Modeling the demographic history of *Drosophila melanogaster* using Approximate Bayesian Computation and Next Generation Sequencing Data

Dissertation

an der Fakultät für Biologie

der Ludwig-Maximilians-Universität

München

vorgelegt von

Pablo Duchén Bocángel

aus

La Paz, Bolivien

München, October 2013

Dekan: Prof. Dr. Heinrich Leonhardt

1. Gutachter: Prof. Dr. Wolfgang Stephan

2. Gutachter: Prof. Dr. John Parsch

Dissertation eingereicht am: Oktober 2013

Datum der Disputation: 18. Dezember 2013

Erklärung:

Diese Dissertation wurde im Sinne von §12 der Promotionsordnung von Prof. Dr. Stephan betreut. Ich erkläre hiermit, dass die Dissertation nicht einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolg unterzogen habe.

Eidesstattliche Erklärung:

Ich erkläre hiermit Eidesstatt, dass die vorgelegte Dissertation von mir selbstständig, ohne unerlaubte Hilfe angefertigt wurde.

München, 17.10.2013

Pablo Duchén Bocángel

A mis padres.

Contents

List of Figures

Declaration of contributions as a co-author

In this dissertation I present the work of my doctoral research from January 2010 to September 2013. It is organized in three chapters. All of them are the result of collaborations with other scientists.

For the work presented in the first chapter, Wolfgang Stephan, Stefan Laurent and myself designed the study. I conducted the statistical analysis. Stephan Hutter and Daniel Živković collaborated with the analysis. I developed the computational tools required for the study. I did the writing and Wolfgang Stephan and Stefan Laurent did the revisions. This chapter has been published under the following title:

Duchen, P., D. Živković, S. Hutter, W. Stephan, and S. Laurent. 2013. Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population. Genetics 193:291–301.

For the second chapter Stefan Laurent and I designed the study. I conducted the analysis and developed the computational tools for it. The manuscript was revised by Wolfgang Stephan. A paper is in preparation containing everything included in this chapter.

For the third chapter John Pool, Chuck Langley, Kristian Stevens, David Begun, Russell Corbett-Detig and Ryuichi Sugino designed and conceived the experiments.

Charis Cardeno, Marc Crepeau and Perot Saelao performed the experiments. Russell Corbett-Detig, John Pool, J. Emerson, Kristian Stevens and myself analyzed the data. In particular, I performed the quality control of the sequences, the bioinformatics, and the assembly of all genomes together with Kristian Stevens. John Pool and Charles Langley wrote the paper. This paper has been published under the following title:

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley. 2012. Population genomics of sub-Saharan *Drosophila melanogaster* : African diversity and non-African admixture. PLoS Genetics 8:e1003080.

Summary

The main goal of this thesis was to develop demographic models of the fruit fly *Drosophila melanogaster* using Approximate Bayesian Computation and Next Generation Sequencing Data. These models were used to reconstruct the history of African, European, and North American populations.

Chapter 1 deals with the demographic history of North American *D. melanogaster*. This project was motivated by the release of full-genome sequences of a North American population, which showed greater diversity than European *D. melanogaster* although the introduction of the fruit fly to North America dates back to only $\sim$ 200 years ago. Here, we tested different demographic models involving populations of Zimbabwe, The Netherlands, and North Carolina (North America). Among the tested models we included variants with and without migration, as well as a model involving admixture between the population of Africa and Europe that generated the population of North America. We found that the admixture model fits best the observed data and we estimated the proportion of European and African admixture in the North American population. This population has 85% European and 15% African ancestry. We also estimated other population parameters including population sizes (current and ancestral) and divergence times. Confirming previous studies we also estimated the divergence between African and European populations to be around 19,000 years ago.

