Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Comparative analysis of multiple genomes in a phylogenetic framework dramatically improves the precision and sensitivity of evolutionary inference, producing more robust results than single-genome analyses can provide. The genomes of 12 *Drosophila* species, ten of which are presented here for the first time (*sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis* and *grimshawi*), illustrate how rates and patterns of sequence divergence across taxa can illuminate evolutionary processes on a genomic scale. These genome sequences augment the formidable genetic tools that have made *Drosophila melanogaster* a pre-eminent model for animal genetics, and will further catalyse fundamental research on mechanisms of development, cell biology, genetics, disease, neurobiology, behaviour, physiology and evolution. Despite remarkable similarities among these *Drosophila* species, we identified many putatively non-neutral changes in protein-coding genes, non-coding RNA genes, and *cis*-regulatory regions. These may prove to underlie differences in the ecology and behaviour of these diverse species.

As one might expect from a genus with species living in deserts, in the tropics, on chains of volcanic islands and, often, commensally with humans, *Drosophila* species vary considerably in their morphology, ecology and behaviour[1]. Species in this genus span a wide range of global distributions: the 12 sequenced species originate from Africa, Asia, the Americas and the Pacific Islands, and also include cosmopolitan species that have colonized the planet (*D. melanogaster* and *D. simulans*) as well as closely related species that live on single islands (*D. sechellia*)[2]. A variety of behavioural strategies is also encompassed by the sequenced species, ranging in feeding habit from generalist, such as *D. ananassae*, to specialist, such as *D. sechellia*, which feeds on the fruit of a single plant species.

Despite this wealth of phenotypic diversity, *Drosophila* species share a distinctive body plan and life cycle. Although only *D. melanogaster* has been extensively characterized, it seems that the most important aspects of the cellular, molecular and developmental biology of these species are well conserved. Thus, in addition to providing an extensive resource for the study of the relationship between sequence and phenotypic diversity, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence. These genome sequences provide an unprecedented dataset to contrast genome structure, genome content, and evolutionary dynamics across the well-defined phylogeny of the sequenced species (Fig. 1).

Genome assembly, annotation and alignment

Genome sequencing and assembly. We used the previously published sequence and updated assemblies for two *Drosophila* species, *D. melanogaster*[3,4] (release 4) and *D. pseudoobscura*[5] (release 2), and generated DNA sequence data for 10 additional *Drosophila* genomes by whole-genome shotgun sequencing[6,7]. These species were chosen to span a wide variety of evolutionary distances, from closely related pairs such as *D. sechellia*/*D. simulans* and *D. persimilis*/*D. pseudoobscura* to the distantly related species of the *Drosophila* and *Sophophora* subgenera. Whereas the time to the most recent common ancestor of the sequenced species may seem small on an evolutionary timescale, the evolutionary divergence spanned by the genus *Drosophila* exceeds that of the entire mammalian radiation when generation time is taken into account, as discussed further in ref. 8. We sequenced seven of the new species (*D. yakuba*, *D. erecta*, *D. ananassae*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi*) to deep coverage (8.4× to 11.0×) to produce high quality draft sequences. We sequenced two species, *D. sechellia* and *D. persimilis*, to intermediate coverage (4.9× and 4.1×, respectively) under the assumption that the availability of a sister species sequenced to high coverage would obviate the need for deep sequencing without sacrificing draft genome quality. Finally, seven inbred strains of *D. simulans* were sequenced to low coverage (2.9× coverage from w^{501} and ~1× coverage of six other strains) to provide population variation data[9]. Further details of the sequencing strategy can be found in Table 1, Supplementary Table 1 and section 1 in Supplementary Information.

We generated an initial draft assembly for each species using one of three different whole-genome shotgun assembly programs (Table 1). For *D. ananassae*, *D. erecta*, *D. grimshawi*, *D. mojavensis*, *D. virilis* and *D. willistoni*, we also generated secondary assemblies; reconciliation of these with the primary assemblies resulted in a 7–30% decrease in the estimated number of misassembled regions and a 12–23% increase in the N50 contig size[10] (Supplementary Table 2). For *D. yakuba*, we generated 52,000 targeted reads across low-quality regions and gaps to improve the assembly. This doubled the mean contig and scaffold sizes and increased the total fraction of high quality bases (quality score (Q) > 40) from 96.5% to 98.5%. We improved the initial 2.9× *D. simulans* w^{501} whole-genome shotgun assembly by filling assembly gaps with contigs and unplaced reads from the ~1× assemblies of the six other *D. simulans* strains, generating a 'mosaic' assembly (Supplementary Table 3). This integration markedly improved the *D. simulans* assembly: the N50 contig size of the mosaic assembly, for instance, is more than twice that of the initial w^{501} assembly (17 kb versus 7 kb).

Finally, one advantage of sequencing genomes of multiple closely related species is that these evolutionary relationships can be exploited to dramatically improve assemblies. *D. yakuba* and *D. simulans* contigs and scaffolds were ordered and oriented using pairwise alignment to the well-validated *D. melanogaster* genome

*A list of participants and affiliations appears at the end of the paper.

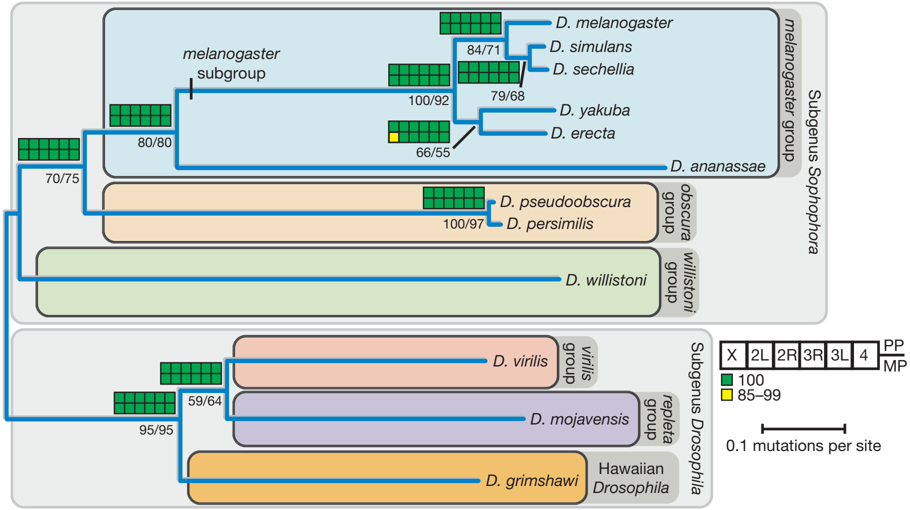


Figure 1 | Phylogram of the 12 sequenced species of *Drosophila*. Phylogram derived using pairwise genomic mutation distances and the neighbour-joining method[152,153]. Numbers below nodes indicate the per cent of genes supporting a given relationship, based on evolutionary distances estimated from fourfold-degenerate sites (left of solidus) and second codon positions (right of solidus). Coloured blocks indicate support from bayesian (posterior probability (PP), upper blocks) and maximum parsimony (MP; bootstrap values, lower blocks) analyses of data partitioned by chromosome arm. Branch lengths indicate the number of mutations per site (at fourfold-degenerate sites) using the ordinary least squares method. See ref. 154 for a discussion of the uncertainties in the *D. yakuba*/*D. erecta* clade.

sequence (Supplementary Information section 2). Likewise, the 4–5× *D. persimilis* and *D. sechellia* assemblies were improved by assisted assembly using the sister species (*D. pseudoobscura* and *D. simulans*, respectively) to validate both alignments between reads and linkage information. For the remaining species, comparative syntenic information, and in some cases linkage information, were also used to pinpoint locations of probable genome misassembly, to assign assembly scaffolds to chromosome arms and to infer their order and orientation along euchromatic chromosome arms, supplementing experimental analysis based on known markers (A. Bhutkar, S. Russo, S. Schaeffer, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Information section 2).

The mitochondrial (mt)DNA of *D. melanogaster*, *D. sechellia*, *D. simulans* (siII), *D. mauritiana* (maII) and *D. yakuba* have been previously sequenced[11,12]. For the remaining species (except *D. pseudoobscura*, the DNA from which was prepared from embryonic nuclei), we were able to assemble full mitochondrial genomes, excluding the A+T-rich control region (Supplementary Information section 2)[13]. In addition, the genome sequences of three *Wolbachia* endosymbionts (*Wolbachia wSim*, *Wolbachia wAna* and *Wolbachia wWil)* were assembled from trace archives, in *D. simulans*, *D. ananassae* and *D. willistoni*, respectively[14]. All of the genome sequences described here are available in FlyBase (www.flybase.org) and GenBank (www.ncbi.nlm.nih.gov) (Supplementary Tables 4 and 5).

Repeat and transposable element annotation. Repetitive DNA sequences such as transposable elements pose challenges for whole-genome shotgun assembly and annotation. Because the best approach to transposable element discovery and identification is still an active and unresolved research question, we used several repeat libraries and computational strategies to estimate the transposable element/repeat content of the 12 *Drosophila* genome assemblies (Supplementary Information section 3). Previously curated transposable element libraries in *D. melanogaster* provided the starting point for our analysis; to limit the effects of ascertainment bias, we also developed *de novo* repeat libraries using PILER-DF[15,16] and ReAS[17]. We used four transposable element/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) in conjunction with these transposable element libraries to identify repetitive elements in non-*melanogaster* species. We assessed the accuracy of each method by calibration with the estimated 5.5% transposable element content in the *D. melanogaster* genome, which is based on a high-resolution transposable element annotation[18] (Supplementary Fig. 1). On the basis of our results, we suggest a hybrid strategy for new genome sequences, employing translated BLAST with general transposable element libraries and RepeatMasker with species-specific ReAS libraries to estimate the upper and lower bound on transposable element content.

Protein-coding gene annotation. We annotated protein-coding sequences in the 11 non-*melanogaster* genomes, using four different *de novo* gene predictors (GeneID[19], SNAP[20], N-SCAN[21] and CONTRAST[22]); three homology-based predictors that transfer annotations from *D. melanogaster* (GeneWise[23], Exonerate[24], GeneMapper[25]); and one predictor that combined *de novo* and homology-based evidence (Gnomon[26]). These gene prediction sets

Table 1 | A summary of sequencing and assembly properties of each new genome

Final assembly	Genome centre	Q20 coverage (×)	Assembly size (Mb)	No. of contigs ≥2 kb	N50 contig ≥2 kb (kb)	Per cent of base pairs with quality >Q40
D. simulans	WUGSC*	2.9	137.8	10,843	17	90.3
D. sechellia	Broad†	4.9	166.6	9,713	43	90.6
D. yakuba	WUGSC*	9.1	165.7	6,344	125	98.5
D. erecta	Agencourt†	10.6	152.7	3,283	458	99.2
D. ananassae	Agencourt†	8.9	231.0	8,155	113	98.5
D. persimilis	Broad†	4.1	188.4	14,547	20	93.3
D. willistoni	JCVI‡	8.4	235.5	6,652	197	97.4
D. virilis	Agencourt†	8.0	206.0	5,327	136	98.7
D. mojavensis	Agencourt†	8.2	193.8	5,734	132	98.6
D. grimshawi	Agencourt†	7.9	200.5	9,632	114	97.1

Contigs, contiguous sequences not interrupted by gaps; N50, the largest length *L* such that 50% of all nucleotides are contained in contigs of size ≥*L*. The Q20 coverage of contigs is based on the number of assembled reads, average Q20 readlength and the assembled size excluding gaps. Assemblers used: *PCAP6, †ARACHNE4.5 and ‡Celera Assembler 7.

Table 2 | A summary of annotated features across all 12 genomes

	Protein-coding gene annotations			Non-coding RNA annotations				Repeat coverage (%)*	Genome size (Mb; assembly†/flow cytometry‡)
	Total no. of protein- coding genes (per cent with *D. melanogaster* homologue)	Coding sequence/ intron (Mb)	tRNA (pseudo)	snoRNA	miRNA	rRNA (5.8S + 5S)	snRNA		
D. melanogaster	13,733 (100%)	38.9/21.8	297 (4)	250	78	101	28	5.35	118/200
D. simulans	15,983 (80.0%)	45.8/19.6	268 (2)	246	70	72	32	2.73	111/162
D. sechellia	16,884 (81.2%)	47.9/21.9	312 (13)	242	78	133	30	3.67	115/171
D. yakuba	16,423 (82.5%)	50.8/22.9	380 (52)	255	80	55	37	12.04	127/190
D. erecta	15,324 (86.4%)	49.1/22.0	286 (2)	252	81	101	38	6.97	134/135
D. ananassae	15,276 (83.0%)	57.3/22.3	472 (165)	194	76	134	29	24.93	176/217
D. pseudoobscura	16,363 (78.2%)	49.7/24.0	295 (1)	203	73	55	31	2.76	127/193
D. persimilis	17,325 (72.6%)	54.0/21.9	306 (1)	199	75	80	31	8.47	138/193
D. willistoni	15,816 (78.8%)	65.4/23.5	484 (164)	216	77	76	37	15.57	187/222
D. virilis	14,680 (82.7%)	57.9/21.7	279 (2)	165	74	294	31	13.96	172/364
D. mojavensis	14,849 (80.8%)	57.8/21.9	267 (3)	139	71	74	30	8.92	161/130
D. grimshawi	15,270 (81.3%)	54.9/22.5	261 (1)	154	82	70	32	2.84	138/231

* Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). †Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. ‡Genome size estimates based on flow cytometry[38].

were combined using GLEAN, a gene model combiner that chooses the most probable combination of start, stop, donor and acceptor sites from the input predictions[27,28]. All analyses reported here, unless otherwise noted, relied on a reconciled consensus set of predicted gene models—the GLEAN-R set (Table 2, and Supplementary Information section 4.1).

