. An extended version of these methods is given as a supplementary file, Supplementary Material online.

Acknowledgements

This work was supported by grants BFU2008-04988 and BFU2011-30476 from Ministerio de Ciencia e Innovación (Spain) to A.R., by a FPI fellowship to Y.G. and a PIF-UAB fellowship to N.R and by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM071813 to E.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. Heredity 99:567-79.

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8:R11.

Barker JSF, Starmer WT, MacIntyre RJ. 1990. Ecological and evolutionary genetics of *Drosophila*. Plenum Press.

Barker JSF, Starmer WT. 1982. Ecological genetics and evolution: the cactus-yeast-*Drosophila* model system. Sidney, Australia: Academic Press

Begon M. 1982. Yeasts and *Drosophila*. In: Ashburner M, Carson HL, Jr Thompson JN, editors. The Genetics and Biology of Drosophila. Vol. 3b. London: Academic Press. p. 3345–3384.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. Cell 136:149–162.

Bergman CM, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome Biol. 3:research0086.

Betran E, Santos M, Ruiz A. 1998. Antagonistic pleiotropic effect of second-chromosome inversions on body size and early life-history traits in *Drosophila buzzatii*. Evolution 52:144-154.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol. 21:1350–1360.

Bloch G, Hazan E, Rafaeli A. 2013. Circadian rhythms and endocrine functions in adult insects. J Insect Physiol. 59:56–69.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579.

Bosco G, Campbell P, Leiva-Neto JT, Markow TA. 2007. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177:1277–1290.

Boutros M, Agaisse H, Perrimon N. 2002. Sequential activation of signaling pathways during innate immune responses in *Drosophila*. Dev Cell 3:711–722.

Breitmeyer CM, Markow TA. 1998. Resource availability and population size in cactophilic *Drosophila*. Funct Ecol. 12:14–21.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2:393–409.

Calvete O, González J, Betrán E, Ruiz A. 2012. Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in *Drosophila*. Mol Biol Evol. 29:1875–1889.

Carson HL, Wasserman M. 1965. A widespread chromosomal polymorphism in a widespread species, *Drosophila buzzatii*. Am Nat. 99:111–115.

Carson HL. 1971. The ecology of *Drosophila* breeding sites. No. 2. University of Hawaii Foundation Lyon Arboretum Fund

Carvunis A-R, et al. 2012. Proto-genes and *de novo* gene birth. Nature 487:370–374.

Celniker SE, et al. 2009. Unlocking the secrets of the genome. Nature 459:927–930.

Charrière GM, et al. 2010. Identification of *Drosophila* Yin and PEPT2 as evolutionarily conserved phagosome-associated muramyl dipeptide transporters. J Biol Chem. 285:20147–20154.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. Science 330:1682-5.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nat Rev Genet. 14:645-60.

Chen ZX, et al. 2014. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. Genome Res. 24:1209-1223.

Claycomb JM, Orr-Weaver TL. 2005. Developmental gene amplification: insights into DNA replication and gene expression. Trends Genet. 21:149–162.

Crone EJ, et al. 2007. Only one esterase of *Drosophila melanogaster* is likely to degrade juvenile hormone in vivo. Insect Biochem Mol Biol. 37:540–549.

Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6:e1001064.

Date P, et al. 2013. Divergence in olfactory host plant preference in *D. mojavensis* in response to cactus host use. PloS One 8:e70027.

David J, Tsacas L. 1980. Cosmopolitan, subcosmopolitan and widespread species: different strategies within the Drosophilid family (Diptera). C R Soc Biogéogr 57:11–26.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinformatics. Chapter 10:Unit 10.3.

Delprat A, Etges WJ, Ruiz A. 2014. Reanalysis of polytene chromosomes in *Drosophila mojavensis* populations from Santa Catalina Island, California, USA. Drosophila Information Service 97 (in press).

Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 13:2213–2219.