Chapter 2 deals with gene flow of *D. melanogaster* between African and European populations. Gene flow in *D. melanogaster* is well acknowledged but has not been

quantified using DNA sequence data. Previous studies from the late 80's based on allozymes found that the number of migrants per generation (Nm) was around 2 between several populations distributed worldwide. Here we used ABC methods and full-genome sequences to estimate the rate of migration between a population from Rwanda in Africa and a population from France. We found that Nm is around 10, which may imply there was a significant increase of gene flow in the last few decades. Our estimates show that the migration rate between these two populations is not necessarily symmetrical, with migration from Europe to Africa being higher than the opposite, although the difference does not seem to be significant. The study of gene flow is relevant because it constitutes an important force in population genetics. Theoretical studies have shown that, under neutrality, it is enough to have one migrant per generation to stop two populations from diverging and speciating, and if migration is strong enough it can also overcome the effect of selection.

Chapter 3 focuses on the sequencing of 130 full genomes of *D. melanogaster* from Africa and 9 from France. This project made use of haploid embryos, a new technique introduced in 2011 that allows the development of haploid *D. melanogaster*, which is then used for sequencing. The main goal of this project was to characterize these populations in terms of their diversity, admixture, and differentiation. We found that the most diverse population comes from Zambia, which is now thought to be much closer to *D. melanogaster*'s center of origin. We also found a significant amount of non-cosmopolitan admixture in several African populations, meaning that there exists a significant amount of back migration from Europe to Africa (corroborating the findings of chapter 2). In order to identify admixture tracts a new method was developed for this purpose, which uses a hidden Markov model to locate admixed regions along the genome. Admixed regions, as well as regions showing high levels of identity by descent were masked for downstream population genetics analyses. These full genomes constitute the second effort of the Drosophila Population

Genomics Project (DPGP 2) and are now available for the scientific community.

Zusammenfassung

Das Hauptziel dieser Doktorarbeit war die Entwicklung demografischer Modelle für die Taufliege *Drosophila melanogaster* basierend auf *Approximate Bayesian Computation* (ABC) und Hochdurchsatz Sequenzdaten. Wir verwendeten diese Modelle, um die Geschichte der afrikanischen, europäischen und nordamerikanischen Populationen zu rekonstruieren.

Kapitel 1 beschäftigt sich mit der demografischen Geschichte der nordamerikanischen *D. melanogaster*. Die Motivation für dieses Projekt war die Veröffentlichung der vollständigen Genomsequenzen einer großen Stichprobe einer nordamerikanischen Population. Diese Population zeigt eine größere genetische Vielfalt als europäische *D. melanogaster*, obwohl die Einführung der Taufliege nach Nordamerika nur rund 200 Jahre zurückliegt. Hier testeten wir verschiedene Modelle, die die Populationen von Simbabwe, den Niederlanden und North Carolina (Nordamerika) umfassen. Unter den getesteten Modellen waren Varianten mit und ohne Migration, sowie ein Modell, in dem die nordamerikanische Population durch *Admixture* zwischen der afrikanischen und europäischen Populationen entsteht. Das *Admixture*-Modell ergab die beste übereinstimmung mit den beobachteten Daten. Wir schätzten, dass die nordamerikanische Population zu 85% europäischer und zu 15% afrikanischer Abstammung ist. Weitere geschätzte Parameter waren aktuelle und ursprüngliche Populationsgrößen, sowie die Divergenzzeit zwischen afrikanischen und europäischen Populationen. Letztere schätzten wir auf rund 19,000 Jahre und damit auf einen ähnlichen Wert wie frühere Studien.