Quality of gene models. As the first step in assessing the quality of the GLEAN-R gene models, we used expression data from microarray experiments on adult flies, with arrays custom-designed for *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*[29] (GEO series GSE6640; Supplementary Information section 4.2). We detected expression significantly above negative controls (false-discovery-rate-corrected Mann–Whitney U (MWU) $P < 0.001$) for 77–93% of assayed GLEAN-R models, representing 50–68% of the total GLEAN-R predictions in each species (Supplementary Table 6). Evolutionarily conserved gene models are much more likely to be expressed than lineage-specific ones (Fig. 2). Although these data cannot confirm the detailed structure of gene models, they do suggest that the majority of GLEAN-R models contain sequence that is part of a poly-adenylated transcript. Approximately 20% of transcription in *D. melanogaster* seems to be unassociated with protein-coding genes[30], and our microarray experiments fail to detect conditionally expressed genes. Thus, transcript abundance cannot conclusively establish the presence or absence of a protein-coding gene. Nonetheless, we believe these expression data increase our confidence in the reliability of the GLEAN-R models, particularly those supported by homology evidence (Fig. 2).

Because the GLEAN-R gene models were built using assemblies that were not repeat masked, it is likely that some proportion of gene models are false positives corresponding to coding sequences of transposable elements. We used RepeatMasker with *de novo* ReAS libraries and PFAM structural annotations of the GLEAN-R gene set to flag potentially transposable element-contaminated gene models (Supplementary Information section 4.2). These procedures suggest that 5.6–32.3% of gene models in non-*melanogaster* species correspond to protein-coding content derived from transposable elements (Supplementary Table 7); these transposable element-contaminated gene models are almost exclusively confined to gene predictions without strong homology support (Fig. 2). Transposable element-contaminated gene models are excluded from the final gene prediction set used for subsequent analysis, unless otherwise noted.

Homology assignment. Two independent approaches were used to assign orthology and paralogy relationships among euchromatic *D. melanogaster* gene models and GLEAN-R predictions. The first approach was a fuzzy reciprocal BLAST (FRB) algorithm, which is an

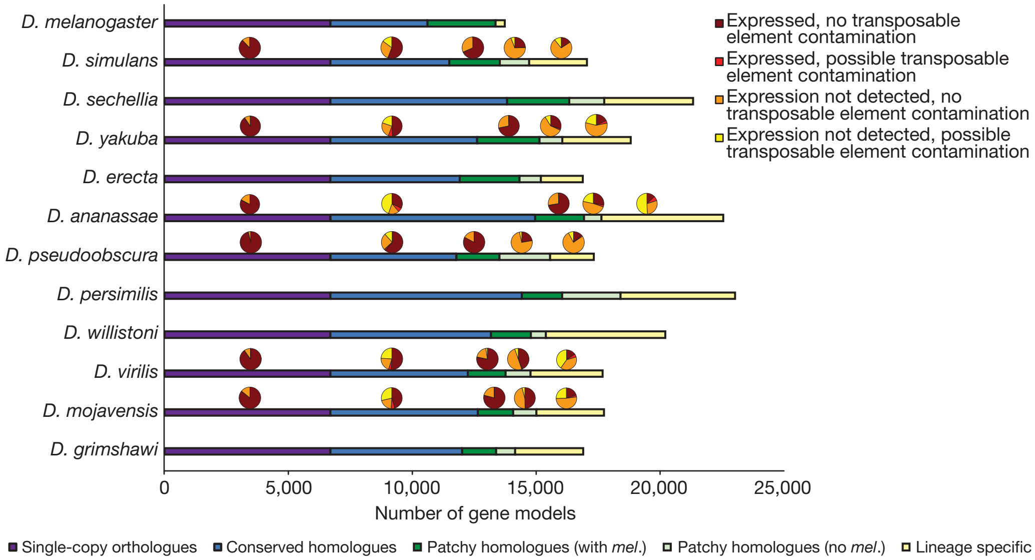


Figure 2 | Gene models in 12 *Drosophila* genomes. Number of gene models that fall into one of five homology classes: single-copy orthologues in all species (single-copy orthologues), conserved in all species as orthologues or paralogues (conserved homologues), a *D. melanogaster* homologue, but not found in all species (patchy homologues with *mel.*), conserved in at least two species but without a *D. melanogaster* homologue (patchy homologues, no *mel.*), and found only in a single lineage (lineage specific). For those species with expression data[29], pie charts indicate the fraction of genes in each homology class that fall into one of four evidence classes (see text for details).

extension of the reciprocal BLAST method[31] applicable to multiple species simultaneously (Supplementary Information section 5.1). Because the FRB algorithm does not integrate syntenic information, we also used a second approach based on Synpipe (Supplementary Information section 5.2), a tool for synteny-aided orthology assignment[32]. To generate a reconciled set of homology calls, pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB + Synpipe set of homology calls forms the basis of our subsequent analyses. There were 8,563 genes with single-copy orthologues in the *melanogaster* group and 6,698 genes with single-copy orthologues in all 12 species; similar numbers of genes were also obtained with an independent approach[33]. Most single-copy orthologues are expressed and are free from potential transposable element contamination, suggesting that the reconciled orthologue set contains robust and high-quality gene models (Fig. 2).

Validation of homology calls. Because both the FRB algorithm and Synpipe rely on BLAST-based methods to infer similarities, rapidly evolving genes may be overlooked. Moreover, assembly gaps and poor-quality sequence may lead to erroneous inferences of gene loss. To validate putative gene absences, we used a synteny-based GeneWise pipeline to find potentially missed homologues of *D. melanogaster* proteins (Supplementary Information section 5.4). Of the 21,928 cases in which a *D. melanogaster* gene was absent from another species in the initial homology call set, we identified plausible homologues for 13,265 (60.5%), confirmed 4,546 (20.7%) as genuine absences, and were unable to resolve 4,117 (18.8%). Because this approach is conservative and only confirms strongly supported absences, we are probably underestimating the number of genuine absences.

Coding gene alignment and filtering. Investigating the molecular evolution of orthologous and paralogous genes requires accurate multi-species alignments. Initial amino acid alignments were generated using TCOFFEE[34] and converted to nucleotide alignments (Supplementary Table 8). To reduce biases in downstream analyses, a simple computational screen was developed to identify and mask problematic regions of each alignment (Supplementary Information section 6). Overall, 2.8% of bases were masked in the *melanogaster* group alignments, and 3.0% of bases were masked in the full 12 species alignments, representing 8.5% and 13.8% of alignment columns, respectively. The vast majority of masked bases are masked in no more than one species (Supplementary Fig. 3), suggesting that the masking procedure is not simply eliminating rapidly evolving regions of the genome. We find an appreciably higher frequency of masked bases in lower-quality *D. simulans* and *D. sechellia* assemblies, compared to the more divergent (from *D. melanogaster*) but higher-quality *D. erecta* and *D. yakuba* assemblies, suggesting a higher error rate in accurately predicting and aligning gene models in lower-quality assemblies (Supplementary Information section 6 and Supplementary Fig. 3). We used masked versions of the alignments, including only the longest *D. melanogaster* transcripts for all subsequent analysis unless otherwise noted.

Annotation of non-coding (nc)RNA genes. Using *de novo* and homology-based approaches we annotated over 9,000 ncRNA genes from recognized ncRNA classes (Table 2, and Supplementary Information section 7). In contrast to the large number of predictions observed for many ncRNA families in vertebrates (due in part to large numbers of ncRNA pseudogenes[35,36]), the number of ncRNA genes per family predicted by RFAM and tRNAscan in *Drosophila* is relatively low (Table 2). This suggests that ncRNA pseudogenes are largely absent from *Drosophila* genomes, which is consistent with the low number of protein-coding pseudogenes in *Drosophila*[37]. The relatively low numbers of some classes of ncRNA genes (for example, small nucleolar (sno)RNAs) in the *Drosophila* subgenus are likely to be an artefact of rapid rates of evolution in these types of genes and the limitation of the homology-based methods used to annotate distantly related species.

Evolution of genome structure

Coarse-level similarities among Drosophilids. At a coarse level, genome structure is well conserved across the 12 sequenced species. Total genome size estimated by flow cytometry varies less than threefold across the phylogeny, ranging from 130 Mb (*D. mojavensis)* to 364 Mb (*D. virilis*)[38] (Table 2), in contrast to the order of magnitude difference between *Drosophila* and mammals. Total protein-coding sequence ranges from 38.9 Mb in *D. melanogaster* to 65.4 Mb in *D. willistoni.* Intronic DNA content is also largely conserved, ranging from 19.6 Mb in *D. simulans* to 24.0 Mb in *D. pseudoobscura* (Table 2). This contrasts dramatically with transposable element-derived genomic DNA content, which varies considerably across genomes (Table 2) and correlates significantly with euchromatic genome size (estimated as the summed length of contigs $>$ 200 kb) (Kendall's $\tau = 0.70$, $P = 0.0016$).

To investigate overall conservation of genome architecture at an intermediate scale, we analysed synteny relationships across species using Synpipe[32] (Supplementary Information section 9.1). Synteny block size and average number of genes per block varies across the phylogeny as expected, with the number of blocks increasing and the average size of blocks decreasing with increasing evolutionary distance from *D. melanogaster* (A. Bhutkar, S. Russo, T. F. Smith and W. M. Gelbart, personal communication) (Supplementary Fig. 4). We inferred 112 syntenic blocks between *D. melanogaster* and *D. sechellia* (with an average of 122 genes per block), compared to 1,406 syntenic blocks between *D. melanogaster* and *D. grimshawi* (with an average of 8 genes per block). On average, 66% of each genome assembly was covered by syntenic blocks, ranging from 68% in *D. sechellia* to 58% in *D. grimshawi.*

Similarity across genomes is largely recapitulated at the level of individual genes, with roughly comparable numbers of predicted protein-coding genes across the 12 species (Table 2). The majority of predicted genes in each species have homologues in *D. melanogaster* (Table 2, Supplementary Table 9). Moreover, most of the 13,733 protein-coding genes in *D. melanogaster* are conserved across the entire phylogeny: 77% have identifiable homologues in all 12 genomes, 62% can be identified as single-copy orthologues in the six genomes of the *melanogaster* group and 49% can be identified as single-copy orthologues in all 12 genomes. The number of functional non-coding RNA genes predicted in each *Drosophila* genome is also largely conserved, ranging from 584 in *D. mojavensis* to 908 in *D. ananassae* (Table 2).

There are several possible explanations for the observed interspecific variation in gene content. First, approximately 700 *D. melanogaster* gene models have been newly annotated since the FlyBase Release 4.3 annotations used in the current study, reducing the discrepancy between *D. melanogaster* and the other sequenced genomes in this study. Second, because low-coverage genomes tend to have more predicted gene models, we suspect that artefactual duplication of genomic segments due to assembly errors inflates the number of predicted genes in some species. Finally, the non-*melanogaster* species have many more predicted lineage-specific genes than *D. melanogaster*, and it is possible that some of these are artefactual. In the absence of experimental evidence, it is difficult to distinguish genuine lineage-specific genes from putative artefacts. Future experimental work will be required to fully disentangle the causes of interspecific variation in gene number.

Abundant genome rearrangements during *Drosophila* evolution. To study the structural relationships among genomes on a finer scale, we analysed gene-level synteny between species pairs. These synteny maps allowed us to infer the history and locations of fixed genomic rearrangements between species. Although *Drosophila* species vary in their number of chromosomes, there are six fundamental chromosome arms common to all species. For ease of denoting chromosomal homology, these six arms are referred to as 'Muller elements' after Hermann J. Muller, and are denoted A–F. Although most pairs of orthologous genes are found on the same Muller element, there is extensive gene shuffling within Muller elements between even moderately diverged genomes (Fig. 3, and Supplementary Information section 9.1).

Previous analysis has revealed heterogeneity in rearrangement rates among close relatives: careful inspection of 29 inversions that differentiate the chromosomes of *D. melanogaster* and *D. yakuba* revealed that 28 were fixed in the lineage leading to *D. yakuba*, and only one was fixed on the lineage leading to *D. melanogaster*[39]. Rearrangement rates are also heterogeneous across the genome among the 12 species: simulations reject a random-breakage model, which assumes that all sites are free to break in inversion events, but fail to reject a model of coldspots and hotspots for breakpoints (S. Schaeffer, personal communication). Furthermore, inversions seem to have played important roles in the process of speciation in at least some of these taxa[40].

One particularly striking example of the dynamic nature of genome micro-structure in *Drosophila* is the homeotic *homeobox (Hox)* gene cluster(s)[41]. *Hox* genes typically occur in genomic clusters, and this clustering is conserved across many vertebrate and invertebrate taxa, suggesting a functional role for the precise and collinear arrangement of these genes. However, several cluster splits have been previously identified in *Drosophila*[42,43], and the 12 *Drosophila* genome sequences provide additional evidence against the functional importance of *Hox* gene clustering in *Drosophila*. There are seven different gene arrangements found across 13 *Drosophila* species (the 12 sequenced genomes and *D. buzzatii*), with no species retaining the inferred ancestral gene order[44]. It thus seems that, in *Drosophila, Hox* genes do not require clustering to maintain proper function, and are a powerful illustration of the dynamism of genome structure across the sequenced genomes.

Transposable element evolution. Mobile, repetitive transposable element sequences are a particularly dynamic component of eukaryotic genomes. Transposable element/repeat content (in scaffolds >200 kb) varies by over an order of magnitude across the genus, ranging from ~2.7% in *D. simulans* and *D. grimshawi* to ~25% in *D. ananassae* (Table 2, and Supplementary Fig. 1). These data support the lower euchromatic transposable element content in *D. simulans* relative to *D. melanogaster*[45], and reveal that euchromatic transposable element/repeat content is generally similar within the *melanogaster* subgroup. Within the *Drosophila* subgenus, *D. grimshawi* has the lowest transposable element/repeat content, possibly relating to its ecological status as an island endemic, which may minimize the chance for horizontal transfer of transposable element families. Finally, the highest levels of transposable element/repeat content are found in *D. ananassae* and *D. willistoni*. These species also have the highest numbers of pseudo-transfer (t)RNA genes (Table 2), indicating a potential relationship between pseudo-tRNA genesis and repetitive DNA, as has been established in the mouse genome[36].

Different classes of transposable elements can vary in abundance owing to a variety of host factors, motivating an analysis of the intragenomic ecology of transposable elements in the 12 genomes. In *D. melanogaster*, long terminal repeat (LTR) retrotransposons have the highest abundance, followed by LINE (long interspersed nuclear element)-like retrotransposons and terminal inverted repeat (TIR) DNA-based transposons[18]. An unbiased, conservative approach (Supplementary Information section 3) for estimating the rank order abundance of major transposable element classes suggests that these abundance trends are conserved across the entire genus (Supplementary Fig. 5). Two exceptions are an increased abundance of TIR elements in *D. erecta* and a decreased abundance of LTR elements in *D. pseudoobscura*; the latter observation may represent an assembly artefact because the sister species *D. persimilis* shows typical LTR abundance. Given that individual instances of transposable element repeats and transposable element families themselves are not conserved across the genus, the stability of abundance trends for different classes of transposable elements is striking and suggests common mechanisms for host–transposable element co-evolution in *Drosophila*.

