Evolutionary insights from large scale resequencing datasets in *Drosophila melanogaster*

Sara Guirao-Rico and Josefa González

Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

sara.guirao@ibe.upf-csic.es
josefa.gonzalez@ibe.upf-csic.es

Corresponding author

Josefa González
ABSTRACT

*Drosophila melanogaster* has long been used as an evolutionary model system. Its small genome size, well-annotated genome, and ease of sampling, also makes it a choice species for genome resequencing studies. Hundreds of genomic samples from populations worldwide are available and are currently being used to tackle a wide range of evolutionary questions. In this review, we focused on three insights that have increased our understanding of the evolutionary history of this species, and that have implications for the study of evolutionary processes in other species as well. Because of technical limitations, most of the studies so far have focused on SNP variants. However, long-read sequencing techniques should allow us in the near future to include other type of genomic variants that also influence genome evolution.
Introduction

The transition from the genetic to the genomic era over a decade ago has been possible thanks to the advances in sequencing technologies. High-throughput sequencing has brought remarkable benefits, such as reducing the costs and time required to sequence a genome, increasing the genomic resolution to its highest point by bringing all types of mutations in all functional categories to light, and increasing the statistical power for teasing apart adaptive and neutral processes. However, it has also brought new challenges such as the need to develop bioinformatics software to deal with massive and complex data, and the need for incremental computing resources and data storage.

Whole-genome resequencing refers to one of the two categories of whole-genome sequencing, besides de novo sequencing, which is aimed at comparing genomic variability of a group of individuals from the same or from different populations. Therefore, the use of a reference genome, from the same species or a closely related one, for read mapping and variant identification, is imperative. Drosophila melanogaster has been extensively used for resequencing experiments due to its small genome size, high quality annotated genome, and the wide range of genetic tools that make this particular species a powerful evolutionary model system. Fly lines from multiple populations have been sequenced at different labs using different sequencing technologies and different starting biological material, such as haploid embryos, isofemale, and inbreed lines (Table 1).
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Location</th>
<th>Populations</th>
<th>Samples</th>
<th>Biological material</th>
<th>Sequencing platform</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DPGP3: Lack et al. [16••]</strong></td>
<td>Africa (Zambia)</td>
<td>1</td>
<td>197</td>
<td>Haploid</td>
<td>Illumina GAIIx</td>
</tr>
<tr>
<td>Kofler et al. [44]</td>
<td>Africa</td>
<td>1</td>
<td>1</td>
<td>Pooled Isofemales(^a)</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td><strong>AGES: Lack et al. [16••]</strong></td>
<td>Africa</td>
<td>12</td>
<td>53</td>
<td>Haploid</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td><strong>DPGP2: Pool et al., Langley et al. [15, 45]</strong></td>
<td>Mostly Africa</td>
<td>22</td>
<td>150</td>
<td>Mostly Haploid</td>
<td>Illumina GAIIx/Illumina GA 1s</td>
</tr>
<tr>
<td>Boitard et al. [46]</td>
<td>Europe</td>
<td>1</td>
<td>2</td>
<td>Pooled F1 wild-caught females(^b)</td>
<td>Illumina GAIIx</td>
</tr>
<tr>
<td>Kofler et al. [47]</td>
<td>Europe</td>
<td>1</td>
<td>1</td>
<td>Pooled isofemales(^c)</td>
<td>Illumina GAIIx</td>
</tr>
<tr>
<td>Kapopoulou et al. [4]</td>
<td>Europe</td>
<td>1</td>
<td>14</td>
<td>Haploid</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td>Bastide et al. [48]</td>
<td>Europe</td>
<td>2</td>
<td>16</td>
<td>Pooled F1 wild-caught females(^d)</td>
<td>Illumina GAIIx/Illumina HiSeq 2000</td>
</tr>
<tr>
<td><strong>DrosEU: Kapun et al. [5••]</strong></td>
<td>Europe</td>
<td>32</td>
<td>48</td>
<td>Pooled wild-caught males(^e)</td>
<td>Illumina NextSeq 500</td>
</tr>
<tr>
<td><strong>DGRP: Mackay et al., Huang et al. [49, 50]</strong></td>
<td>North America</td>
<td>1</td>
<td>205</td>
<td>Inbred</td>
<td>Illumina GAIIx/Illumina HiSeq 2000</td>
</tr>
<tr>
<td>Gilks et al. [51]</td>
<td>North America</td>
<td>1</td>
<td>223</td>
<td>F1 hemiclones</td>
<td>Illumina HiSeq 2500</td>
</tr>
<tr>
<td>Fabian et al. [52]</td>
<td>North America</td>
<td>3</td>
<td>3</td>
<td>Pooled Isofemales(^f)</td>
<td>Illumina GAIIx</td>
</tr>
<tr>
<td>Bergland et al. [41•]</td>
<td>North America</td>
<td>5</td>
<td>13</td>
<td>Pooled Isofemales(^g)</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td>Campo et al., Kao et al. [53],[31••]</td>
<td>North America</td>
<td>13</td>
<td>58</td>
<td>Inbred</td>
<td>Illumina GAIIx/Illumina HiSeq 2000/Illumina HiSeq 2000</td>
</tr>
<tr>
<td><strong>DrosRTEC: Machado et al. [6••]</strong></td>
<td>North America/Europe</td>
<td>20</td>
<td>60</td>
<td>Pooled wild-caught males(^h)</td>
<td>Illumina HiSeq 2000/Illumina HiSeq 4000</td>
</tr>
<tr>
<td>Reinhardt et al. [27]</td>
<td>North America/Oceania</td>
<td>4</td>
<td>4</td>
<td>Pooled Isofemales(^i)</td>
<td>Illumina GAII</td>
</tr>
<tr>
<td>Kolaczkowski et al. [54]</td>
<td>Oceania</td>
<td>2</td>
<td>2</td>
<td>Pooled Isofemales$^d$</td>
<td>Solexa/Illumina GA1s</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------</td>
<td>----</td>
<td>----</td>
<td>-----------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Bergman et al. [55]</td>
<td>Africa/Europe/North America</td>
<td>3</td>
<td>50</td>
<td>Isofemales</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td>King et al. [56]</td>
<td>Worldwide</td>
<td>15</td>
<td>15</td>
<td>Inbred</td>
<td>Illumina GAII</td>
</tr>
</tbody>
</table>