Kapitel 2 befasst sich mit dem Genfluss zwischen afrikanischen und europäischen Populationen von *D. melanogaster*. Dass solcher Genfluss stattfindet ist bekannt, aber er wurde bisher nicht mit DNA-Sequenzdaten quantifiziert. Studien aus den späten 80er Jahren (basierend auf Allozymen) schätzten die Zahl der Migranten pro Generation (Nm) zwischen mehreren Populationen weltweit auf rund zwei. Hier verwendeten wir ABC-Methoden und vollständige Genomsequenzen, um die Migrationsrate zwischen einer Population aus Ruanda in Afrika und einer Population aus Frankreich zu schätzen. Wir schätzten Nm auf etwa zehn, ein signifikant höherer Wert als in früheren Studien, was auf eine Zunahme des Genflusses innerhalb der letzten Jahrzehnte hindeuten könnte. Unsere Schätzungen zeigen, dass die Migrationsrate zwischen den beiden Populationen nicht symmetrisch ist. Migration von Europa nach Afrika scheint häufiger zu sein als Migration in die andere Richtung, wobei der Unterschied aber nicht signifikant war. Die Relevanz dieser Studie ergibt sich aus der Rolle von Genfluss als wichtige populationsgenetische Kraft. Theoretische Studien haben gezeigt, dass unter Neutralität ein einziger Migrant pro Generation genügt, um die Divergenz zweier Populationen und damit Artbildung zu stoppen. Wenn die Migration stark genug ist, kann sie auch die Wirkung der Selektion überwinden.

Kapitel 3 befasst sich mit der Sequenzierung von 139 vollständigen Genomen von *D. melanogaster*, 130 aus Afrika und 9 aus Frankreich. Dieses Projekt nutzte eine im Jahr 2011 eingeführte neue Technik: haploide Embryonen, die sich zu haploiden Fliegen entwickeln und dann für die Sequenzierung verwendet werden können. Das Hauptziel dieses Projekts war es, die verschiedenen Populationen in ihrer genetischen Vielfalt, *Admixture* und Differenzierung zu charakterisieren. Wir fanden die größte Vielfalt in der Population aus Sambia, so dass nun angenommen wird, dass der Ursprungsort von *D. melanogaster* in der Nähe dieser Population liegt. Wir fanden auch eine erhebliche Anzahl nicht-afrikanischer Allele in mehreren afrikanischen

Populationen, was bedeutet, dass es eine erhebliche Menge an Migration von Europa nach Afrika geben muss (in übereinstimmung mit den Ergebnissen von Kapitel 2). Um Genomregionen mit *Admixture* zu identifizieren, entwickelten wir ein neues Verfahren basierend auf einem "Hidden Markov-Modell". Regionen mit *Admixture* und solche mit hoher Abstammungsgleichheit wurden für weitere populationsgenetische Analysen maskiert. Diese vollständig sequenzierten Genome bilden die zweite Phase des *Drosophila Population Genomics Project* (DPGP2) und stehen nun der wissenschaftlichen Gemeinschaft zur Verfügung.

General Introduction

One of the major aims of population genetics is to understand the way evolutionary forces act on populations. Among these forces natural selection and genetic drift play a major role in determining the fate of alleles. Genetic drift is the random sampling of gametes chosen to reproduce and continue to the next generation (Kimura, 1983). This random picking of gametes changes the frequency of a given allele along its history and eventually results in either a fixation or extinction. Genetic drift will be stronger in small populations, provided that for larger populations it will take longer for an allele to fix or go extinct (Kimura, 1983). However, natural selection will also have a significant effect on a population. Depending on the strength of selection large populations can also be significantly affected and this effect can be seen very fast, especially if selection is strong (textbook examples include: Kettlewell, 1958; Grant and Grant, 2006). Natural selection can take multiple mechanisms of action, including positive, negative and balancing selection (Hartl and Clark, 2007).