Drosophila 12 Genomes Consortium, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Dworkin I, Jones CD. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics 181:721–736.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Etges W, et al. 2014. Deciphering life history transcriptomes in different environments. Mol Ecol. doi: 10.1111/mec.13017.

Etges WJ, Johnson WR, Duncan GA, Huckins G, Heed WB. 1999. Ecological Genetics of Cactophilic *Drosophila*. In: Ecology of Sonoran Desert plants and plant communities. University of Arizona Press. p. 164–214.

Fellows DP, Heed WB. 1972. Factors Affecting Host Plant Selection in Desert-Adapted Cactiphilic *Drosophila*. Ecology 53:850–858.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331–368.

Fogleman JC, Armstrong L. 1989. Ecological aspects of cactus triterpene glycosides I. Their effect on fitness components of *Drosophila mojavensis*. J Chem Ecol. 15:663–676.

Fogleman JC, Danielson PB. 2001. Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. Am Zool. 41:877–889.

Fogleman JC, Kircher HW. 1986. Differential effects of fatty acid chain length on the viability of two species of cactophilic *Drosophila*. Comp Biochem Physiol A Physiol. 83:761–764.

Fonseca NA, et al. 2013. *Drosophila americana* as a model species for comparative studies on the molecular basis of phenotypic variation. Genome Biol Evol. 5:661–679.

Fontdevila A, Ruiz A, Alonso G, Ocaña J. 1981. Evolutionary history of *Drosophila buzzatii*. I. Natural chromosomal polymorphism in colonized populations of the Old World. Evolution 35:148-157.

Frank MR, Fogleman JC. 1992. Involvement of cytochrome P450 in host-plant utilization by Sonoran Desert *Drosophila*. Proc Natl Acad Sci U S A. 89:11998–12002.

Fry JD, Saweikis M. 2006. Aldehyde dehydrogenase is essential for both adult and larval ethanol resistance in *Drosophila melanogaster*. Genet Res. 87:87–92.

Garcia-Fernàndez J. 2005. The genesis and evolution of homeobox gene clusters. Nat Rev Genet. 6:881–892.

González J, et al. 2005. A BAC-based physical map of the *Drosophila buzzatii* genome. Genome Res. 15:885–889.

Graveley BR, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473–479.

Gregory TR, Johnston JS. 2008. Genome size diversity in the family Drosophilidae. Heredity 101:228–238.

Guillén Y, Ruiz A. 2012. Gene alterations at *Drosophila* inversion breakpoints provide *prima facie* evidence for natural selection as an explanation for rapid chromosomal evolution. BMC Genomics 13:53.

Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 *Drosophila* genomes. PLoS Genet. 3:e197.

Han MV, Hahn MW. 2012. Inferring the history of interchromosomal gene transposition in *Drosophila* using n-dimensional parsimony. Genetics 190:813–825.

Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30:1987–1997.

Hasson E, Naveira H, Fontdevila A. 1992. The breeding sites of Argentinian cactophilic species of the *Drosophila mulleri* complex (subgenus Drosophila-repleta group). Rev Chil Hist Nat. 65:319–326.

Hasson E, et al. 1995. The evolutionary history of *Drosophila buzzatii*. XXVI. Macrogeographic patterns of inversion polymorphism in New World populations. J Evol Biol. 8:369–384.

Havens JA, Etges WJ. 2013. Premating isolation is determined by larval rearing substrates in cactophilic *Drosophila mojavensis*. IX. Host plant and population specific epicuticular hydrocarbon expression influences mate choice and sexual selection. J Evol Biol. 26:562–576.

Heed WB, Mangan RL. 1986. Community ecology of the Sonoran Desert *Drosophila*. In: The genetics and biology of Drosophila. Vol. 3e. M. Ashburner, H. L. Carson, J. N. Thompson. London: Academic Press.

Hoffmann AA, Sørensen JG, Loeschcke V. 2003. Adaptation of *Drosophila* to temperature extremes: bringing together quantitative and molecular approaches. J Therm Biol. 28:175–216.