$^a$554 flies/pool; $^b$97 flies/pool; $^c$5 flies pooled from each of 113 isofemale lines; $^d$100-160 flies/pool; $^e$~33-40 flies/pool; $^f$39-102 flies/pool; $^g$~50-200 flies/pool; $^h$27-164 flies/pool; $^i$~16 flies/pool; $^j$38-40 flies/pool.
In 2016, the Drosophila Genome Nexus resource reassembled the published *D. melanogaster* individual genomes available at that time, and added a set of new unpublished ones using common mapping and variant calling pipelines [1]. As such, the Drosophila Genome Nexus represents a valuable effort to homogenize the alignment and filtering process and thus allows the direct comparison between heterogeneous samples [1, 2]. Two new datasets of individual lines, both of them from European natural populations, have been made available since the publication of the Drosophila Genome Nexus resource[3, 4]. Moreover, two big datasets based on pool-sequencing of natural populations are also available now [5••],[6••].

In addition to genome sequences, other -omic datasets such as transcriptomics, and epigenomics are also available for *D. melanogaster* [7]. Resequencing strategies have also been applied to laboratory populations such as those used in evolve and resequencing experiments [8] or a panel of recombinant inbred lines [9]. Furthermore, resequencing data for other Drosophilid species are starting to emerge [10-14].

In this review, we focused on the genomic insights obtained from resequencing datasets of natural populations of *D. melanogaster*. Given the wide range of studies that can be performed using resequencing data, we focused on three evolutionary questions related to the demographic history of *D. melanogaster*, and to the spatial and temporal scales of natural variation (Figure 1).
Figure 1. Evolutionary insights from resequencing datasets in *D. melanogaster*. (a) Schematic representation of the historical colonization processes of *D. melanogaster* from its ancestral range. Numbers in each continent denote the number of localities from which the samples used in resequencing data studies come from. Red and blue arrows depict migration events from ancestral and derived populations, respectively. (b) Example of the change in frequency of a clinal allele and its correlation with latitude. Red and blue colours in the pie-chart symbols denote the relative frequency of two alleles segregating at the same locus. (c) Example of the fluctuation in the frequency of seasonal alleles across time (seasons and years). Each blue line corresponds to the change in frequency of a specific seasonal SNP. S, spring; F, fall; 1, 2 and 3: year 1, 2 and 3.