Selection versus demography

At the molecular level if we sample some chromosomes from a population and look at their alignment we will notice the existence of standing variation in the form of single nucleotide polymorphisms (SNPs). If one of these variants is beneficial then selection will increase its frequency with time, as well as the frequency of surrounding SNPs linked to the selected one. This effect is also known as a "selective sweep" (Maynard-

Smith and Haigh, 1974). If we analyze patterns of variation along the chromosome we will see that regions subjected to positive selection will have lower levels of variation as a result of selective sweeps (Stephan et al., 1992). A similar pattern can be seen from negative selection, also known as purifying selection (Charlesworth et al., 1993; Stephan, 2010), whereas balancing selection is mostly going to favor the presence of two or more alleles segregating in a population (Clarke, 1964; Clarke and O'donald, 1964; Charlesworth, 2006). All in all, natural selection (and drift) will leave noticeable patterns in the genome, which are targeted by genome scans (e.g. Sabeti et al., 2006; Li and Stephan, 2006; Zayed and Whitfield, 2008). Such patterns and selective footprints could be easily found if the population's demographic history remained constant over time. However, demographic histories of populations almost never remain constant.

One of the main challenges when searching for footprints of adaptation is that the signatures of selection can be very easily confounded with signatures of demography. Typical signatures of positive selection include reduction of heterozygosity, excess of low-frequency variants (singletons), and an excess of high-frequency variants (Maynard-Smith and Haigh, 1974; Stephan et al., 1992). Weak negative selection also generates an excess of singletons and an overall reduction of heterozygosity (Fu and Li, 1993). Balancing selection often produces an excess of intermediate frequency variants (Charlesworth, 2006). However, typical sweep or weak negative selection patterns can also be generated by a population bottleneck, i.e. a drastic reduction in the number of individuals comes together with a reduction in genetic diversity. When this smaller population starts to expand new variants will be in low frequency, generating excess in the singleton class, the same as in positive or negative selection (e.g. Haddrill et al., 2005; Li and Stephan, 2006). Excess in intermediate frequency variants can also be generated by population admixture, resembling patterns of balancing selection.

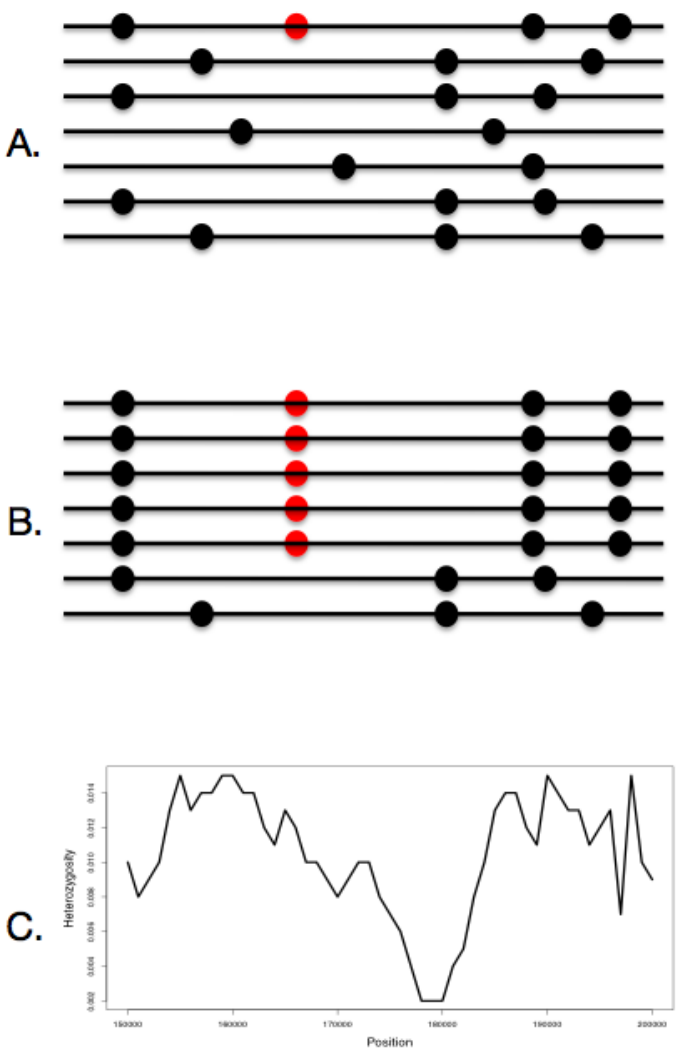

Figure 1: Graphical representation of a selective sweep. A) A beneficial mutation (red) pops up in the population. B) This beneficial mutation is positively selected and increases in frequency in the population together with other sites linked to it. This increase in frequency will result in a reduction in variability in the surroundings of the selected site. C) A hypothesized selective sweep: the x-axis represents the position along a chromosome and the y-axis represents heterozygosity.