Huang DW, Sherman BT, Lempicki RA. 2009a. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 4:44–57.

Huang DW, Sherman BT, Lempicki RA. 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37:1–13.

Huang DW, et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35:W169–W175.

Hughes CL, Kaufman TC. 2002. Hox genes and the evolution of the arthropod body plan. Evol Dev. 4:459–499.

Irimia M, Maeso I, Garcia-Fernàndez J. 2008. Convergent evolution of clustering of Iroquois homeobox genes across metazoans. Mol Biol Evol. 25:1521–1525.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Kelleher ES, Markow TA. 2009. Duplication, selection and gene conversion in a *Drosophila mojavensis* female reproductive protein family. Genetics 181:1451–1465.

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

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115:49–63.

Kircher HW. 1982. Chemical composition of cacti and its relationship to Sonoran Desert Drosophila. In: Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System. J.S.F. Barker and W. T. Starmer. Sydney, Australia: Academic Press. p. 143–158.

Kong EC, et al. 2010. Ethanol-regulated genes that contribute to ethanol sensitivity and rapid tolerance in Drosophila. Alcohol Clin Exp Res. 34:302–316.

Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics. 17 Suppl 1:S140–148.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. 36:D271–275.

Kuhn GCS, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. Chromosome Res 16:307–324.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. Genome Biol. 12:R118.

Lai EC, Bodner R, Posakony JW. 2000. The enhancer of split complex of Drosophila includes four Notch-regulated members of the bearded gene family. Development 127:3441–3455.

Lang M, et al. 2012. Mutations in the *neverland* gene turned *Drosophila pachea* into an obligate specialist species. Science 337:1658–1661.

Lang M, et al. 2014. Radiation of the *Drosophila nannoptera* species group in Mexico. J Evol Biol. 27:575–584.

Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.

Lavergne S, Muenke NJ, Molofsky J. 2010. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. Ann Bot. 105:109–116.

Lee CE. 2002. Evolutionary genetics of invasive species. Trends Ecol Evol. 17:386–391.

Lewontin RC. 1965. Selection for colonizing ability. In: Baker HG, Stebbins, editors. The genetics of colonizing species. New York: Academic Press.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. BMC Evol Biol. 2:20.

Loeschcke V, Krebs RA, Dahlgaard J, Michalak P. 1997. High-temperature stress and the evolution of thermal resistance in *Drosophila*. EXS 83:175–190.

Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1:18.

Machado CA, Matzkin LM, Reed LK, Markow TA. 2007. Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. Mol Ecol. 16:3009–3024.

Mackay TFC, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. Nature 482:173–178.

Manfrin MH, Sene FM. 2006. Cactophilic *Drosophila* in South America: a model for evolutionary studies. Genetica 126:57–75.

Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Markow TA, O'Grady P. 2008. Reproductive ecology of *Drosophila*. Funct Ecol. 22:747–759.

Matzkin LM, Markow TA. 2013. Transcriptional differentiation across the four subspecies of *Drosophila mojavensis*. In: Speciation: Natural Processes, Genetics and Biodiversity. New York: Nova Scientific Publishers.

Matzkin LM, Watts TD, Bitler BG, Machado CA, Markow TA. 2006. Functional genomics of cactus host shifts in *Drosophila mojavensis*. Mol Ecol. 15:4635–4643.

Matzkin LM. 2014. Ecological genomics of host shifts in *Drosophila mojavensis*. Adv Exp Med Biol. 781:233–247.

Melters DP, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol. 14:R10.

Milligan B. 1998. Total DNA isolation. In: Molecular Genetic Analysis of Population: A practical approach. 2nd ed. Oxford, NY, Tokyo: Oxford University Press. p. 29–64.

Misawa K, Kikuno RF. 2010. GeneWaltz–A new method for reducing the false positives of gene finding. BioData Min. 3:6.

Morales-Hojas R, Vieira J. 2012. Phylogenetic patterns of geographical and ecological diversification in the subgenus Drosophila. PloS One 7:e49552.

Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. BMC Bioinformatics 13 Suppl 14:S8.

Nardon C, et al. 2005. Is genome size influenced by colonization of new environments in dipteran species? Mol Ecol. 14:869–878.

Negre B, et al. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. Genome Res. 15:692–700.

Negre B, Ruiz A. 2007. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? Trends Genet. 23:55–59.

Negre B, Simpson P. 2009. Evolution of the *achaete-scute* complex in insects: convergent duplication of proneural genes. Trends Genet. 25:147–152.

Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet. 39:197-218.

Oliveira DCSG, et al. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. Mol Phylogenet Evol. 64:533–544.

Oliveros J. 2007. VENNY. An interactive tool for comparing lists with Venn diagrams. BioinfoGP CNB-CSIC.

Ometto L, Cestaro A, Ramasamy S, et al. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive *Drosophila* pest. Genome Biol Evol. 5:745–757.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. eLife 3:e01311.

Parisi M, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in Drosophila melanogaster adults. Genome Biol. 5:R40.

Parsons P. 1983. The Evolutionary Biology of Colonizing Species. New York: Cambridge University Press

Piñol J, Francino O, Fontdevila A, Cabré O. 1988. Rapid isolation of *Drosophila* high molecular weight DNA to obtain genomic libraries. Nucleic Acids Res. 16:2736.

Pohl JB, et al. 2013. Circadian genes differentially affect tolerance to ethanol in Drosophila. Alcohol. Clin. Exp. Res. 37:1862–1871.

Poptsova MS, Gogarten JP. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology 156:1909–1917.

Prada CF 2010. Evolución cromosómica del cluster *Drosophila martensis*: origen de las inversiones y reutilización de los puntos de rotura. PhD thesis. Universitat Autònoma de Barcelona.

Rajpurohit S, Oliveira CC, Etges WJ, Gibbs AG. 2013. Functional genomic and phenotypic responses to desiccation in natural populations of a desert drosophilid. Mol Ecol. 22:2698–2715.

Rauschenbach IY, Bogomolova EV, Karpova EK, Shumnaya LV, Gruntenko NE. 2012. The role of D1 like receptors in the regulation of juvenile hormone synthesis in *Drosophila* females with increased dopamine level. Dokl Biochem Biophys. 446:231–234.

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

Reinhardt JA, et al. 2013. *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet 9:e1003860.

Ruiz A, Cansian AM, Kuhn GC, Alves MA, Sene FM. 2000. The *Drosophila serido* speciation puzzle: putting new pieces together. Genetica 108:217–227.

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

Ruiz A, Heed WB. 1988. Host-plant specificity in the cactophilic *Drosophila* mulleri species complex. J Anim Ecol. 57:237–249.

Ruiz A, Wasserman M. 1993. Evolutionary cytogenetics of the *Drosophila buzzatii* species complex. Heredity 70:582–596.

Ruiz-Ruano FJ, et al. 2011. DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. Cytogenet. Genome Res. 134:120–126.

St Pierre SE, Ponting L, Stefancsik R, McQuilton P, FlyBase Consortium. 2014. FlyBase 102–advanced approaches to interrogating FlyBase. Nucleic Acids Res. 42(Database issue):D780-8.

Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proc Natl Acad Sci U S A. 104:6504–6510.

Schaeffer SW, et al. 2008. Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics 179:1601–1655.

Schneider A, et al. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 1:114–118.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 5:e1000495.

Singh ND, Larracuente AM, Sackton TB, Clark AG. 2009. Comparative genomics on the Drosophila phylogenetic tree. Annu Rev Ecol Evol Syst. 40:459–480.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

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

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2:ii215–225.

Starmer WT. 1981. A comparison of Drosophila habitats according to the physiological attributes of the associated yeast communities. Evolution 35:38-52.

Stein LD, et al. 2002. The generic genome browser: a building block for a model organism system database. Genome Res. 12:1599–1610.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet. 12:692–702.