Although comprehensive analysis of the structural and evolutionary relationships among families of transposable elements in the 12 genomes remains a major challenge for *Drosophila* genomics, some initial insights can be gleaned from analysis of particularly well-characterized transposable element families. Previous analysis has shown variable dynamics for the most abundant transposable element family *(DINE-1)*[46] in the *D. melanogaster* genome[18,47]: although inactive in *D. melanogaster*[48], *DINE-1* has experienced a recent transpositional burst in *D. yakuba*[49]. Our analysis confirms that this element is highly abundant in all of the other sequenced genomes of *Drosophila*, but is not found outside of Diptera[50,51]. Moreover, the inferred phylogenetic relationship of *DINE-1* paralogues from several *Drosophila* species suggests vertical transmission as the major mechanism for *DINE-1* propagation. Likewise, analysis of the *Galileo*

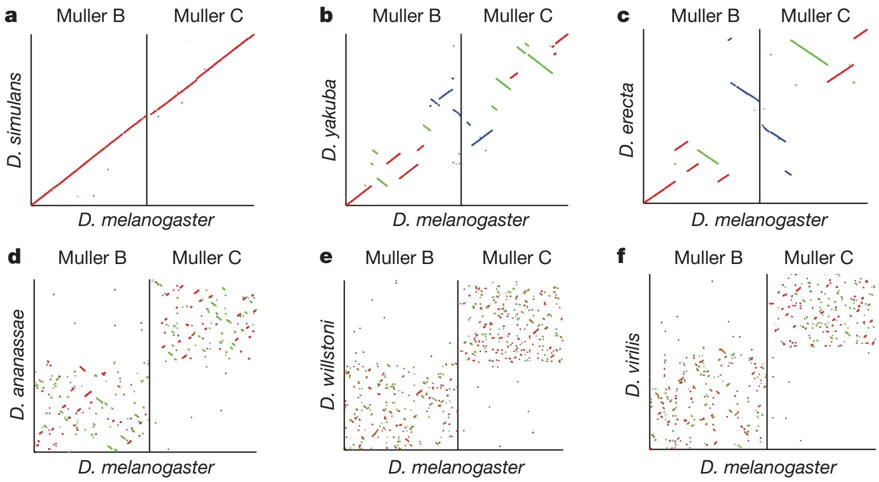


Figure 3 | Synteny plots for Muller elements B and C with respect to *D. melanogaster* gene order. The horizontal axis shows *D. melanogaster* gene order for Muller elements B and C, and the vertical axis maps homologous locations[32,155] in individual species (**a**–**f** in increasing evolutionary distance from *D. melanogaster)*. Left to right on the *x* axis is from telomere to centromere for Muller element B, followed by Muller element C from centromere to telomere. Red and green lines represent syntenic segments in the same or reverse orientation along the chromosome relative to *D. melanogaster*, respectively. Blue segments show gene transposition of genes from one element to the other.

and *1360* transposons reveals a widespread but discontinuous phylogenetic distribution for both families, notably with both families absent in the geographically isolated Hawaiian species, *D. grimshawi*[52]. These results are consistent with an ancient origin of the *Galileo* and *1360* families in the genus and subsequent horizontal transfer and/or loss in some lineages.

The use of these 12 genomes also facilitated the discovery of transposable element lineages not yet documented in *Drosophila*, specifically the P instability factor (*PIF*) superfamily of DNA transposons. Our analysis indicates that there are four distinct lineages of this transposon in *Drosophila*, and that this element has indeed colonized many of the sequenced genomes[53]. This superfamily is particularly intriguing given that *PIF*-transposase-like genes have been implicated in the origin of at least seven different genes during the *Drosophila* radiation[53], suggesting that not only do transposable elements affect the evolution of genome structure, but that their domestication can play a part in the emergence of novel genes.

D. melanogaster maintains its telomeres by occasional targeted transposition of three telomere-specific non-LTR retrotransposons (*HeT-A*, *TART* and *TAHRE*) to chromosome ends[54,55] and not by the more common mechanism of telomerase-generated G-rich repeats[56]. Multiple telomeric retrotransposons have originated within the genus, where they now maintain telomeres, and recurrent loss of most of the ORF2 from telomeric retrotransposons (for example, *TAHRE*) has given rise to half-telomeric-retrotransposons (for example, *HeT-A*) during *Drosophila* evolution[57]. The phylogenetic relationship among these telomeric elements is congruent with the species phylogeny, suggesting that they have been vertically transmitted from a common ancestor[57].

ncRNA gene family evolution. Using ncRNA gene annotations across the 12-species phylogeny, we inferred patterns of gene copy number evolution in several ncRNA families. Transfer RNA genes are the most abundant family of ncRNA genes in all 12 genomes, with 297 tRNAs in *D. melanogaster* and 261–484 tRNA genes in the other species (Table 2). Each genome encodes a single selenocysteine tRNA, with the exception of *D. willistoni*, which seems to lack this gene (R. Guigo, personal communication). Elevated tRNA gene counts in *D. ananassae* and *D. willistoni* are explained almost entirely by pseudo-tRNA gene predictions. We infer from the lack of pseudo-tRNAs in most *Drosophila* species, and from similar numbers of tRNAs obtained from an analysis of the chicken genome ($n = 280$)[58], that the minimal metazoan tRNA set is encoded by ~300 genes, in contrast to previous estimates of 497 in human and 659 in *Caenorhabditis elegans*[59,60]. Similar numbers of snoRNAs are predicted in the *D. melanogaster* subgroup ($n = 242$–255), in which sequence similarity is high enough for annotation by homology, with fewer snoRNAs ($n = 194$–216) annotated in more distant members of the *Sophophora* subgenus, and even fewer snoRNAs ($n = 139$–165) predicted in the *Drosophila* subgenus, in which annotation by homology becomes much more difficult.

Of 78 previously reported micro (mi)RNA genes, 71 (91%) are highly conserved across the entire genus, with the remaining seven genes (*mir-2b-1*, *-289*, *-303*, *-310*, *-311*, *-312* and *-313*) restricted to the subgenus *Sophophora* (Supplementary Information section 7.2). All the species contain similar numbers of spliceosomal snRNA genes (Table 2), including at least one copy each of the four U12-dependent (minor) spliceosomal RNAs, despite evidence for birth and death of these genes and the absence of stable subtypes[61]. The unusual, lineage-specific expansion in size of U11 snRNA, previously described in *Drosophila*[61,62], is even more extreme in *D. willistoni*. We annotated 99 copies of the 5S ribosomal (r)RNA gene in a cluster in *D. melanogaster*, and between 13 and 73 partial 5S rRNA genes in clusters in the other genomes. Finally, we identified members of several other classes of ncRNA genes, including the RNA components of the RNase P (1 per genome) and the signal recognition particle (SRP) RNA complexes (1–3 per genome), suggesting that these functional RNAs are involved in similar biological processes throughout the genus. We were only able to locate the *roX* (RNA on X)[63,64] genes involved in dosage compensation using nucleotide homology in the *melanogaster* subgroup, although analyses incorporating structural information have identified *roX* genes in other members of the genus[65].

We investigated the evolution of rRNA genes in the 12 sequenced genomes, using trace archives to locate sequence variants within the transcribed portions of these genes. This analysis revealed moderate levels of variation that are not distributed evenly across the rRNA genes, with fewest variants in conserved core coding regions, more variants in coding expansion regions, and higher still variant abundances in non-coding regions. The level and distribution of sequence variation in rRNA genes are suggestive of concerted evolution, in which recombination events uniformly distribute variants throughout the rDNA loci, and selection dictates the frequency to which variants can expand[66].

Protein-coding gene family evolution. For a general perspective on how the protein-coding composition of these 12 genomes has changed, we examined gene family expansions and contractions in the 11,434 gene families (including those of size one in each species) predicted to be present in the most recent common ancestor of the two subgenera. We applied a maximum likelihood model of gene gain and loss[67] to estimate rates of gene turnover. This analysis suggests that gene families expand or contract at a rate of 0.0012 gains and losses per gene per million years, or roughly one fixed gene gain/loss across the genome every 60,000 yr[68]. Many gene families (4,692 or 41.0%) changed in size in at least one species, and 342 families showed significantly elevated ($P < 0.0001$) rates of gene gain and loss compared to the genomic average, indicating that non-neutral processes may play a part in gene family evolution. Twenty-two families exhibit rapid copy number evolution along the branch leading to *D. melanogaster* (eighteen contractions and four expansions; Supplementary Table 10). The most common Gene Ontology (GO) terms among families with elevated rates of gain/loss include 'defence response', 'protein binding', 'zinc ion binding', 'proteolysis', and 'trypsin activity'. Interestingly, genes involved in 'defence response' and 'proteolysis' also show high rates of protein evolution (see below). We also found heterogeneity in overall rates of gene gain and loss across lineages, although much of this variation could result from interspecific differences in assembly quality[68].

Lineage-specific genes. The vast majority of *D. melanogaster* proteins that can be unambiguously assigned a homology pattern (Supplementary Information section 5) are inferred to be ancestrally present at the genus root (11,348/11,644, or 97.5%). Of the 296 non-ancestrally present genes, 252 are either *Sophophora*-specific, or have a complicated pattern of homology requiring more than one gain and/or loss on the phylogeny, and are not discussed further. The remaining 44 proteins include 14 present in the *melanogaster* group, 23 present only in the *melanogaster* subgroup, 3 unique to the *melanogaster* species complex, and 4 found in *D. melanogaster* only. Because we restricted this analysis to unambiguous homologues of high-confidence protein-coding genes in *D. melanogaster*[8], we are probably undercounting the number of genes that have arisen *de novo* in any particular lineage. However, ancestrally heterochromatic genes that are currently euchromatic in *D. melanogaster* may spuriously seem to be lineage-specific.

The 44 lineage-specific genes (Supplementary Table 11) differ from ancestrally present genes in several ways. They have a shorter median predicted protein length (lineage-specific median 177 amino acids, other median 421 amino acids, MWU, $P = 3.6 \times 10^{-13}$), are more likely to be intronless (Fisher's exact test (FET), $P = 6.2 \times 10^{-6}$), and are more likely to be located in the intron of another gene on the opposite strand (FET, $P = 3.5 \times 10^{-4}$). In addition, 18 of these 44 genes are testis- or accessory-gland-specific in *D. melanogaster*, a significantly greater fraction than is found in the ancestral set (FET, $P = 1.25 \times 10^{-4}$). This is consistent with previous observations that novel genes are often testis-specific in *Drosophila*[69–73] and expression studies on seven of the species show that species-restricted genes are more likely to exhibit male-biased expression[29]. Further, these genes are significantly more tissue-specific in expression (as measured by τ; ref. 74) (MWU, $P = 9.6 \times 10^{-6}$), and this pattern is not solely driven by genes with testis-specific expression patterns.

Protein-coding gene evolution

Positive selection and selective constraints in *Drosophila* genomes. To study the molecular evolution of protein-coding genes, we estimated rates of synonymous and non-synonymous substitution in 8,510 single-copy orthologues within the six *melanogaster* group species using PAML[75] (Supplementary Information section 11.1); synonymous site saturation prevents analysis of more divergent comparisons. We investigate only single-copy orthologues because when paralogues are included, alignments become increasingly problematic. Rates of amino acid divergence for single-copy orthologues in all 12 species were also calculated; these results are largely consistent with the analysis of non-synonymous divergence in the *melanogaster* group, and are not discussed further.

To understand global patterns of divergence and constraint across functional classes of genes, we examined the distributions of ω ($= d_N/d_S$, the ratio of non-synonymous to synonymous divergence) across Gene Ontology categories (GO)[76], excluding GO annotations based solely on electronic support (Supplementary Information section 11.2). Most functional categories of genes are strongly constrained, with median estimates of ω much less than one. In general, functionally similar genes are similarly constrained: 31.8% of GO categories have significantly lower variance in ω than expected (*q*-value true-positive test[77]). Only 11% of GO categories had statistically significantly elevated ω (relative to the median of all genes with GO annotations) at a 5% false-discovery rate (FDR), suggesting either positive selection or a reduction in selective constraint. The GO categories with elevated ω include the biological process terms 'defence response', 'proteolysis', 'DNA metabolic process' and 'response to biotic stimulus'; the molecular function terms 'transcription factor activity', 'peptidase activity', 'receptor binding', 'odorant binding', 'DNA binding', 'receptor activity' and 'G-protein-coupled receptor activity'; and the cellular location term 'extracellular' (Fig. 4, and Supplementary Table 12). Similar results are obtained when d_N is compared across GO categories, suggesting that in most cases differences in ω among GO categories is driven by amino acid rather than synonymous site substitutions. The two exceptions are the molecular function terms 'transcription factor activity' and 'DNA binding activity', for which we observe significantly decelerated d_S (FDR $= 7.2 \times 10^{-4}$ for both; Supplementary Information section 11.2) and no significant differences in d_N.