The only pattern for which there is no known demographic effect able to produce it is an excess of high-frequency variants. However, it has been shown that ancestral misidentification can also produce a fake excess of high-frequency variants (Hernandez et al., 2007). Ancestral misidentification is the erroneous assignment of ancestral or derived categories on a particular site, due to back mutations in the lineages leading to the outgroup or ingroup of the species being studied. In general, all the facts presented in the last two paragraphs show how both demographic and selective instances can generate similar footprints. This is why we need to highlight the importance of an exhaustive understanding of the demography of a population in order to study its patterns of adaptation.

Demography of *Drosophila melanogaster*

The study of demography is not only important as a null model for selection scans. The study of demography gives us a better understanding of the history of a population or a species and this, in turn, contributes to the knowledege of the ecology of the species. This knowledge will have evolutionary, biogeographical, and conservational implications. Among the ecological implications we have the case of invasive species. A good example of this is the invasion of the fruit fly *Drosophila melanogaster* in North America some 200 years ago (Johnson, 1913; Sturtevant, 1920; Keller, 2007). Before 1875 there were no collections of *D. melanogaster* among the very well known dipteran fauna from United States and Canada. In that year the first specimens were collected along ports in New York, New Hampshire and Montreal (Johnson, 1913). Fifty years later *D. melanogaster* was the most common insect in North America (Keller, 2007). This rapid expansion was accompanied by an increase in diversity and a new variety of habitats to occupy and adapt to. The first chapter of this thesis deals with the analysis of demographic models for the North American population of *D. melanogaster*. There we reconstruct the history of colonization in

North America from its two possible source populations, namely Africa and Europe.

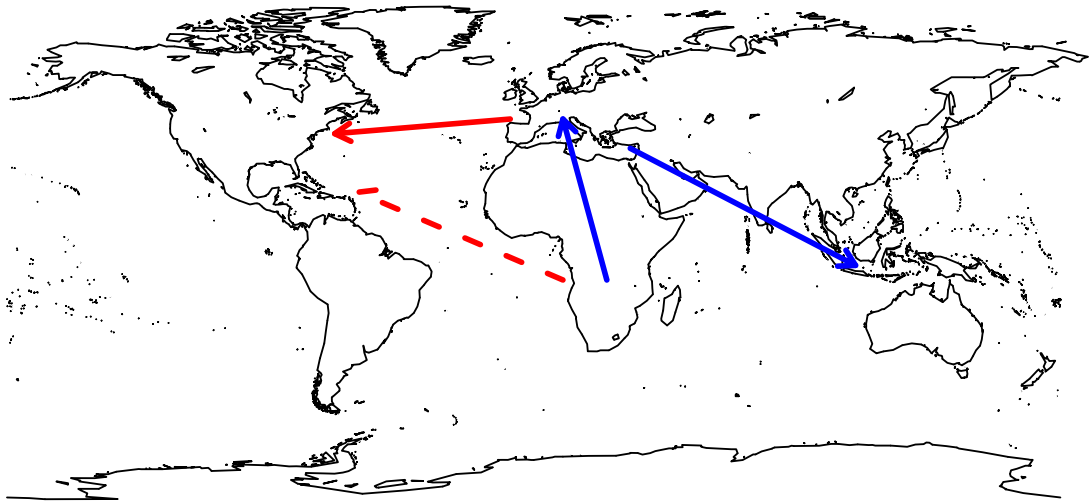

Figure 2: Demographic history of *D. melanogaster* since its origin in sub-Saharan Africa, as inferred from allozyme, morphometric and physiological data. Blue arrows represent inferred old colonizations, red arrows represent witnessed (solid) / hypothesized (dashed) recent colonizations. Modified from David and Capy (1988).