Tesler G. 2002. GRIMM: genome rearrangements web server. Bioinformatics 18:492–493.

Throckmorton L. 1975. The phylogeny, ecology and geography of Drosophila. In: King R, editor. Handbook of Genetics. Vol. 3. New York: Plenum Press. p. 421–469.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

Trapnell C, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28:511–515.

Ugarković Đ. 2009. Centromere-competent DNA: structure and evolution. In: Ugarkovic D, editor. Centromere. Progress in Molecular and Subcellular Biology. Springer Berlin Heidelberg. p. 53–76.

Vanaphan N, Dauwalder B, Zufall RA. 2012. Diversification of takeout, a male-biased gene family in Drosophila. Gene 491:142–148.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data.Annu Rev Genet. 47:97-120.

Wall DP, Deluca T. 2007. Ortholog detection using the reciprocal smallest distance algorithm. Methods Mol Biol. 396:95–110.

Wang J, et al. 2003. Vertebrate gene predictions and the problem of large genes. Nat Rev Genet. 4:741–749.

Wasserman M. 1992. Cytological evolution of the *Drosophila repleta* species group. In: *Drosophila* inversion polymorphism. C.B Krimbas and J.R. Powell. Boca Raton, FL: CRC Press. p. 455–552.

Weake VM, Workman JL. 2010. Inducible gene expression: diverse regulatory mechanisms. Nat Rev Genet. 11:426–437.

Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041-51.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

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

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472-9.

Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. Bioessays 34:982-91.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. Science 337:341–345.

Zhou Q, et al. 2012. Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. BMC Genomics 13:109.

Table 1. Summary of assembly statistics for the genome of *Drosophila buzzatii.*

Assembly	**Freeze 1**	**SOAPdenovo**
Number of scaffolds (>3kb)	826	10949
Coverage	~22x	~76x
Assembly size (bp)	161490851	144184967
Scaffold N50 index	30	2035
Scaffold N50 length (bp)	1380942	18900
Scaffold N90 index	158	7509
Scaffold N90 length (bp)	161757	5703
Contig N50 index	1895	2820
Contig N50 length (bp)	17678	3101

Table 2. Transposable element content of *D. buzzatii* genome. The classification follows Wicker et al. (2007).

Class	Order	Annotated bp	Genome coverage (%)
I (retrotransposons)	LTR	2366439	1.47
	DIRS	55	0.00
	LINE	2541645	1.57
II (DNA transposons)	TIR	2017167	1.25
	Helitron	5531009	3.42
	Maverick	189267	0.12
	Unknown	973759	0.60
TOTAL		**13619341**	**8.43**

Table 3. Satellite DNAs identified in the *D. buzzatii* genome.

Tandem repeat family	Repeat length	GC content (%)	Genome coverage (%)[a]	Consensus Sequence[b]	Distribution
pBuM189	189	29	0.039	GCAAAAGACTCCGTCAATTAGAAAACAAAAA ATGTTATAGTTTTGAGGATTAACCGGCAAAAA CCGTATTATTTGTTATATGATTTCTGTATGGAA TACCGTTTTAGAAGCGTCTTTTATCGTATTACT CAGATATATCTTAAGATTTAGCATAATCTAAGA ACTTTTTGAAATATTCACATTTGTCCA	*D. buzzatii* cluster species *D. mojavensis*
DbuTR198	198	34	0.027	AAGGTAGAAAGGTAGTTGGTGAGATAAACCA GAAAAAGAGCTAAAAACGGCTAAAAACGGCT AGAAAATAGCCAGAAAGGTAGATTGAACATTA ATGGGCAAATGGATGGATAAATAAGACTGGT CATCATCCAATGAACAGAATCATGATTAAGAG ATAGAAATATGATTAGAAAGTAGGATAGAAAG GTTAGAAAG	*D. buzzatii*

[a]Genome fraction was calculated assuming a genome size of 163.547.398 bp (version 1 freeze of all contigs).