A short glimpse at some other evolutionary questions that can be addressed by using whole-genome resequencing datasets is given in Box1.
**BOX1.** Evolutionary questions that have been addressed by using whole-genome resequencing datasets

**Characterization of intra and interpopulation genetic variation:**
Identifying and quantifying the extent of genetic variability at intra/inter population level of different types of mutations:
- Structural Variants (SV): Inversions [57]; Transposable Elements (TEs) [58].
- Single Nucleotide Polymorphisms (SNPs) [5••].
- Mitochondrial [5••].
- Copy number variation (CNVs) [59].

**Direct estimation of the per generation mutation rate:**
- Using pairs of outbred individuals and their offspring [60].
- Using mutation accumulation lines [61].

**Determining population history and the role of non-selective forces:**
- Role of non-adaptive forces: migration (introgression, hybridization, admixture), genetic drift [29, 30].
- Demographic inference: reconstructing demographic history and estimating its parameters because of its own intrinsic interest, but also to include the demographic history into null models to test for selection [4, 18].

**Detecting natural selection:**
- Positive, negative and balancing selection [5, 6, 28].
- Selection on new mutations or on standing genetic variants [27, 62].
- Monogenic, oligogenic, or polygenic selection [63].
- Local, global, or parallel selection [3, 27].

**Establishing genotype-environment-phenotype map connections:**
- Genome-environment association (GEA) analyses: identifying the genetic variants associated with adaptation to local conditions [41•].
- Genetic architecture of phenotypes: identifying the total number of genes contributing to a given trait, their location, effect size, heritability, and the interactions among them, that is additivity, dominance, epistasis, pleiotropy, and with their environment [64].
- Genome-wide association studies (GWAS): developing a dense data sets of markers as a platform for genotyping thousands of recombinant lines or individuals required for accurate mapping of QTLs [65].

**Generating a framework/“test bench” for statistical testing hypothesis (model-based):**
- Models for background selection [66].
- Models for Recombination rate [67].
- Models for testing the fitness effect of the different types of DNA sequence variants [68].

**Studying experimental evolution:**
- Evolve and Resequence (E&R) experiments: comparing the change in frequency between a reference and an experimental population adapted to a specific laboratory condition to identify targets of selection, the different aspects of the adaptive process, and how they shape the phenotype [8].


The African origin of *D. melanogaster*

Unravelling the geographical origin of the *D. melanogaster* ancestral population is essential to disentangle the relative role of population history and natural selection in shaping genetic diversity in worldwide populations of this species. Studies performed using the resequenced populations in the Drosophila Genome Nexus resource provided new lines of evidence that narrowed and displaced the origin of the ancestral range population of *D. melanogaster* from sub-Saharan populations further south to Zambia and nearby southern African populations. These populations and especially Zambia showed classical signals of being the origin of the ancestral range, that is, highest genomic diversity for all types of mutations and lowest linkage disequilibrium levels [15],[16••],[17]. The authors also investigated the admixture levels in African populations and found that Zambia had very low levels of admixture while other African populations, such as northern sub-Saharan populations, exhibited very heterogeneous levels of admixture with derived populations, probably as a consequence of urbanization [15].

Resequencing datasets have also provided more information on the colonization events of this species. Kapopoulou *et al.* [18] dated for the first time the colonization of West Africa that was estimated to have happened 72 kya (C.I. = 62.6 - 79.4 kya). In another recent paper Kapopoulou *et al.* [4] suggested that the divergence time between African and European *D. melanogaster* populations is older than previously reported. The authors proposed that the split happened 30 kya before the previously reported 10 - 16 kya estimates based on 105 - 250 noncoding loci on the X chromosome [19, 20]. Note, however, that these two studies were based on the analysis a limited number of strains from a reduced number of populations.

Overall, African and European resequencing datasets have allowed us to identify the ancestral range of the species, and to estimate the split times of the first two colonization events in the demographic history of *D. melanogaster*.