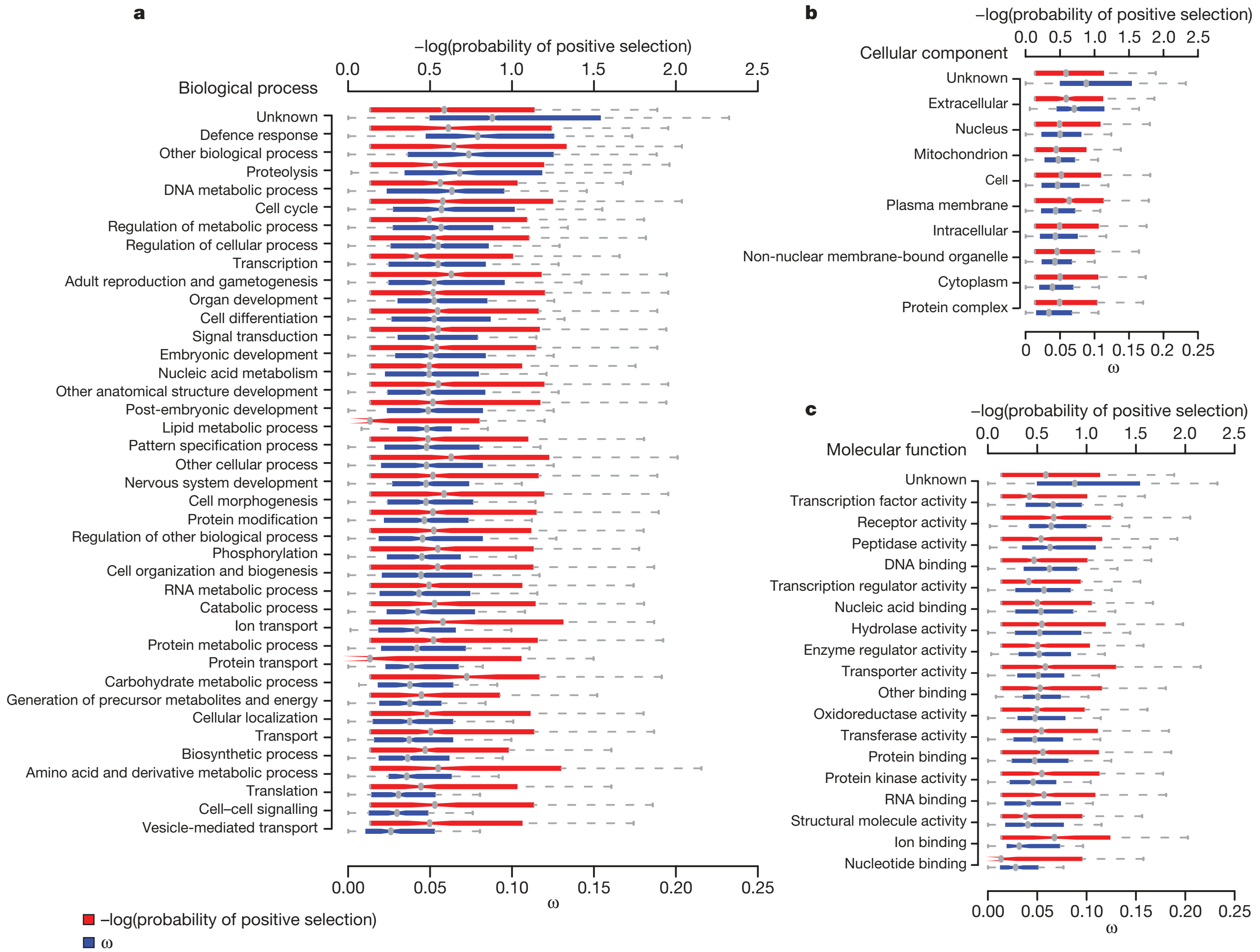


Figure 4 | Patterns of constraint and positive selection among GO terms. Distribution of average ω per gene and the negative $\log_{10}$ of the probability of positive selection (Supplementary Information section 11.2) for genes annotated with: **a**, biological process GO terms; **b**, cellular component GO terms; and **c**, molecular function GO terms. Only GO terms with 200 or more genes annotated are plotted. See Supplementary Table 12 for median values and significance. Note that most genes evolve under evolutionary constraint at most of their sites, leading to low values of ω; even genes that experience positive selection do not typically have an average ω across all codons that exceeds one.

To distinguish possible positive selection from relaxed constraint, we tested explicitly for genes that have a subset of codons with signatures of positive selection, using codon-based likelihood models of molecular evolution, implemented in PAML[78,79] (Supplementary Information section 11.1). Although this test is typically regarded as a conservative test for positive selection, it may be confounded by selection at synonymous sites. However, selection at synonymous sites (that is, codon bias, see below) is quite weak. Moreover, variability in ω presented here tends to reflect variability in d_N. We therefore believe that it is appropriate to treat synonymous sites as nearly neutral and sites with $\omega > 1$ as consistent with positive selection. Despite a number of functional categories with evidence for elevated ω, 'helicase activity' is the only functional category significantly more likely to be positively selected (permutation test, $P = 2 \times 10^{-4}$, FDR = 0.007; Supplementary Table 12); the biological significance of this finding merits further investigation. Furthermore, within each GO class, there is greater dispersion among genes in their probability of positive selection than in their estimate of ω (MWU one-tailed, $P = 0.011$; Supplementary Information section 11.1), suggesting that although functionally similar genes share patterns of constraint, they do not necessarily show similar patterns of positive selection (Fig. 4).

Interestingly, protein-coding genes with no annotated ('unknown') function in the GO database seem to be less constrained (permutation test, $P < 1 \times 10^{-4}$, FDR = 0.006)[80] and to have on average lower P-values for the test of positive selection than genes with annotated functions (permutation test, $P = 0.001$, FDR = 0.058). It is unlikely that this observation results entirely from an over-representation of mis-annotated or non-protein-coding genes in the 'unknown' functional class, because this finding is robust to the removal of all *D. melanogaster* genes predicted to be non-protein-coding in ref. 8. The bias in the way biological function is ascribed to genes (to laboratory-induced, easily scorable functions) leaves open the possibility that unannotated biological functions may have an important role in evolution. Indeed, genes with characterized mutant alleles in FlyBase evolve significantly more slowly than other genes (median $\omega_{\text{with alleles}} = 0.0525$ and $\omega_{\text{without alleles}} = 0.0701$; MWU, $P < 1 \times 10^{-16}$).

Previous work has suggested that a substantial fraction of non-synonymous substitutions in *Drosophila* were fixed through positive selection[81–85]. We estimate that 33.1% of single-copy orthologues in the *melanogaster* group have experienced positive selection on at least a subset of codons (q-value true-positive tests[77]) (Supplementary Information section 11.1). This may be an underestimate, because we have only examined single-copy orthologues, owing to difficulties in producing accurate alignments of paralogues by automated methods. On the basis of the 878 genes inferred to have experienced positive selection with high confidence (FDR < 10%), we estimated that an average of 2% of codons in positively selected genes have $\omega > 1$. Thus, several lines of evidence, based on different methodologies, suggest that patterns of amino acid fixation in *Drosophila* genomes have been shaped extensively by positive selection.

The presence of functional domains within a protein may lead to heterogeneity in patterns of constraint and adaptation along its length. Among genes inferred to be evolving by positive selection at a 10% FDR, 63.7% (q-value true-positive tests[77]) show evidence for spatial clustering of positively selected codons (Supplementary Information section 11.2). Spatial heterogeneity in constraint is further supported by contrasting ω for codons inside versus outside defined InterPro domains (genes lacking InterPro domains are treated as 'outside' a defined InterPro domain). Codons within InterPro domains were significantly more conserved than codons outside InterPro domains (median ω: 0.062 InterPro domains, 0.084 outside InterPro domains; MWU, $P < 2.2 \times 10^{-16}$; Supplementary Information section 11.2). Similarly, there were significantly more positively selected codons outside of InterPro domains than inside domains (FET $P < 2.2 \times 10^{-16}$), suggesting that in addition to being more constrained, codons in protein domains are less likely to be targets of positive selection (Supplementary Fig. 6).

Factors affecting the rate of protein evolution in *Drosophila*. The sequenced genomes of the *melanogaster* group provide unprecedented statistical power to identify factors affecting rates of protein evolution. Previous analyses have suggested that although the level of gene expression consistently seems to be a major determinant of variation in rates of evolution among proteins[86,87], other factors probably play a significant, if perhaps minor, part[88–91]. In *Drosophila*, although highly expressed genes do evolve more slowly, breadth of expression across tissues, gene essentiality and intron number all also independently correlate with rates of protein evolution, suggesting that the additional complexities of multicellular organisms are important factors in modulating rates of protein evolution[78]. The presence of repetitive amino acid sequences has a role as well: non-repeat regions in proteins containing repeats evolve faster and show more evidence for positive selection than genes lacking repeats[92].

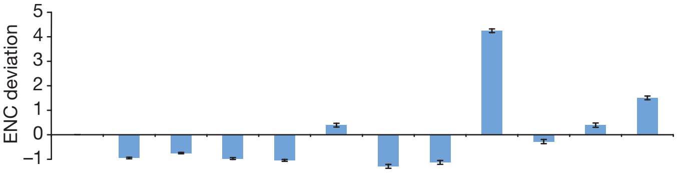
These data also provide a unique opportunity to examine the impact of chromosomal location on evolutionary rates. Population genetic theory predicts that for new recessive mutations, both purifying and positive selection will be more efficient on the X chromosome given its hemizygosity in males[93]. In contrast, the lack of recombination on the small, mainly heterochromatic dot chromosome[94,95] is expected to reduce the efficacy of selection[96]. Because codon bias, or the unequal usage of synonymous codons in protein-coding sequences, reflects weak but pervasive selection, it is a sensitive metric for evaluating the efficacy of purifying selection. Consistent with expectation, in all 12 species, we find significantly elevated levels of codon bias on the X chromosome and significantly reduced levels of codon bias on the dot chromosome[97]. Furthermore, X-chromosome-linked genes are marginally over-represented within the set of positively selected genes in the *melanogaster* group (FET, $P = 0.055$), which is consistent with increased rates of adaptive substitution on this chromosome. This analysis suggests that chromosomal context also serves to modulate rates of molecular evolution in protein-coding genes.

To examine further the impact of genomic location on protein evolution, we examined the subset of genes that have moved within or between chromosome arms[32,98]. Genes inferred to have moved between Muller elements have a significantly higher rate of protein evolution than genes inferred to have moved within a Muller element (MWU, $P = 1.32 \times 10^{-14}$) and genes that have maintained their genomic position (MWU, $P = 0.008$) (Supplementary Fig. 7). Interestingly, genes that move within Muller elements have a significantly lower rate of protein evolution than those for which genomic locations have been maintained (MWU, $P = 3.85 \times 10^{-14}$). It remains unclear whether these differences reflect underlying biases in the types of genes that move inter- versus intra-chromosomally, or whether they are due to *in situ* patterns of evolution in novel genomic contexts.

Codon bias. Codon bias is thought to enhance the efficiency and/or accuracy of translation[99–101] and seems to be maintained by mutation–selection–drift balance[101–104]. Across the 12 *Drosophila* genomes, there is more codon bias in the *Sophophora* subgenus than in the *Drosophila* subgenus, and a previously noted[105–109] striking reduction in codon bias in *D. willistoni*[110,111] (Fig. 5). However, with only minor exceptions, codon preferences for each amino acid seem to be conserved across 11 of the 12 species. The striking exception is *D. willistoni*, in which codon usage for 6 of 18 redundant amino acids has diverged (Fig. 5). Mutation alone is not sufficient to explain codon-usage bias in *D. willistoni*, which is suggestive of a lineage-specific shift in codon preferences[111,112]. We found evidence for a lineage-specific genomic reduction in codon bias in *D. melanogaster* (Fig. 5), as has been suggested previously[113–119]. In addition, maximum-likelihood estimation of the strength of selection on synonymous sites in 8,510 *melanogaster* group single-copy orthologues revealed a marked reduction in the number of genes under selection for increased codon bias in *D. melanogaster* relative to its sister species *D. sechellia*[120].

Evolution of genes associated with ecology and reproduction. Given the ecological and environmental diversity encompassed by the 12 *Drosophila* species, we examined the evolution of genes and gene families associated with ecology and reproduction. Specifically, we selected genes with roles in chemoreception, detoxification/metabolism, immunity/defence, and sex/reproduction for more detailed study.

Chemoreception. *Drosophila* species have complex olfactory and gustatory systems used to identify food sources, hazards and mates, which depend on odorant-binding proteins, and olfactory/odorant and gustatory receptors (*Or*s and *Gr*s). The *D. melanogaster* genome has approximately 60 *Or*s, 60 *Gr*s and 50 odorant-binding protein genes. Despite overall conservation of gene number across the 12 species and widespread evidence for purifying selection within the *melanogaster* group, there is evidence that a subset of *Or* and *Gr* genes experiences positive selection[121–123]. Furthermore, clear lineage-specific differences are detectable between generalist and specialist species within the *melanogaster* subgroup. First, the two independently evolved specialists (*D. sechellia* and *D. erecta*) are losing *Gr* genes approximately five times more rapidly than the generalist species[121,124]. We believe this result is robust to sequence quality, because all pseudogenes and deletions were verified by direct re-sequencing and synteny-based orthologue searches, respectively. Generalists are expected to encounter the most diverse set of tastants and seem to have maintained the greatest diversity of gustatory receptors. Second, *Or* and *Gr* genes that remain intact in *D. sechellia* and *D. erecta* evolve significantly more rapidly along these two lineages ($\omega = 0.1556$ for *Or*s and 0.1874 for *Gr*s) than along the generalist lineages ($\omega = 0.1049$ for *Or*s and 0.1658 for *Gr*s; paired Wilcoxon, $P = 0.0003$ and 0.003, respectively[124]). There is some evidence that odorant-binding protein genes also evolve significantly faster in specialists compared to generalists[122]. This elevated ω reflects a trend observed throughout the genomes of the two specialists and is likely to result, at least in part, from demographic phenomena. However, the difference between specialist and generalist ω for *Or*/*Gr* genes (0.0292) is significantly greater than the difference for genes across the genome (0.0091; MWU, $P = 0.0052$)[121], suggesting a change in selective regime. Moreover, the observation that elevated ω as well as accelerated gene loss disproportionately affect groups of *Or* and *Gr* genes that respond to specific chemical ligands and/or are expressed during specific life stages suggests that rapid evolution at *Or*/*Gr* loci in specialists is related to the ecological shifts these species have sustained[121].