[b] Consensus sequence generated after clustering TRF results (see Materials and Methods).

Table 4. GO analysis of putative genes under positive selection detected by both site models (SM) and branch-site models (BSM). Only categories showing an enrichment with a p-value < 1.0e-03 are included.

Codon substitution Models	Lineage (branch number)	Number of candidates	GO enrichment					
			Molecular Function		Biological Process		Interpro domain	
			Id	Fold enrichment	Id	Fold enrichment	Id	Fold enrichment
Site Models (SM)	*D. buzzatii* vs *D. mojavensis*	772	Antiporter activity	1.77	Regulation of transcription	4.90	Src Homology-3 domain	1.60
			Transcription factor activity	1.56				
Branch Site Models (BSM)	*D. buzzatii* #1	350	DNA binding	1.36	Regulation of transcription DNA dependent	1.36	Immunoglobulin-like	1.33
					Phosphate Metabolic Process	0.72		
	D. mojavensis #2	172	Dopamine beta-monooxigenase activity	2.35	Heterocycle catabolic process	2.35	DOMON (DOpamine beta-MOnooxygenase N-terminal domain)	2.35
					Cation transport	0.98		
					Histidine family amino acid catabolic process	2.35		
	Cactophilic #3	458	Zinc ion binding	2.01	Cytoeskeleton organization	1.67	Zinc Finger, PHD-type	1.93
			Transition Metal Ion Binding	2.01	Regulation of transcription DNA dependent	1.06	Proteinase inhibitor I1 kazal	2.20
			DNA binding	1.66				

Figure legends

Figure 1. (a) Phylogenetic relationship of fruit fly species considered in our comparative analysis and their host preference. (b) Geographical distribution of cactophilic species *D. buzzatii* (red) and *D. mojavensis* (green) in America.

Figure 2. Developmental expression profile of *D. buzzatii* genes. (a) Number of expressed PCGs (red) and ncRNA genes (blue) along five developmental stages. (b) Classification of PCGs and ncRNA genes according to the number of stages where they are expressed.

Figure 3. Venn diagram showing the number of genes under positive selection detected by two different methods, site models (SM) and branch-site models (BSM) using three different lineages as foreground branches.

Figure 4. Patterns of divergence in orphan and non-orphan genes. Orphan genes (blue) have significantly higher dn and ω values compared to that of non-orphan genes (red). Non-orphan genes show significantly higher ds.