The spatial scale of genomic variation

The analysis of nucleotide variability at the genomic level also allows us to explore more deeply the role of spatially varying selection and demography in the generation of clinal patterns of variation. Previous analysis based on a limited number of loci or tiling arrays showed that derived populations displayed signals of population bottlenecks such as reduced diversity, different levels of linkage disequilibrium, and an increase in
population differentiation among populations [21]. Some of these populations also showed latitudinal variation in genetic markers across geographical transects [22-24]. However, North American populations were also found to exhibit several genomic hallmarks consistent with secondary contact between European and African populations suggesting that demography could play a role in the generation of the observed clinal patterns [25, 26].

New data from *D. melanogaster* population resequencing projects have confirmed that most of the derived populations have suffered bottlenecks [16••] and that patterns of clinal variation at the genomic level in North America might have been generated as a consequence of secondary contact [27-30]. Bergland *et al.* [29] suggested that secondary contact could have also happened in Australian populations. In the case of the North American populations, Kao *et al.* [31••] and others [28, 29] provided new genomic evidence that these populations are the result of a two-wave colonization process, creating a secondary contact between African and European flies in south-east US and Caribbean populations, and generating a latitudinal cline of ancestry along the East Cost of the US. But the most relevant result came from the emergence of new evidence of the action of spatially varying selection shaping the frequency of some loci [6, 29]. Thus, the patterns of genomic variation along latitudinal clines are probably the result of the combined effect of historical demographic events and local adaptation (i.e., gene flow from Caribbean populations and clinally varying selection).

Most of the analyses of spatial variation at the genome-wide level have so far been done with North American and Australian populations while the European continent remained largely unexplored until very recently [3]. Indeed, patterns of variation within and between *D. melanogaster* populations sampled across 32 European locations have recently been analysed at multiple levels: from nuclear to mitochondrial DNA, and from single nucleotide polymorphisms (SNPs) to structural variants (inversions and transposable elements) [5••]. European *D. melanogaster* populations exhibited similar amounts of genetic variation and similar latitudinal clines for inversion polymorphisms as those found in other derived populations [5••]. However, new genomic regions with signals of selective sweeps that predate the colonization of the continent, and some others that pointed out to local adaptation have emerged. One of the most remarkable findings is that the analysis of neutral variation revealed a previously unknown
longitudinal pattern of population differentiation, which is mimicked by a longitudinal cline of two polymorphic inversions, the frequency of which decreases from East to West. A similar longitudinal pattern has also been observed for human populations suggesting co-migration of the two species [32, 33]. Overall, this new genomic data was a crucial step in developing a more general overview of the worldwide dynamics of spatial variation. However, some aspects regarding the patterns of variation of European samples remain elusive, for example the causes of the observed longitudinal pattern. In addition, incorporating data on populations from geographical regions currently underrepresented, such as South America and East Asia, should help us gain a more comprehensive picture of spatial variation in this species. Finally, samples of natural populations collected at a single time point offer only a snapshot of the patterns of genomic variation, as they miss the temporal axis.

The temporal scale of genomic variation

Studying the temporal dynamic of genomic variation is essential from an evolutionary point of view since a substantial fraction of the changes in frequency of different classes of mutations might be reflecting evolutionary responses to environmental changes through time. Environments are not just spatially heterogeneous but also vary through time and in many cases, these changes are very fast and abrupt, such as for instance, the rapid turnover in the use of pesticides. There are several well-documented examples of rapid adaptation in the literature [34]. Performing a series of temporal sampling, that is sampling the same population at several time points, allows for the detection of directional allele frequency changes even across very short timescales [35-37]. However, environmental changes can also be cyclic, with abiotic and biotic factors fluctuating through time (i.e., temperature, rainfall, photoperiod, food, competitors, and predators). Seasons are an example of cyclic annual fluctuations that influence living organisms from temperate environments by imposing different selective pressures with which they have to deal. Several examples of *D. melanogaster* biological traits that vary with seasons have been reported (e.g., [38, 39]). In a seasonal scenario, multivoltine organisms, that is those with multiple generations per year, might adapt by fluctuating the frequency of the loci under selection to allocate alternating selective pressures (e.g., some alleles may be beneficial during a specific season and deleterious during another). In contrast to long-term directional selection, which depletes variation, fluctuating seasonal selection will maintain polymorphism in populations across time. Despite its
importance, studies addressing how populations adapt to environmental seasonal changes are scarce (e.g., [40]).