	D. mel	D. sim	D. sec	D. yak	D. ere	D. ana	D. pse	D. per	D. wil	D. vir	D. moj	D. gri
Arg	CGC	•	•	•	•	•	•	•	CGT	•	•	•
Leu	CTG	•	•	•	•	•	•	•	TTG	•	•	•
Ser	TCC	•	•	•	•	•	AGC	AGC	•	AGC	AGC	AGC
Asn	AAC	•	•	•	•	•	•	•	AAT	•	•	•
His	CAC	•	•	•	•	•	•	•	GAC	•	•	•
Asp	GAC	•	•	•	•	•	•	•	GAT	•	•	•
Tyr	TAC	•	•	•	•	•	•	•	TAT	•	•	•

Figure 5 | Deviations in codon bias from *D. melanogaster* in 11 *Drosophila* species. The upper panel depicts differences in ENC (effective number of codons) between *D. melanogaster* and the 11 non-*melanogaster* species, calculated on a gene-by-gene basis. Note that increasing levels of ENC indicates a decrease in codon bias. The *Sophophora* subgenus in general has higher levels of codon bias than the *Drosophila* subgenus with the exception of *D. willistoni,* which shows a dramatic reduction in codon bias. The lower panel shows the 7 codons for which preference changes across the 12 *Drosophila* species. A dot indicates identical codon preference to *D. melanogaster*; otherwise the preferred codon is indicated.

Detoxification/metabolism. The larval food sources for many *Drosophila* species contain a cocktail of toxic compounds, and consequently *Drosophila* genomes encode a wide variety of detoxification proteins. These include members of the cytochrome P450 (P450), carboxyl/choline-esterase (CCE) and glutathione *S*-transferase (GST) multigene families, all of which also have critical roles in resistance to insecticides[125–127]. Among the P450s, the five enzymes associated with insecticide resistance are highly dynamic across the phylogeny, with 24 duplication events and 4 loss events since the last common ancestor of the genus, which is in striking contrast to genes with known developmental roles, eight of which are present as a single copy in all 12 species (C. Robin, personal communication). As with chemoreceptors, specialists seem to lose detoxification genes at a faster rate than generalists. For instance, *D. sechellia* has lost the most P450 genes; these 14 losses comprise almost one-third of all P450 loss events (Supplementary Table 13) (C. Robin, personal communication). Positive selection has been implicated in detoxification-gene evolution as well, because a search for positive selection among GSTs identified the parallel evolution of a radical glycine to lysine amino acid change in GSTD1, an enzyme known to degrade DDT[128]. Finally, although metabolic enzymes in general are highly constrained (median $\omega = 0.045$ for enzymes, 0.066 for non-enzymes; MWU, $P = 5.7 \times 10^{-24}$), enzymes involved in xenobiotic metabolism evolve significantly faster than other enzymes (median $\omega = 0.05$ for the xenobiotic group versus $\omega = 0.045$ overall, two-tailed permutation test, $P = 0.0110$; A. J. Greenberg, personal communication).

Metazoans deal with excess selenium in the diet by sequestration in selenoproteins, which incorporate the rare amino acid selenocysteine (Sec) at sites specified by the TGA codon. The recoding of the normally terminating signal TGA as a Sec codon is mediated by the selenocystein insertion sequence (SECIS), a secondary structure in the 3′ UTR of selenoprotein messenger RNAs. All animals examined so far have selenoproteins; three have been identified in *D. melanogaster* (SELG, SELM and SPS2[129,130]). Interestingly, although the three known *melanogaster* selenoproteins are all present in the genomes of the other *Drosophila* species, in *D. willistoni* the TGA Sec codons have been substituted by cysteine codons (TGT/TGC). Consistent with this finding, analysis of the seven genes implicated to date in selenoprotein synthesis including the Sec-specific tRNA suggests that most of these genes are absent in *D. willistoni* (R. Guigo, personal communication). *D. willistoni* thus seems to be the first animal known to lack selenoproteins. If correct, this observation is all the more remarkable given the ubiquity of selenoproteins and the selenoprotein biosynthesis machinery in metazoans, the toxicity of excess selenium, and the protection from oxidative stress mediated by selenoproteins. However, it remains possible that this species encodes selenoproteins in a different way, and this represents an exciting avenue of future research.

Immunity/defence. *Drosophila*, like all insects, possesses an innate immune system with many components analogous to the innate immune pathways of mammals, although it lacks an antibody-mediated adaptive immune system[131]. Immune system genes often evolve rapidly and adaptively, driven by selection pressures from pathogens and parasites[132–134]. The genus *Drosophila* is no exception: immune system genes evolve more rapidly than non-immune genes, showing both high total divergence rates and specific signs of positive selection[135]. In particular, 29% of receptor genes involved in phagocytosis seem to evolve under positive selection, suggesting that molecular co-evolution between *Drosophila* pattern recognition receptors and pathogen antigens is driving adaptation in the immune system[135]. Somewhat surprisingly, genes encoding effector proteins such as antimicrobial peptides are far less likely to exhibit adaptive sequence evolution. Only 5% of effector genes (and no antimicrobial peptides) show evidence of adaptive evolution, compared to 10% of genes genome-wide. Instead, effector genes seem to evolve by rapid duplication and deletion. Whereas 49% of genes genome-wide, 63%

of genes involved in pathogen recognition and 81% of genes implicated in immune-related signal transduction can be found as single-copy orthologues in all 12 species, only 40% of effector genes exist as single-copy orthologues across the genus ($\chi2 = 41.13$, $P = 2.53 \times 10^{-8}$), suggesting rapid radiation of effector protein classes along particular lineages[135]. Thus, much of the *Drosophila* immune system seems to evolve rapidly, although the mode of evolution varies across immune-gene functional classes.

Sex/reproduction. Genes encoding sex- and reproduction-related proteins are subject to a wide array of selective forces, including sexual conflict, sperm competition and cryptic female choice, and to the extent that these selective forces are of evolutionary consequence, this should lead to rapid evolution in these genes[136] (for an overview see refs 137, 138). The analysis of 2,505 sex- and reproduction-related genes within the *melanogaster* group indicated that male sex- and reproduction-related genes evolve more rapidly at the protein level than genes not involved in sex or reproduction or than female sex- and reproduction-related genes (Supplementary Fig. 8). Positive selection seems to be at least partially responsible for these patterns, because genes involved in spermatogenesis have significantly stronger evidence for positive selection than do non-spermatogenesis genes (permutation test, $P = 0.0053$). Similarly, genes that encode components of seminal fluid have significantly stronger evidence for positive selection than 'non-sex' genes[139]. Moreover, protein-coding genes involved in male reproduction, especially seminal fluid and testis genes, are particularly likely to be lost or gained across *Drosophila* species[29,139].

Evolutionary forces in the mitochondrial genome. Functional elements in mtDNA are strongly conserved, as expected: tRNAs are relatively more conserved than the mtDNA overall (average pairwise nucleotide distance = 0.055 substitutions per site for tRNAs versus 0.125 substitutions per site overall). We observe a deficit of substitutions occurring in the stem regions of the stem-loop structure in tRNAs, consistent with strong selective pressure to maintain RNA secondary structure, and there is a strong signature of purifying selection in protein-coding genes[13]. However, despite their shared role in aerobic respiration, there is marked heterogeneity in the rates of amino acid divergence between the oxidative phosphorylation enzyme complexes across the 12 species (NADH dehydrogenase, 0.059 > ATPase, 0.042 > CytB, 0.037 > cytochrome oxidase, 0.020; mean pairwise d_N), which contrasts with the relative homogeneity in synonymous substitution rates. A model with distinct substitution rates for each enzyme complex rather than a single rate provides a significantly better fit to the data ($P < 0.0001$), suggesting complex-specific selective effects of mitochondrial mutations[13].

Non-coding sequence evolution

ncRNA sequence evolution. The availability of complete sequence from 12 *Drosophila* genomes, combined with the tractability of RNA structure predictions, offers the exciting opportunity to connect patterns of sequence evolution directly with structural and functional constraints at the molecular level. We tested models of RNA evolution focusing on specific ncRNA gene classes in addition to inferring patterns of sequence evolution using more general datasets that are based on predicted intronic RNA structures.

The exquisite simplicity of miRNAs and their shared stem-loop structure makes these ncRNAs particularly amenable to evolutionary analysis. Most miRNAs are highly conserved within the *Drosophila* genus: for the 71 previously described miRNA genes inferred to be present in the common ancestor of these 12 species, mature miRNA sequences are nearly invariant. However, we do find a small number of substitutions and a single deletion in mature miRNA sequences (Supplementary Table 14), which may have functional consequences for miRNA–target interactions and may ultimately help identify targets through sequence covariation. Pre-miRNA sequences are also highly conserved, evolving at about 10% of the rate of synonymous sites.

To link patterns of evolution with structural constraints, we inferred ancestral pre-miRNA sequences and deduced secondary structures at each ancestral node on the phylogeny (Supplementary Information section 12.1). Although conserved miRNA genes show little structural change (little change in free energy), the five *melanogaster* group-specific miRNA genes (*miR-303* and the *mir-310/311/312/313* cluster) have undergone numerous changes across the entire pre-miRNA sequence, including the ordinarily invariant mature miRNA. Patterns of polymorphism and divergence in these lineage-specific miRNA genes, including a high frequency of derived mutations, are suggestive of positive selection[140]. Although lineage-specific miRNAs may evolve under less constraint because they have fewer target transcripts in the genome, it is also possible that recent integration into regulatory networks causes accelerated rates of miRNA evolution.

We further investigated patterns of sequence evolution for the subset of 38 conserved pre-miRNAs with mature miRNA sequences at their 3′ end by calculating evolutionary rates in distinct site classes (Fig. 6, and Supplementary Information section 12.2). Outside the mature miRNA and its complementary sequence, loops had the highest rate of evolution, followed by unpaired sites, with paired sites having the lowest rate of evolution. Inside the mature miRNA, unpaired sites evolve more slowly than paired sites, whereas the opposite is true for the sequence complementary to the mature miRNA. Surprisingly, a large fraction of unpaired bulges or internal loops in the mature miRNA seem to be conserved—a pattern which may have implications for models of miRNA biogenesis and the degree of mismatch allowed in miRNA–target prediction methods. Overall these results support the qualitative model proposed in ref. 141 for the canonical progression of miRNA evolution, and show that functional constraints on the miRNA itself supersede structural constraints imposed by maintenance of the hairpin-loop.

To assess constraint on stem regions of RNA structures more generally, we compared substitution rates in stems (*S*) to those in nominally unconstrained loop regions (*L*) in a wide variety of ncRNAs (Supplementary Information section 12.3). We estimated substitution rates using a maximum likelihood framework, and compared the observed *L*/*S* ratio with the average *L*/*S* ratio estimated from published secondary structures in RFAM, which we normalized to 1.0. *L*/*S* ratios for *Drosophila* ncRNA families range from a highly constrained 2.57 for the nuclear RNase P family to 0.56 for the 5S ribosomal RNA (Supplementary Table 15).

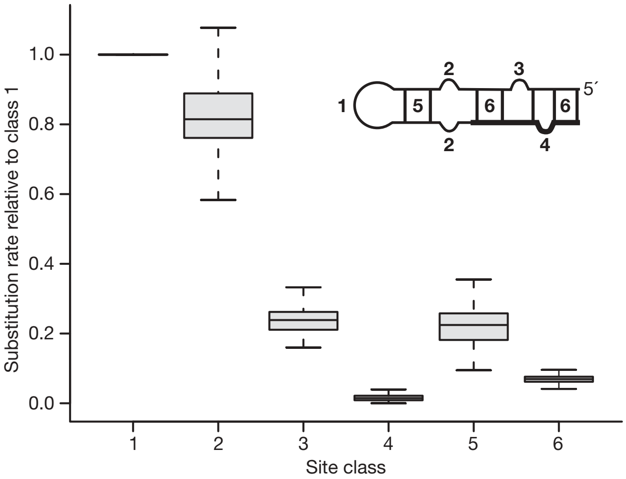


Figure 6 | Substitution rate of site classes within miRNAs. Bootstrap distributions of miRNA substitution rates. Structural alignments of miRNA precursor hairpins were partitioned into six site-classes (inset): (1) hairpin loops; unpaired sites (2) outside, (3) in the complementary region of, and (4) inside the miRNA; and base pairs (5) adjacent to and (6) involving the miRNA. Whiskers show approximate 95% confidence intervals for median differences, boxes show interquartile range.

Finally, we predicted a set of conserved intronic RNA structures and analysed patterns of compensatory nucleotide substitution in *D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis* (Supplementary Information section 13). Signatures of compensatory evolution in RNA helices are detected as covarying nucleotide sites or 'covariations' (that is, two Watson–Crick bases that interact in species A replaced by a different Watson–Crick pair in species B). The number of covariations (per base pair of a helix) depends on the physical distance between the interacting nucleotides (Supplementary Fig. 9), as has been observed for the RNA helices in the *Drosophila bicoid* 3′ UTR region[142]. Short-range pairings exhibit a higher average number of covariations with a larger variance among helices than longer-range pairings. The decrease in rate of covariation with increasing distance may be explained by physical properties of a helix, which may impose selective constraints on the evolution of covarying nucleotides within a helix. Alternatively, if individual mutations at each locus are deleterious but compensated by mutations at a second locus, given sufficiently strong selection against the first deleterious mutation these epistatic fitness interactions could generate the observed distance effect[143].

Evolution of *cis*-regulatory DNAs. Comparative analyses of *cis*-regulatory sequences may provide insights into the evolutionary forces acting on regulatory components of genes, shed light on the constraints of the *cis*-regulatory code and aid in annotation of new regulatory sequences. Here we rely on two recently compiled databases, and present results comparing *cis*-regulatory modules[144] and transcription factor binding sites (derived from DNase I footprints)[145] between *D. melanogaster* and *D. simulans* (Supplementary Information section 8). We estimated mean selective constraint (*C*, the fraction of mutations removed by natural selection) relative to the 'fastest evolving intron' sites at the 5′ end of short introns, which represent putatively unconstrained neutral standards (Supplementary Information section 8.2)[146]. Note that this approach ignores the contribution of positively selected sites, potentially underestimating the fraction of functionally relevant sites[147].

Consistent with previous findings, *Drosophila cis*-regulatory sequences are highly constrained[148,149]. Mean constraint within *cis*-regulatory modules is 0.643 (95% bootstrap confidence interval = 0.621–0.662) and within footprints is 0.692 (0.655–0.723), both of which are significantly higher than mean constraint in non-coding DNA overall (0.555 (0.546–0.563)) and significantly lower than constraint at non-degenerate coding sites (0.862 (0.856–0.868)) and ncRNA genes (0.864 (0.846–0.880)) (Supplementary Fig. 10). The high level of constraint in *cis*-regulatory sequences also extends into flanking sequences, only declining to constraint levels typical of non-coding DNA 40 bp away. This is consistent with previous findings that transcription factor binding sites tend to be found in larger blocks of constraint that cluster to form *cis*-regulatory modules[150]. To understand selective constraints on nucleotides within *cis*-regulatory sequences that have direct contact with transcription factors, we estimated the selective constraint for the best match to position weight matrices within each footprint[151]; core motifs in transcription-factor-binding sites have a mean constraint of 0.773 (0.729–0.814), significantly greater than the mean for the footprints as a whole, and approaching the level of constraint found at non-degenerate coding sites and in ncRNA genes (Supplementary Fig. 10).