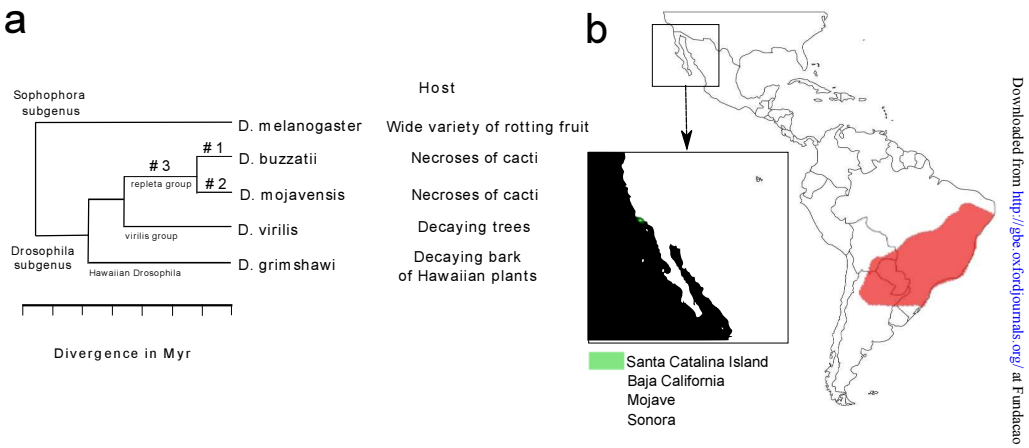

Figure 1

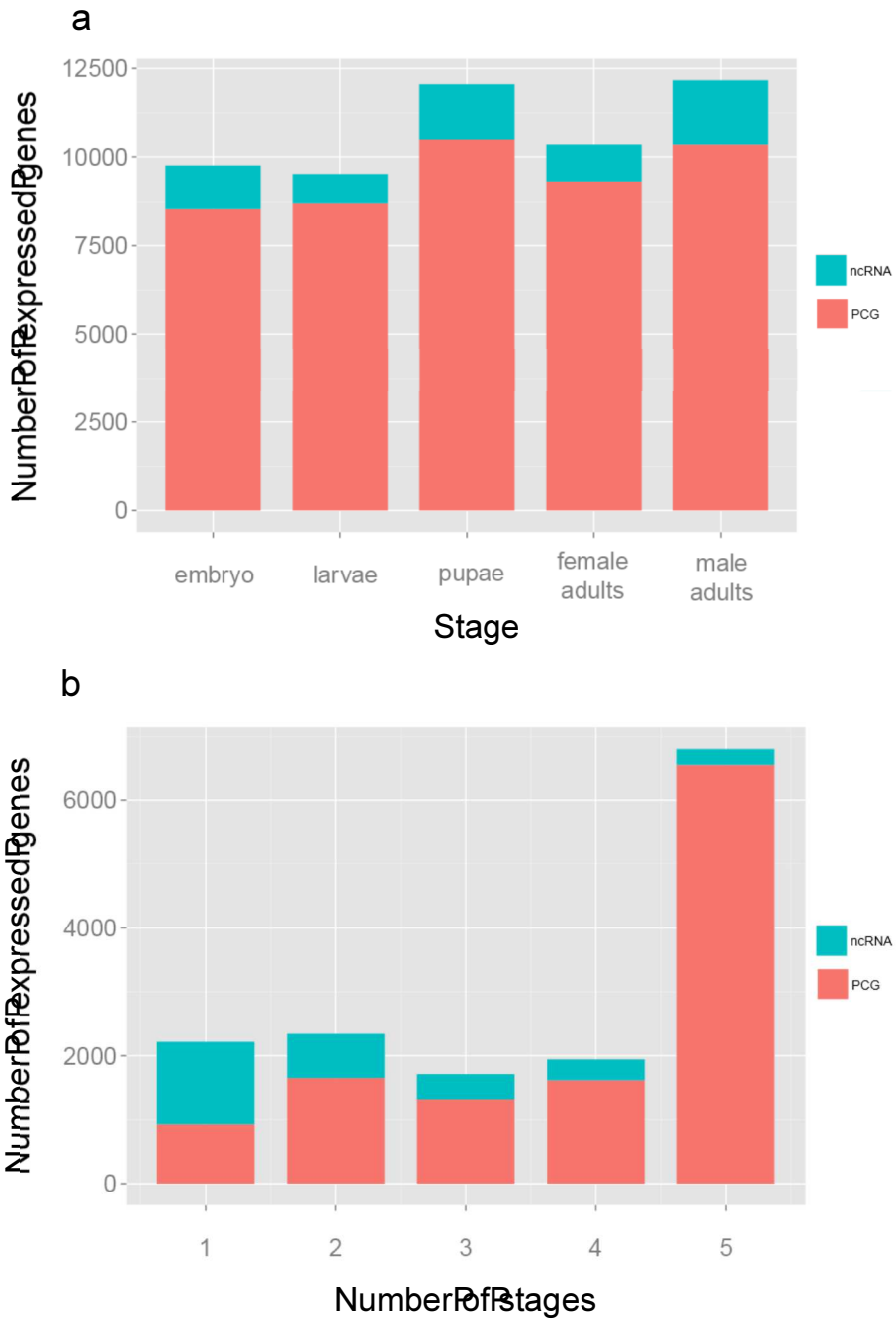

Figure 2