The second chapter of the thesis tackles another important evolutionary force: gene flow. This force plays a crucial role in divergence between populations and determines the strength of selection. Haldane (1930) showed that if the ratio between the migration rate and the selection coefficient is bigger than one then the effect of migration overcomes the effect of selection. Conversely, Wright (1931) showed that if the product between the migration rate (between two populations) and the effective population size is bigger than one then these two populations no longer diverge from each other. Again, *D. melanogaster* constitutes an ideal study system for migration and gene flow. As already mentioned above this fruit fly is a skilled colonizer and the rapid spread throughout North America is not the only example. Starting from its origin in sub-Saharan Africa (Tsacas and Lachaise, 1974; Lachaise et al., 1988; Begun and Aquadro, 1993; Andolfatto, 2001; Stephan and Li, 2007) *D. melanogaster*

first colonized the whole African continent (David and Capy, 1988), then Europe $\sim$ 19,000 years ago (Baudry et al., 2004; Li and Stephan, 2006; Thornton and Andolfatto, 2006; Laurent et al., 2011; Duchen et al., 2013), Asia $\sim$ 2500 years ago (Laurent et al., 2011), Australia $\leq$ 1000 years ago (David and Capy, 1988) and America just recently. Although *D. melanogaster* diverged from *D. simulans* 2.3 million years ago (Li et al., 1999) the spread throughout the world happened only in the last few thousand years. This burst of recent migration might be explained by the fact that this species is a human commensal, that is, most of the fruit fly's movement is human-driven (Lachaise and Silvain, 2004). For this reason, the study of migration in this model organism is relevant and applicable to understand *D. melanogaster*'s ecology and, indirectly, applicable to better understand human dispersal. Interestingly, there are very few studies quantifying the amount of gene flow in this species (Singh and Rhomberg, 1987; Kennington et al., 2003). The goal in this chapter was to quantify gene flow between several populations of *D. melanogaster* distributed worldwide, taking advantage of the demographic models developed in the first chapter.

Approximate Bayesian Computation

Two things are shared in the development of the first two chapters regarding the use of new methods and technologies: Next Generation Sequencing and Approximate Bayesian Computation. We will start with Approximate Bayesian Computation (ABC), which is used in Chapters 1 and 2. ABC was originally developed by Tavaré et al. (1997), and then improved by Pritchard et al. (1999) and Beaumont et al. (2002), among others. This method is very flexible when dealing with complex demographic scenarios. The main goal of ABC is to estimate population parameters of any given demographic model, such as population sizes, split times between populations, time of population size changes, migration rates or admixture events. The

advantage of ABC is that it directly approximates the posterior probability of each parameter via simulations without the need of calculating the likelihood of the data. Calculating this likelihood analytically for complex demographic models is usually not possible, and using numerical techniques often takes a lot of time.