We next examined the variation in selective constraint across *cis*-regulatory sequences. Surprisingly, we find no evidence that selective constraint is correlated with predicted transcription-factor-binding strength (estimated as the position weight matrix score *P*-value) (Spearman's $r = 0.0681$, $P = 0.0609$). We observe significant variation in constraint both among target genes (Kruskal–Wallis tests, footprints, $P < 0.0001$; and position weight matrix matches within footprints, $P = 0.0023$) and among chromosomes (*cis*-regulatory modules, $P = 0.0186$; footprints, $P = 0.0388$; and position weight matrix matches within footprints, $P = 0.0108$; Supplementary Table 16).

Discussion and conclusion

Each new genome sequence affords novel opportunities for comparative genomic inference. What makes the analysis of these 12 *Drosophila* genomes special is the ability to place every one of these genomic comparisons on a phylogeny with a taxon separation that is ideal for asking a wealth of questions about evolutionary patterns and processes. It is without question that this phylogenomic approach places additional burdens on bioinformatics efforts, multiplying the amount of data many-fold, requiring extra care in generating multi-species alignments, and accommodating the reality that not all genome sequences have the same degree of sequencing or assembly accuracy. These difficulties notwithstanding, phylogenomics has extraordinary advantages not only for the analyses that are possible, but also for the ability to produce high-quality assemblies and accurate annotations of functional features in a genome by using closely related genomes as guides. The use of multi-species orthology provides especially convincing evidence in support of particular gene models, not only for protein-coding genes, but also for miRNA and other ncRNA genes.

Many attributes of the genomes of *Drosophila* are remarkably conserved across species. Overall genome size, number of genes, distribution of transposable element classes, and patterns of codon usage are all very similar across these 12 genomes, although *D. willistoni* is an exceptional outlier by several criteria, including its unusually skewed codon usage, increased transposable element content and potential lack of selenoproteins. At a finer scale, the number of structural changes and rearrangements is much larger; for example, there are several different rearrangements of genes in the *Hox* cluster found in these *Drosophila* species.

The vast majority of multigene families are found in all 12 genomes, although gene family size seems to be highly dynamic: almost half of all gene families change in size on at least one lineage, and a noticeable fraction shows rapid and lineage-specific expansions and contractions. Particularly notable are cases consistent with adaptive hypotheses, such as the loss of *Gr* genes in ecological specialists and the lineage-specific expansions of antimicrobial peptides and other immune effectors. All species were found to have novel genes not seen in other species. Although lineage-specific genes are challenging to verify computationally, we can confirm at least 44 protein-coding genes unique to the *melanogaster* group, and these proteins have very different properties from ancestral proteins. Similarly, although the relative abundance of transposable element subclasses across these genomes does not differ dramatically, total genomic transposable element content varies substantially among species, and several instances of lineage-specific transposable elements were discovered.

There is considerable variation among protein-coding genes in rates of evolution and patterns of positive selection. Functionally similar proteins tend to evolve at similar rates, although variation in genomic features such as gene expression level, as well as chromosomal location, are also associated with variation in evolutionary rate among proteins. Whereas broad functional classes do not seem to share patterns of positive selection, and although very few GO categories show excesses of positive selection, a number of genes involved in interactions with the environment and in sex and reproduction do show signatures of adaptive evolution. It thus seems likely that adaptation to changing environments, as well as sexual selection, shape the evolution of protein-coding genes.

Annotation of ncRNA genes across all 12 species allows comprehensive analysis of the evolutionary divergence of these genes. MicroRNA genes in particular are more conserved than protein-coding genes with respect to their primary DNA sequence, and the substitutions that do occur often have compensatory changes such that the average estimated free energy of the folding structures remains remarkably constant across the phylogeny. Surprisingly, mismatches in miRNAs seem to be highly conserved, which may impact models of miRNA biogenesis and target recognition. Lineage-restricted miRNAs, however, have considerably elevated rates of change, suggesting either reduced constraint due to novel miRNAs having fewer targets, or adaptive evolution of evolutionarily young miRNAs.

Virtually any question about the function of genome features in *Drosophila* is now empowered by being embedded in the context of this 12 species phylogeny, allowing an analysis of the ways by which evolution has tuned myriad biological processes across the hundreds of millions of years spanned in total by this phylogeny. The analyses presented herein have generated more questions than they have answered, and these results represent a small fraction of that which is possible. Because much of this rich and extraordinary comparative genomic dataset remains to be explored, we believe that these 12 *Drosophila* genome sequences will serve as a powerful tool for gleaning further insight into genetic, developmental, regulatory and evolutionary processes.

METHODS

The full methods for this paper are described in Supplementary Information. Here, we describe the datasets generated by this project and their availability.

Genomic sequence. Scaffolds and assemblies for all genomic sequence generated by this project are available from GenBank (Supplementary Tables 4 and 5), and FlyBase (ftp://ftp.flybase.net/12_species_analysis/). Genome browsers are available from UCSC (http://genome.ucsc.edu/cgi-bin/hgGateway?hgsid = 98180333&clade = insect&org = 0&db = 0) and Flybase (http://flybase.org/cgi-bin/gbrowse/dmel/). BLAST search of these genomes is available at FlyBase (http://flybase.org/blast).

Predicted gene models. Consensus gene predictions for the 11 non-*melanogaster* species, produced by combining several different GLEAN runs that weight homology evidence more or less strongly, are available from FlyBase as GFF files for each species (ftp://ftp.flybase.net/12_species_analysis/). These gene models can also be accessed from the Genome Browser in FlyBase (Gbrowse; http://flybase.org/cgi-bin/gbrowse/dmel/). Predictions of non-protein-coding genes are also available in GFF format for each species, from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Homology. Multiway homology assignments are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/), and also in the Genome Browser (Gbrowse).

Alignments. All alignment sets produced are available in FASTA format from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

PAML parameters. Output from PAML models for the alignments of single copy orthologues in the *melanogaster* group, including the *q*-value for the test for positive selection, are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

Received 19 July; accepted 5 October 2007.

1. Markow, T. A. & O'Grady, P. M. Drosophila biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
2. Powell, J. R. *Progress and Prospects in Evolutionary Biology: The* Drosophila *Model* (Oxford Univ. Press, Oxford, 1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287**, 2185–2195 (2000).
4. Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3**, research0079.1–0079.14 (2002).
5. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15**, 1–18 (2005).
6. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila. Science* **287**, 2196–2204 (2000).
7. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
8. Stark et al.. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* doi:10.1038/nature06340 (this issue).
9. Begun, D. J. *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans. PLoS Biol.* **5**, e310, doi:10.1371/journal.pbio.0050310 (2007).
10. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* (in the press).
11. Clary, D. O. & Wolstenholme, D. R. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* **22**, 252–271 (1985).
12. Ballard, J. W. When one is not enough: introgression of mitochondrial DNA in *Drosophila. Mol. Biol. Evol.* **17**, 1126–1130 (2000).
13. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of *Drosophila. Mol. Biol. Evol.* (submitted).
14. Salzberg, S. *et al.* Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biol.* **6**, R23 (2005).
15. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152–i158 (2005).
16. Smith, C. D. *et al.* Improved repeat identification and masking in Dipterans. *Gene* **389**, 1–9 (2007).
17. Li, Q. *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PloS Comput. Biol.* **1**, e43 (2005).
18. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol.* **7**, R112 (2006).
19. Guigo, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *J. Mol. Biol.* **226**, 141–157 (1992).
20. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
21. Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13**, 379–393 (2006).
22. Gross, S. S., Do, C. B. & Batzoglou, S. in *BCATS 2005 Symposium Proc.* 82 (2005).
23. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
24. Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
25. Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biol.* **7**, R29 (2006).
26. Souvorov, A. *et al.* in *NCBI News Fall/Winter, NIH Publication No. 04-3272* (eds Benson, D & Wheeler, D) (2006).
27. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949 (2006).
28. Elsik, C. G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
29. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* doi:10.1038/nature06323 (this issue).
30. Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of *Drosophila melanogaster. Nature Genet.* **38**, 1151–1158 (2006).
31. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
32. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics* **17**, 152–161 (2006).
33. Heger, A. & Ponting, C. Evolutionary rate analyses of orthologues and paralogues from twelve *Drosophila* genomes. doi:10.1101/gr6249707 *Genome Res*. (in the press).
34. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
35. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
36. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
37. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P. & Gerstein, M. Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res.* **31**, 1033–1037 (2003).
38. Bosco, G., Campbell, P., Leiva-Neto, J. & Markow, T. Analysis of *Drosophila* species genome size and satellite DNA content reveals significant differences among strains as well as between species. *Genetics* doi:10.1534/Genetics107.075069 (in the press).
39. Ranz, J. *et al.* Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol.* **5**, e152, doi:10.1371/journal.pbio.0050152 (2007).
40. Noor, M. A. F., Garfield, D. A., Schaeffer, S. W. & Machado, C. A. Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* doi:10.1534/genetics.107.070672 (in the press).
41. Lewis, E. B. A gene complex controlling segmentation in *Drosophila. Nature* **276**, 565–570 (1978).
42. Negre, B., Ranz, J. M., Casals, F., Caceres, M. & Ruiz, A. A new split of the *Hox* gene complex in *Drosophila*: relocation and evolution of the gene labial. *Mol. Biol. Evol.* **20**, 2042–2054 (2003).
43. Von Allmen, G. *et al.* Splits in fruitfly *Hox* gene complexes. *Nature* **380**, 116 (1996).
44. Negre, B. & Ruiz, A. HOM-C evolution in *Drosophila*: is there a need for *Hox* gene clustering? *Trends Genet.* **23**, 55–59 (2007).
45. Dowsett, A. P. & Young, M. W. Differing levels of dispersed repetitive DNA among closely related species of *Drosophila. Proc. Natl Acad. Sci.* **79**, 4570–4574 (1982).
46. Kapitonov, V. V. & Jurka, J. DNAREP1_DM. (Repbase Update Release 3.4, 1999).
47. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).

48. Singh, N. D., Arndt, P. F. & Petrov, D. A. Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster. Genetics* **169,** 709–722 (2004).

49. Yang, H.-P., Hung, T.-L., You, T.-L. & Yang, T.-H. Genomewide comparative analysis of the highly abundant transposable element *DINE-1* suggests a recent transpositional burst in *Drosophila yakuba*. *Genetics* **173,** 189–196 (2006).

50. Yang, H.-P. & Barbash, D. Abundant and species-specific miniature inverted-repeat transposable elements in 12 *Drosophila* genomes. *Genome Biol.* (submitted).

51. Wilder, J. & Hollocher, H. Mobile elements and the genesis of microsatellites in dipterans. *Mol. Biol. Evol.* **18,** 384–392 (2001).

52. Marzo, M., Puig, M. & Ruiz, A. The foldback-like element *Galileo* belongs to the P superfamily of DNA transposons and is widespread within the genus *Drosophila. Proc. Natl Acad. Sci. USA* (submitted).

53. Casola, C., Lawing, A., Betran, E. & Feschotte, C. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* **24,** 1872–1888 (2007).

54. Abad, J. P. *et al.* Genomic analysis of *Drosophila melanogaster* telomeres: full-length copies of *HeT-A* and *TART* elements at telomeres. *Mol. Biol. Evol.* **21,** 1613–1619 (2004).

55. Abad, J. P. *et al. TAHRE*, a novel telomeric retrotransposon from *Drosophila melanogaster*, reveals the origin of *Drosophila* telomeres. *Mol. Biol. Evol.* **21,** 1620–1624 (2004).

56. Blackburn, E. H. Telomerases. *Annu. Rev. Biochem.* **61,** 113–129 (1992).

57. Villasante, A. *et al. Drosophila* telomeric retrotransposons derived from an ancestral element that as recruited to replace telomerase. *Genome Res.* (in the press).

58. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).

59. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

60. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).

61. Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K. & Makalowski, W. Spliceosomal small nuclear RNA genes in 11 insect genomes. *RNA* **13,** 5–14 (2007).

62. Schneider, C., Will, C. L., Brosius, J., Frilander, M. J. & Luhrmann, R. Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in *Drosophila. Proc. Natl Acad. Sci. USA* **101,** 9584–9589 (2004).

63. Deng, X. & Meller, V. H. Non-coding RNA in fly dosage compensation. *Trends Biochem. Sci.* **31,** 526–532 (2006).

64. Amrien, H. & Axel, R. Genes expressed in neurons of adult male *Drosophila. Cell* **88,** 459–469 (1997).

65. Park, S.-W. *et al.* An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* (in the press).

66. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of 12 *Drosophila* species. *Genome Res.* (in the press).

67. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15,** 1153–1160 (2005).

68. Hahn, M. W., Han, M. V. & Han, S.-G. Gene family evolution across 12 *Drosophila* genomes. *PLoS Biol.* **3,** e197 (2007).

69. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103,** 9935–9939 (2006).

70. Ponce, R. & Hartl, D. L. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster. Gene* **376,** 174–183 (2006).

71. Arguello, J. R., Chen, Y., Tang, S., Wang, W. & Long, M. Originiation of an X-linked testes chimeric gene by illegitimate recombination in *Drosophila. PLoS Genet.* **2,** e77 (2006).

72. Begun, D. J., Lindfore, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172,** 1675–1681 (2006).

73. Betran, E., Thornton, K. & Long, M. Retroposed new genes out of the X in *Drosophila. Genome Res.* **12,** 1854–1859 (2002).

74. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21,** 650–659 (2005).

75. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).

76. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).

77. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64,** 479–498 (2002).

78. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila. Trends Genet.* (submitted).

79. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449 (2000).

80. Bergman, C. M. *et al.* Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3,** research0086.1–0086.20 (2002).

81. Bierne, N. & Eyre Walker, A. C. The genomic rate of adaptive amino acid substitution in *Drosophila. Mol. Biol. Evol.* **21,** 1350–1360 (2004).