Whole-genome resequencing data can help shed light on the tempo and mode of evolution over ecologically relevant temporal scales. Bergland et al. [41••] explored different genomic and evolutionary aspects of alleles that undergo rapid seasonal adaptation by analysing the allele frequency changes of a D. melanogaster population sampled in spring and fall over three consecutive years in a single North American location. This work was a turning point as it was the first study, to the best of our knowledge, to quantify the number of loci that are responding to seasonal selection and to reveal that the changes in the frequency of the SNPs at those loci are predictable. The frequencies of hundreds of putatively adaptive SNPs were found to oscillate among seasons and some of the oscillating SNPs were already present before the split between D. melanogaster and Drosophila simulans. Moreover, some of these SNPs were associated with phenotypes that were previously reported to vary seasonally in D. melanogaster populations, that is winter favoured allele had faster chill coma recovery time.

To deepen the understanding of the seasonal adaptation process, Machado et al. [6••] increased the number of sampled D. melanogaster populations to up to 20 different localities from two continents, North America and Europe, with some of the localities sampled for more than two consecutive years. As in Bergland et al. [41••], the study focused on the analysis of allele frequency fluctuations in order to search for signals of seasonal selection. By increasing the range of the two axis of environmental heterogeneity, space and time, more evidence that seasonal adaptation could be predicted by weather conditions, in only few weeks before sampling, came to light [6••]. It was also shown that changes in SNP frequency through seasons could be mirrored by changes in SNP frequency through space: environmental factors that vary with seasons also vary with latitude. Moreover, it was revealed that seasonal selection affects allele frequencies at ~1.0 - 2.5% of SNPs, and this change in frequency could reach up to 10% between spring and fall. The authors concluded that seasonal adaptation is a general and predictable evolutionary force in D. melanogaster populations living in temperate areas. Nevertheless, a larger-scale sampling is needed for a deeper understanding and a more accurate quantification of the proportion of
seasonal SNPs, since the analysed populations are heterogeneously distributed and hence, the whole geographic range has not been explored. Lastly, functional analysis will eventually be needed to reveal the molecular effects of the seasonal adaptive variants.

Future prospects
While most of our knowledge so far comes from analyses based on short-read sequencing, long-read sequencing technologies are starting to be applied to Drosophila species including *D. melanogaster* [42, 43]. These so-called third generation sequencing techniques hold promise for incorporating the knowledge on other types of genomic variants, such as structural variants to our current understanding of the evolution of *D. melanogaster*. This is crucial, as structural variants also generate variation that significantly contributes to genome function and genome evolution.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest

Acknowledgements
We thank González lab members and Alejandro Sánchez-Gracia for comments on the manuscript. JG is funded by the European Commission (H2020-ERC-2014-CoG-647900).
REFERENCES


In this work, the authors analysed the patterns of variation in a continent-wide scale sample of European D. melanogaster population at multiple levels. The analysis revealed, among other remarkable results, a previously unknown longitudinal pattern of population differentiation.


This is the most comprehensive study to date on seasonal adaptation in D. melanogaster. The analysis of the allele frequencies of samples from several North American populations showed that local adaptation to temporally varying selection pressures is a major force driving the evolution of D. melanogaster.


In this work, the authors reassembled published *D. melanogaster* genomes and added new data including several sub-Saharan populations. One of these populations, Zambia, shows strong signals of being the ancestral *D. melanogaster* population.


This study provides genome-wide evidence of African and European admixture in south-east US and populations from the Caribbean. The authors showed that the proportion of African ancestry decreases clinally with higher latitude.


[34] Pelissie B, Crossley MS, Cohen ZP, Schoville SD: Rapid evolution in insect pests: the importance of space and time in population genomics studies. Curr Opin Insect Sci 2018, 26:8-16.


This is the first study analysing the genetic basis of temporally variable selection. The authors found that seasonal adaptation is a relevant feature of D. melanogaster populations and suggest that its direction can be predicted by weather conditions.