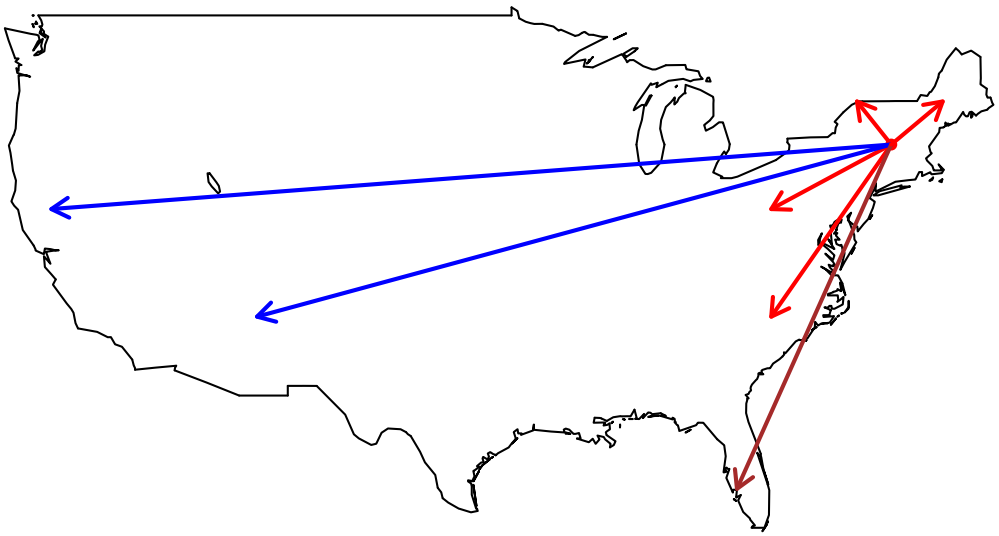

Figure 3: The expansion of *D. melanogaster* in North America according to Keller (2007). The first appearance was in 1875 in New York State. Red arrows represent colonizations within 10 years after 1875, brown arrows within 20 years, and blue arrows within 40 years.

Another difference with full-likelihood methods is that ABC does not use the full data but uses summaries of the data instead, called summary statistics. Among the summary statistics used in chapters 1 and 2 we have: the number of segregating sites S, Watterson's Θ_W (Watterson, 1975), Π, the number of haplotypes K (Depaulis et al., 1998), haplotype diversity, Tajima's D (Tajima, 1989), the linkage disequilibrium statistic Z_{nS} (Kelly, 1997), F_{st}, the site frequency spectrum (SFS), and the joint site frequency spectrum (JSFS) if more than one population is analyzed. By using summary statistics to estimate population parameters (instead of

the full data) ABC performs faster than full likelihood methods. However, this does not necessarily mean that the estimates will be more accurate or precise. In general, full-likelihood methods should be preferred over approximations whenever possible, unless it is unfeasible time-wise. For those cases ABC will perform much faster and with comparably good results.

ABC performs the following algorithm: 1) generate a vector of summary statistics from a target dataset, 2) simulate a dataset and generate the same vector of summary statistics, 3) calculate the Euclidean distance between these two vectors and accept the simulation if the distance is smaller than a given delta value, 4) repeat steps 2 and 3 until "enough" simulated vectors are accepted (Pritchard et al., 1999). Given that the population parameters are known for the simulated datasets it is then possible to construct the posterior distribution of each parameter with the set of accepted simulations. This is called the ABC rejection method. There exist enhancements to the classical ABC (rejection method): since it is difficult to get smaller distances with increasing number of summary statistics Beaumont et al. (2002) proposed a "regression" method. With this method a larger number of simulations will be accepted but there will be different weights given to the distances depending on how close they are to the target vector. With this new set of weighted distances a local regression is performed and the new parameter values are obtained, from which the posterior distribution is generated. Another enhancement has been proposed by Wegmann et al. (2009). They tackled the problem of noise generated by multidimensionality when using several and, often, correlated summary statistics. They proposed a reduction of dimensionality using partial least squares, which is similar to principal component analysis. This way noise is reduced and variance is maintained. For the demographic models analyzed in chapters 1 and 2 all the above enhancements are used. Other ways of improving the estimation have been suggested by Joyce and Marjoram (2008) and Fearnhead and Prangle (2012). They

proposed algorithms to choose only the most informative summary statistics in order to reduce dimensionality. Blum and François (2010) proposed a combination of machine learning and importance sampling to improve the estimation of posterior densities.