82. Sawyer, S. A., Kulathinal, R. J., Bustamante, C. D. & Hartl, D. L. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.* **57** (suppl. 1), S154–S164 (2003).

83. Sawyer, S. A., Parsch, J., Zhang, Z. & Hartl, D. L. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila. Proc. Natl Acad. Sci. USA* **104,** 6504–6510 (2007).

84. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila. Nature* **415,** 1022–1024 (2002).

85. Welch, J. J. Estimating the genomewide rate of adaptive protein evolution in *Drosophila. Genetics* **173,** 821–837 (2006).

86. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102,** 14338–14343 (2005).

87. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23,** 327–337 (2006).

88. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158,** 927–931 (2001).

89. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7,** 337–348 (2006).

90. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA* **102,** 5483–5488 (2005).

91. Rocha, E. P. The quest for the universals of protein evolution. *Trends Genet.* **22,** 412–416 (2006).

92. Huntley, M. A. & Clark, A. G. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol. Biol. Evol.* (in the press).

93. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130,** 113–146 (1987).

94. Larsson, J. & Meller, V. H. Dosage compensation, the origin and the afterlife of sex chromosomes. *Chromosome Res.* **14,** 417–431 (2006).

95. Riddle, N. C. & Elgin, S. C. The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. *Chromosome Res.* **14,** 405–416 (2006).

96. Gordo, I. & Charlesworth, B. Genetic linkage and molecular evolution. *Curr. Biol.* **11,** R684–R686 (2001).

97. Singh, N. D., Larracuente, A. M. & Clark, A. G. Contrasting the efficacy of selection on the X and autosomes in *Drosophila. Mol. Biol. Evol.* (submitted).

98. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome scale analysis of positionally relocated genes. *Genome Res.* (in the press).

99. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Curr. Opin. Genet. Dev.* **8,** 688–693 (1998).

100. Akashi, H., Kliman, R. M. & Eyre-Walker, A. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila. Genetica (Dordrecht)* **102–103,** 49–60 (1998).

101. Bulmer, M. The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129,** 897–908 (1991).

102. McVean, G. A. T. & Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. *Genet. Res.* **74,** 145–158 (1999).

103. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* **24,** 28–38 (1986).

104. Akashi, H. & Schaeffer, S. W. Natural selection and the frequency distributions of ''silent'' DNA polymorphism in *Drosophila. Genetics* **146,** 295–307 (1997).

105. Powell, J. R., Sezzi, E., Moriyama, E. N., Gleason, J. M. & Caccone, A. Analysis of a shift in codon usage in *Drosophila. J. Mol. Evol.* **57,** S214–S225 (2003).

106. Anderson, C. L., Carew, E. A. & Powell, J. R. Evolution of the *Adh* locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. *Mol. Biol. Evol.* **10,** 605–618 (1993).

107. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* **153,** 339–350 (1999).

108. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Evidence for a high ancestral GC content in *Drosophila. Mol. Biol. Evol.* **17,** 1710–1717 (2000).

109. Rodriguez-Trelles, F., Tarrio, R. & Ayala, F. J. Fluctuating mutation bias and the evolution of base composition in *Drosophila. J. Mol. Evol.* **50,** 1–10 (2000).

110. Heger, A. & Ponting, C. Variable strength of translational selection among twelve *Drosophila* species. *Genetics* (in the press).

111. Vicario, S., Moriyama, E. N. & Powell, J. R. Codon Usage in Twelve Species of *Drosophila. BMC Evol. Biol.* (submitted).

112. Singh, N. D., Arndt, P. F. & Petrov, D. A. Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. *BMC Biol.* 4, 10.1186/1741-7007-4-37 (2006).

113. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).

114. Akashi, H. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster. Genetics* **144,** 1297–1307 (1996).

115. Akashi, H. *et al.* Molecular evolution in the *Drosophila melanogaster* species subgroup: Frequent parameter fluctuations on the timescale of molecular divergence. *Genetics* **172,** 1711–1726 (2006).
116. Bauer DuMont, V., Fay, J. C., Calabrese, P. P. & Aquadro, C. F. DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167,** 171–185 (2004).
117. McVean, G. A. & Vieira, J. The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J. Mol. Evol.* **49,** 63–75 (1999).
118. Nielsen, R., Bauer DuMont, V., Hubisz, M. J. & Aquadro, C. F. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila. Mol. Biol. Evol.* **24,** 228–235 (2007).
119. Begun, D. J. The frequency distribution of nucleotide variation in *Drosophila* simulans. *Mol. Biol. Evol.* **18,** 1343–1352 (2001).
120. Singh, N. S., Bauer DuMont, V. L., Hubisz, M. J., Nielsen, R. & Aquadro, C. F. Patterns of mutation and selection at synonymous sites in *Drosophila. Mol. Biol. Evol.* doi:10.1093/mobev.msm196 (in the press).
121. McBride, C. S. & Arguello, J. R. Five *Drosophila* genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics* (in the press).
122. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: Purifying selection and birth-and-death evolution. *Genome Biol.* **8,** 235 (2007).
123. Gardiner, A., Barker, D., Butilne, R. K., Jordan, W. C. & Ritchie, M. G. *Drosophila* chemoreceptor evolution: Selection, specialisation and genome size. *Genome Biol.* (submitted).
124. McBride, C. S. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc. Natl Acad. Sci. USA* **104,** 4996–5001 (2007).
125. Ranson, H. *et al.* Evolution of supergene families associated with insecticide resistance. *Science* **298,** 179–181 (2002).
126. Tijet, N., Helvig, C. & Feyereisen, R. The cytochrome P450 gene superfamily in *Drosophila melanogaster. Gene* **262,** 189–198 (2001).
127. Claudianos, C. *et al.* A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15,** 615–636 (2006).
128. Low, W. L. *et al.* Molecular evolution of glutathione *S*-transferases in the genus *Drosophila. Genetics* (in the press).
129. Castellano, S. *et al. In silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep.* **2,** 697–702 (2001).
130. Martin-Romero, F. J. *et al.* Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J. Biol. Chem.* **276,** 29798–29804 (2001).
131. Lemaitre, B. & Hoffmann, J. The host defense of *Drosophila melanogaster. Annu. Rev. Immunol.* **25,** 697–743 (2007).
132. Hughes, A. L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335,** 167–170 (1988).
133. Murphy, P. M. Molecular mimicry and the generation of host defense protein diversity. *Cell* **72,** 823–826 (1993).
134. Schlenke, T. A. & Begun, D. J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164,** 1471–1480 (2003).
135. Sackton, T. B. *et al.* The evolution of the innate immune system across *Drosophila. Nature Genet.* (submitted).
136. Civetta, A. & Singh, R. S. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41,** 1085–1095 (1995).
137. Civetta, A. Shall we dance or shall we fight? Using DNA sequence data to untangel controversies surrounding sexual selection. *Genome* **46,** 925–929 (2003).
138. Clark, N. L., Aagard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. *Reproduction* **131,** 11–22 (2006).
139. Haerty, W. *et al.* Evolution in the fast lane: rapidly evolving sex- and reproduction-related genes in *Drosophila* species. *Genetics* (in the press).
140. Lu, J. *et al.* Adaptive evolution of newly-emerged microRNA genes in *Drosophila. Mol. Biol. Evol.* (submitted).
141. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4,** R42 (2003).
142. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154,** 909–921 (2000).
143. Stephan, W. The rate of compensatory evolution. *Genetics* **144,** 419–426 (1996).
144. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for *Drosophila. Bioinformatics* **22,** 381–383 (2006).
145. Bergman, C. M., Carlson, J. W. & Celniker, S. E. *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster. Bioinformatics* **21,** 1747–1749 (2005).
146. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* **16,** 875–884 (2006).
147. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila. Nature* **437,** 1149–1152 (2005).
148. Bird, C. P., Stranger, B. E. & Dermitzakis, E. T. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.* **16,** 559–564 (2006).
149. Wittkopp, P. J. Evolution of *cis*-regulatory sequence and function in Diptera. *Heredity* **97,** 139–147 (2006).
150. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of *eve stripe 2* enhancer evolution in *Drosophila. Development* **125,** 949–958 (1998).
151. Down, A. T. A., Bergman, C. M., Su, J. & Hubbard, T. J. P. Large scale discovery of promoter motifs in *Drosophila melanogaster. PloS Comput. Biol.* **3,** e7 (2007).
152. Tamura, K., Subramanian, S. & Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **21,** 36–44 (2004).
153. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5,** 150–163 (2004).
154. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* **2,** e173 (2006).
155. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A *Drosophila* case study. *Genome Biol.* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Agencourt Bioscience Corporation, The Broad Institute of MIT and Harvard and the Washington University Genome Sequencing Center were supported by grants and contracts from the National Human Genome Research Insititute (NHGRI). T.C. Kaufman acknowledges support from the Indian Genomics Initiative.

Author Contributions The laboratory groups of A. G. Clark (including A. M. Larracuente, T. B. Sackton, and N. D. Singh) and Michael B. Eisen (including V. N. Iyer and D. A. Pollard) played the part of coordinating the primary writing and editing of the manuscript with the considerable help of D. R. Smith, C. M. Bergman, W. M. Gelbart, B. Oliver, T. A. Markow, T. C. Kaufman and M. Kellis. D. R. Smith served as primary coordinator for the assemblies. The remaining authors contributed either through their efforts in sequence production, assembly and annotation, or in the analysis of specific topics that served as the focus of more than 40 companion papers.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.C. (ac347@cornell.edu), M.B.E. (mbeisen@lbl.gov), D.R.S. (douglas.smith@agencourt.com), C.M.B. (casey.bergman@manchester.ac.uk), W.G. (gelbart@morgan.harvard.edu), B.O. (oliver@helix.nih.gov), T.A.M. (tmarkow@public.arl.arizona.edu), T.C.K. (kaufman@indiana.edu), M.K. (manoli@mit.edu), V.N.I. (venky@berkeley.edu), T.B.S. (tbs7@cornell.edu), A.M.L. (aml69@cornell.edu), D.A.P. (danielapollard@alum.bowdoin.edu), N.D.S. (nds25@cornell.edu), or collectively to 12flies@morgan.harvard.edu.

***Drosophila* 12 Genomes Consortium**

Project Leaders Andrew G. Clark[1], Michael B. Eisen[2,3], Douglas R. Smith[4], Casey M. Bergman[5], Brian Oliver[6], Therese A. Markow[7], Thomas C. Kaufman[8], Manolis Kellis[9,10] & William Gelbart[11,12]

Annotation Coordination Venky N. Iyer[13] & Daniel A. Pollard[14]

Analysis/Writing Coordination Timothy B. Sackton[1,15], Amanda M. Larracuente[1] & Nadia D. Singh[1]

Sequencing, Assembly, Annotation and Analysis Contributors Jose P. Abad[16], Dawn N. Abt[17], Boris Adryan[18], Montserrat Aguade[19], Hiroshi Akashi[20], Wyatt W. Anderson[21], Charles F. Aquadro[1], David H. Ardell[22], Roman Arguello[23], Carlo G. Artieri[24], Daniel A. Barbash[1], Daniel Barker[25], Paolo Barsanti[26], Phil Batterham[27], Serafim Batzoglou[28], Dave Begun[29], Arjun Bhutkar[11,30], Enrico Blanco[31], Stephanie A. Bosak[4], Robert K. Bradley[32], Adrianne D. Brand[4], Michael R. Brent[33], Angela N. Brooks[13], Randall H. Brown[33], Roger K. Butlin[34], Corrado Caggese[26], Brian R. Calvi[35], A. Bernardo de Carvalho[36], Anat Caspi[32], Sergio Castrezana[37], Susan E. Celniker[2], Jean L. Chang[10], Charles Chapple[31], Sourav Chatterji[38,39], Asif Chinwalla[40], Alberto Civetta[41], Sandra W. Clifton[40], Josep M. Comeron[42], James C. Costello[43], Jerry A. Coyne[23], Jennifer Daub[44], Robert G. David[4], Arthur L. Delcher[45], Kim Delehaunty[40], Chuong B. Do[28], Heather Ebling[4], Kevin Edwards[46], Thomas Eickbush[47], Jay D. Evans[48], Alan Filipski[49], Sven Findeiß[49,50], Eva Freyhult[22], Lucinda Fulton[40], Robert Fulton[40], Ana C. L. Garcia[51], Anastasia Gardiner[25], David A. Garfield[52], Barry E. Garvin[4], Greg Gibson[53], Don Gilbert[8], Sante Gnerre[10], Jennifer Godfrey[40], Robert Good[27], Valer Gotea[20], Brenton Gravely[54], Anthony J. Greenberg[1], Sam Griffiths-Jones[5,44], Samuel Gross[28], Roderic Guigo[31,55], Erik A. Gustafson[4], Wilfried Haerty[24], Matthew W. Hahn[8,43], Daniel L. Halligan[56], Aaron L. Halpern[57], Gillian M. Halter[20], Mira V. Han[43], Andreas Heger[58,59], LaDeana Hillier[40], Angie S. Hinrichs[60], Ian Holmes[32], Roger A. Hoskins[2], Melissa J. Hubisz[61], Dan Hultmark[62], Melanie A. Huntley[1], David B. Jaffe[10], Santosh Jagadeeshan[24], William R. Jeck[63], Justin Johnson[57], Corbin D. Jones[63], William C. Jordan[64], Gary H. Karpen[13,65], Eiko Kataoka[66], Peter D. Keightley[56], Pouya Kheradpour[9], Ewen F. Kirkness[57], Leonardo B. Koerich[36], Karsten Kristiansen[67], Dave