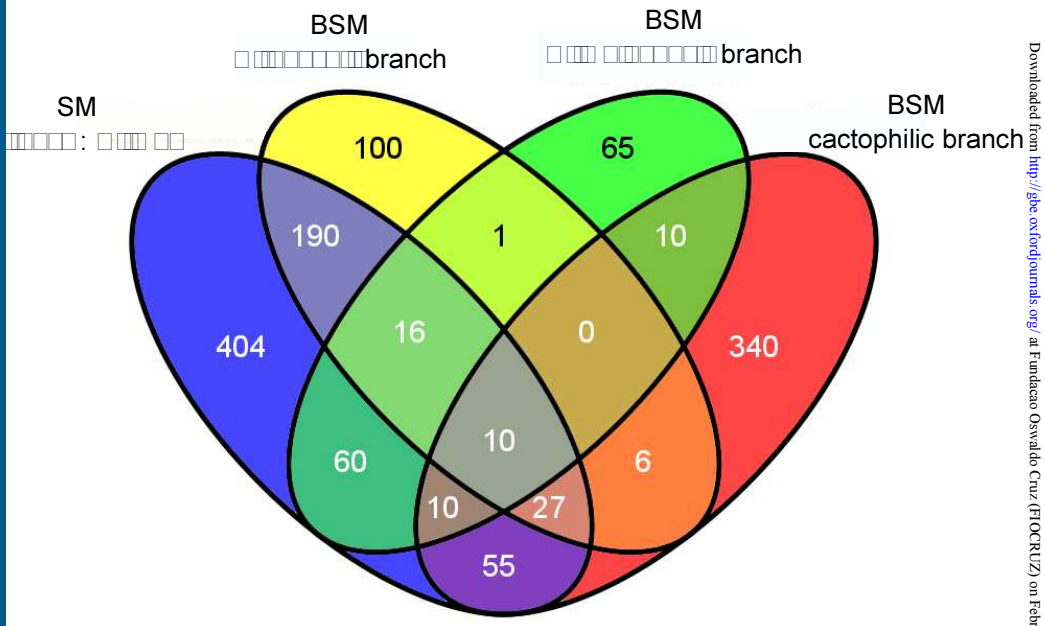

Figure 3

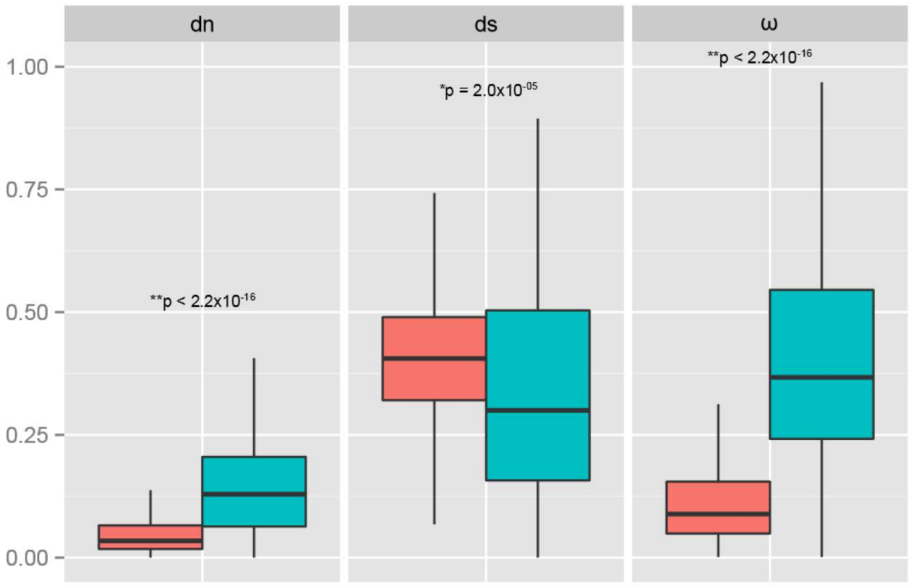

Figure 4