Kudrna[68], Rob J. Kulathinal[69], Sudhir Kumar[49,70], Roberta Kwok[8], Eric Lander[10], Charles H. Langley[29], Richard Lapoint[71], Brian P. Lazzaro[72], So-Jeong Lee[68], Lisa Levesque[41], Ruiqiang Li[67,73], Chiao-Feng Lin[20], Michael F. Lin[9,10], Kerstin Lindblad-Toh[10], Ana Llopart[42], Manyuan Long[23], Lloyd Low[27], Elena Lozovsky[69], Jian Lu[23], Meizhong Luo[68], Carlos A. Machado[7], Wojciech Makalowski[20], Mar Marzo[74], Muneo Matsuda[66], Luciano Matzkin[7], Bryant McAllister[42], Carolyn S. McBride[29], Brendan McKernan[4], Kevin McKernan[4], Maria Mendez-Lago[16], Patrick Minx[40], Michael U. Mollenhauer[20], Kristi Montooth[17], Stephen M. Mount[45,75], Xu Mu[20], Eugene Myers[76], Barbara Negre[77], Stuart Newfeld[70], Rasmus Nielsen[78], Mohamed A. F. Noor[52], Patrick O'Grady[71], Lior Pachter[38], Montserrat Papaceit[19], Matthew J. Parisi[4], Michael Parisi[6], Leopold Parts[9], Jakob S. Pedersen[60,79], Graziano Pesole[80], Adam M. Phillippy[45], Chris P. Ponting[58,59], Mihai Pop[45], Damiano Porcelli[26], Jeffrey R. Powell[81], Sonja Prohaska[49,82], Kim Pruitt[83], Marta Puig[74], Hadi Quesneville[84], Kristipati Ravi Ram[1], David Rand[17], Matthew D. Rasmussen[9], Laura K. Reed[53], Robert Reenan[85], Amy Reily[40], Karin A. Remington[57], Tania T. Rieger[86], Michael G. Ritchie[25], Charles Robin[27], Yu-Hui Rogers[57], Claudia Rohde[87], Julio Rozas[19], Marc J. Rubenfield[4], Alfredo Ruiz[74], Susan Russo[11,12], Steven L. Salzberg[45], Alejandro Sanchez-Gracia[19,88], David J. Saranga[4], Hajime Sato[66], Stephen W. Schaeffer[20], Michael C. Schatz[45], Todd Schlenke[89], Russell Schwartz[90], Carmen Segarra[19], Rama S. Singh[24], Laura Sirot[1], Marina Sirota[91], Nicholas B. Sisneros[68], Chris D. Smith[65,92], Temple F. Smith[30], John Spieth[40], Deborah E. Stage[47], Alexander Stark[9,10], Wolfgang Stephan[93], Robert L. Strausberg[57], Sebastian Strempel[93], David Sturgill[6], Granger Sutton[57], Granger G. Sutton[57], Wei Tao[4], Sarah Teichmann[18], Yoshiko N. Tobari[94], Yoshihiko Tomimura[95], Jason M. Tsolas[4], Vera L. S. Valente[51], Eli Venter[57], J. Craig Venter[57], Saverio Vicario[81], Filipe G. Vieira[19], Albert J. Vilella[19,96], Alfredo Villasante[16], Brian Walenz[57], Jun Wang[67,73], Marvin Wasserman[97], Thomas Watts[7], Derek Wilson[18], Richard K. Wilson[40], Rod A. Wing[68], Mariana F. Wolfner[1], Alex Wong[1], Gane Ka-Shu Wong[73,98], Chung-I Wu[23], Gabriel Wu[32], Daisuke Yamamoto[99], Hsiao-Pei Yang[1], Shiaw-Pyng Yang[40], James A. Yorke[100], Kiyohito Yoshida[101], Evgeny Zdobnov[102], Peili Zhang[11,12], Yu Zhang[6], Aleksey V. Zimin[100], Broad Institute Genome Sequencing Platform* & Broad Institute Whole Genome Assembly Team*

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [3]Center for Integrative Genomics, Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [4]Agencourt Bioscience Corporation, Beverly, Massachusetts 01915, USA. [5]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [6]Laboratory of Cellular and Developmental Biology, National Institutes of Health, Bethesda, Maryland 20892, USA. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [8]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [9]Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA. [10]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [11]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [12]FlyBase, The Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA. [14]Biophysics Graduate Group, University of California at Berkeley, Berkeley, California 94720, USA. [15]Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [16]Centro de Biología Molecular Severo Ochoa, Universidad Autónoma de Madrid, Madrid 28049, Spain. [17]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island 02912, USA. [18]Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK. [19]Departament de Genètica, Universitat de Barcelona, Barcelona 08071, Spain. [20]Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [21]Department of Genetics, University of Georgia, Athens, Georgia 30602, USA. [22]Linnaeus Centre for Bioinformatics, Uppsala Universitet, Uppsala, SE-75124, Sweden. [23]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA. [24]Department of Biology, McMaster University, Hamilton, Ontario, L8S 4K1, Canada. [25]School of Biology, University of St. Andrews, Fife KY16 9TH, UK. [26]Dipartimento di Genetica e Microbiologia dell'Università di Bari, Bari, 70126, Italy. [27]Department of Genetics, University of Melbourne, Melbourne 3010, Australia. [28]Computer Science Department, Stanford University, Stanford, California 94305, USA. [29]Section of Evolution and Ecology and Center for Population Biology, University of California at Davis, Davis, California 95616, USA. [30]BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts 02215, USA. [31]Research Group in Biomedical Informatics, Institut Municipal d'Investigacio Medica, Universitat Pompeu Fabra, Barcelona 08003, Catalonia, Spain. [32]Department of Bioengineering, University of California at Berkeley, Berkeley, California 94720, USA. [33]Laboratory for Computational Genomics, Washington University, St Louis, Missouri 63108, USA. [34]Animal and Plant Sciences, The University of Sheffield, Sheffield S10 2TN, UK. [35]Department of Biology, Syracuse University, Syracuse, New York 13244, USA. [36]Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21944-970, Brazil. [37]Tucson Stock Center, Tucson, Arizona 85721, USA. [38]Department of Mathematics, University of California at Berkeley, Berkeley, California 94720, USA. [39]Genome Center, University of California at Davis, Davis, California 95616, USA. [40]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [41]Department of Biology, University of Winnipeg, Winnipeg, Manitoba R3B 2E9, Canada. [42]Department of Biological Sciences, University of Iowa, Iowa City, Iowa 52242, USA. [43]School of Informatics, Indiana University, Bloomington, Indiana 47405, USA. [44]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [45]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA. [46]Department of Biological Sciences, Illinois State University, Normal, Illinois 61790, USA. [47]Department of Biology, University of Rochester, Rochester, New York 14627, USA. [48]Bee Research Lab, USDA-ARS, Beltsville, Maryland 20705, USA. [49]Center for Evolutionary Functional Genomics, Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA. [50]Department of Computer Science, University of Leipzig, Leipzig 04107, Germany. [51]Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS 68011, Brazil. [52]Department of Biology, Duke University, Durham, New Carolina 27708, USA. [53]Department of Genetics, North Carolina State University, Raleigh, North Caroline 27695, USA. [54]Health Center, University of Connecticut, Farmington, Connecticut 06030, USA. [55]Center of Genomic Regulation, Barcelona 8003, Catalonia, Spain. [56]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK. [57]J. Craig Venter Institute, Rockville, Maryland 20850, USA. [58]MRC Functional Genetics Unit, University of Oxford, Oxford OX1 3QX, UK. [59]Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK. [60]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California 95064, USA. [61]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA. [62]Umeå Center for Molecular Pathogenesis, Umeå University, Umeå SE-90187, Sweden. [63]Department of Biology and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [64]Institute of Zoology, Regent's Park, London NW1 4RY, UK. [65]Drosophila Heterochromatin Genome Project, Department of Genome and Computational Biology, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [66]Kyorin University, School of Medicine, Mitaka, Tokyo 181-8611, Japan. [67]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark [68]Arizona Genomics Institute, Department of Plant Sciences and BIO5, University of Arizona, Tucson, Arizona 85721, USA. [69]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [70]School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. [71]Department of Environmental Science, Policy and Management, University of California at Berkeley, Berkeley, California 94720, USA. [72]Department of Entomology, Cornell University, Ithaca, New York 14853, USA. [73]Beijing Genomics Institute at ShenZhen, ShenZhen 518083, China. [74]Departament Genètica i Microbiologia, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain. [75]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA. [76]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, Viginia 20147-2408, USA. [77]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [78]Institue of Biology, University of Copenhagen, DK-2100 Copenhagen Ø, Denmark. [79]Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark. [80]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari 70126, Italy. [81]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA. [82]Department of Biomedical Informatics, Arizona State University, Tempe, Arizona 85287, USA. [83]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [84]Bioinformatics and Genomics Laboratory, Institut Jacques Monod, Paris, 75251, France. [85]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island 02912, USA. [86]Departamento de Genética, Centro de Ciências Biológicas, Universidade Federal de Pernambuco, Recife/PE 68011, Brazil. [87]Centro Acadêmico de Vitória, Universidade Federal de Pernambuco, Vitória de Santo Antão/PE, Brazil. [88]Cajal Institute, CSIC, Madrid 28002, Spain. [89]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [90]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. [91]Biomedical Informatics, Stanford University, Stanford, California 94305, USA. [92]Department of Biology, San Francisco State University, San Francisco, California 94132, USA. [93]Department of Biology, University of Munich, 82152 Planegg-Martinsried, Germany. [94]Institute of Evolutionary Biology, Setagaya-ku, Tokyo 158-0098, Japan. [95]Shiba Gakuen, Minato-ku, Tokyo 105-0011, Japan. [96]European Bioinformatics Institute, Hinxton, CB10 1SD, UK. [97]Department of Biology, City University of New York at Queens, Flushing, New York 11367, USA. [98]Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E9, Canada. [99]Department of Developmental Biology and Neurosciences, Tohoku University, Sendai 980-8578, Japan. [100]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA. [101]Hokkaido University, EESBIO, Sapporo, Hokkaido 060-0810, Japan. [102]Faculty of Medecine, Universite de Geneve, Geneva CH-1211, Switzerland.

***Broad Institute Genome Sequencing Platform** Jennifer Baldwin[10], Amr Abdouelleil[10], Jamal Abdulkadir[10], Adal Abebe[10], Brikti Abera[10], Justin Abreu[10], St Christophe Acer[10], Lynne Aftuck[10], Allen Alexander[10], Peter An[10], Erica Anderson[10], Scott Anderson[10], Harindra Arachi[10], Marc Azer[10], Pasang Bachantsang[10], Andrew Barry[10], Tashi Bayul[10], Aaron Berlin[10], Daniel Bessette[10], Toby Bloom[10], Jason Blye[10], Leonid Boguslavskiy[10], Claude Bonnet[10], Boris Boukhgalter[10], Imane Bourzgui[10], Adam Brown[10], Patrick Cahill[10], Sheridon Channer[10], Yama Cheshatsang[10], Lisa Chuda[10], Mieke Citroen[10], Alville Collymore[10], Patrick Cooke[10], Maura Costello[10], Katie D'Aco[10], Riza Daza[10], Georgius De Haan[10], Stuart DeGray[10], Christina DeMaso[10], Norbu Dhargay[10], Kimberly Dooley[10], Erin Dooley[10], Missole Doricent[10], Passang Dorje[10], Kunsang Dorjee[10], Alan Dupes[10], Richard Elong[10], Jill Falk[10], Abderrahim Farina[10], Susan Faro[10], Diallo Ferguson[10], Sheila Fisher[10], Chelsea D. Foley[10], Alicia Franke[10], Dennis Friedrich[10], Loryn Gadbois[10], Gary Gearin[10], Christina R. Gearin[10], Georgia Giannoukos[10], Tina Goode[10], Joseph Graham[10], Edward Grandbois[10], Sharleen Grewal[10], Kunsang Gyaltsen[10], Nabil Hafez[10], Birhane Hagos[10], Jennifer Hall[10], Charlotte Henson[10], Andrew Hollinger[10], Tracey Honan[10], Monika D. Huard[10], Leanne Hughes[10], Brian Hurhula[10], M Erii Husby[10], Asha Kamat[10], Ben Kanga[10], Seva Kashin[10], Dmitry Khazanovich[10], Peter Kisner[10], Krista Lance[10], Marcia Lara[10], William Lee[10], Niall Lennon[10], Frances Letendre[10], Rosie LeVine[10], Alex Lipovsky[10], Xiaohong Liu[10], Jinlei Liu[10], Shangtao Liu[10], Tashi Lokyitsang[10], Yeshi Lokyitsang[10], Rakela Lubonja[10], Annie Lui[10], Pen MacDonald[10], Vasilia Magnisalis[10], Kebede Maru[10], Charles Matthews[10], William McCusker[10], Susan McDonough[10], Teena Mehta[10], James Meldrim[10], Louis Meneus[10], Oana Mihai[10], Atanas Mihalev[10], Tanya Mihova[10], Rachel Mittelman[10], Valentine Mlenga[10], Anna Montmayeur[10], Leonidas Mulrain[10], Adam Navidi[10], Jerome Naylor[10], Tamrat Negash[10], Thu Nguyen[10], Nga Nguyen[10], Robert Nicol[10], Choe Norbu[10], Nyima Norbu[10], Nathaniel Novod[10], Barry O'Neill[10], Sahal Osman[10], Eva Markiewicz[10], Otero L. Oyono[10], Christopher Patti[10], Pema Phunkhang[10], Fritz Pierre[10], Margaret Priest[10], Sujaa Raghuraman[10], Filip Rege[10], Rebecca Reyes[10], Cecil Rise[10], Peter Rogov[10], Keenan Ross[10], Elizabeth Ryan[10], Sampath Settipalli[10], Terry Shea[10], Ngawang Sherpa[10], Lu Shi[10], Diana Shih[10], Todd Sparrow[10], Jessica Spaulding[10], John Stalker[10], Nicole Stange-Thomann[10], Sharon Stavropoulos[10], Catherine Stone[10], Christopher Strader[10], Senait Tesfaye[10], Talene Thomson[10], Yama Thoulutsang[10], Dawa Thoulutsang[10], Kerri Topham[10], Ira Topping[10], Tsamla Tsamla[10], Helen Vassiliev[10], Andy Vo[10], Tsering Wangchuk[10], Tsering Wangdi[10], Michael Weiand[10], Jane Wilkinson[10], Adam Wilson[10], Shailendra Yadav[10], Geneva Young[10], Qing Yu[10], Lisa Zembek[10], Danni Zhong[10], Andrew Zimmer[10] & Zac Zwirko[10] **Broad Institute Whole Genome Assembly Team** David B. Jaffe[10], Pablo Alvarez[10], Will Brockman[10], Jonathan Butler[10], CheeWhye Chin[10], Sante Gnerre[10], Manfred Grabherr[10], Michael Kleber[10], Evan Mauceli[10] & Iain MacCallum[10]