Population genomics in *D. melanogaster*

The final chapter deals with the population genomics of sub-Saharan *D. melanogaster*. Although sub-Saharan *D. melanogaster* was already studied the novelty of this research lies in the use of full-genome sequences obtained by Illumina Next Generation Sequencing (NGS) technology. NGS has nowadays become the method of choice since it produces a huge amount of data in the form of short overlapping reads that cover the entire genome. These reads are then mapped to a reference genome or, alternatively, are assembled de novo. Here, an assembly to *D. melanogaster*'s reference genome was produced. It is important to remember the significance of acquiring full genomes in population genetics. With full genomes sequenced we have access to huge amounts data from which it is possible to cherry pick the regions of interest. NGS also allows us to sequence in parallel several samples from several populations. Datasets generated with NGS are very valuable for downstream population genetics analyses like the ones presented in chapters 1 and 2.

Aims

In general, the main goal of this thesis was to generate full-genome assemblies from *D. melanogaster* NGS data and then use ABC methods to study the demography of this organism. The demography of *D. melanogaster* is now available and ready to use for genomic scans for selection. I took part in the assembly of these genomes at the University of California Davis and the demographic analyses were performed

by myself at the University of Munich.

Chapter 1: Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population

Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen,[1] Daniel Živković, Stephan Hutter, Wolfgang Stephan, and Stefan Laurent
Evolutionary Biology, University of Munich, 82152 Planegg-Martinsried, Germany

ABSTRACT *Drosophila melanogaster* spread from sub-Saharan Africa to the rest of the world colonizing new environments. Here, we modeled the joint demography of African (Zimbabwe), European (The Netherlands), and North American (North Carolina) populations using an approximate Bayesian computation (ABC) approach. By testing different models (including scenarios with continuous migration), we found that admixture between Africa and Europe most likely generated the North American population, with an estimated proportion of African ancestry of 15%. We also revisited the demography of the ancestral population (Africa) and found—in contrast to previous work—that a bottleneck fits the history of the population of Zimbabwe better than expansion. Finally, we compared the site-frequency spectrum of the ancestral population to analytical predictions under the estimated bottleneck model.

TO date, several studies have confirmed that *Drosophila melanogaster* originated in sub-Saharan Africa and spread to the rest of the world (Lachaise *et al.* 1988; David and Capy 1988; Begun and Aquadro 1993; Andolfatto 2001; Stephan and Li 2007). With its cosmopolitan distribution we expect that different populations have evolved and adapted differently to distinct environments, making *D. melanogaster* a perfect study system for both adaptation and population history. Extensive research has been performed to detect signatures of adaptation at the genome level (Sabeti *et al.* 2006; Li and Stephan 2006; Zayed and Whitfield 2008). Such detection usually depends on the underlying demographic scenario, since demographic events can leave similar patterns on the genome as adaptive (selective) events (Kim and Stephan 2002; Glinka *et al.* 2003; Jensen *et al.* 2005; Nielsen *et al.* 2005; Pavlidis *et al.* 2008, 2010a). Therefore, a better understanding of the demography of a population will not only allow us to estimate past and present population sizes and the times of the population size changes but will also decrease the rate of false positives of signatures of adaptation. Here we study the demography of African, European, and North American populations, with an emphasis on the North American population.

There is evidence that *D. melanogaster* colonized North America <200 years ago (Johnson 1913; Sturtevant 1920; Keller 2007). *D. melanogaster* (then known as *D. ampelophila*) was first reported in New York in 1875 by New York State entomologist Lintner (Lintner 1882; Keller 2007). In the year 1879 several articles were published indicating the appearance of *D. melanogaster* in several parts of eastern North America, including Connecticut and Massachusetts (Johnson 1913). At that time the dipteran fauna was very well described. It is therefore unlikely that entomologists would have overlooked *D. melanogaster* for long (Keller 2007). Less than 25 years after its introduction, *D. melanogaster* became the most common dipteran species in North America (Howard 1900). Johnson (1913) suggested that North America could have been colonized from the tropics, since the first specimen of *D. melanogaster* in the new world was first described from Cuba (possibly following routes from Central or South America). However, the same author also suggests that the first individuals could have come in vessels from southern Europe during the Spanish regime or from western Africa during the slave trade.

Even if there is agreement regarding the origin of *D. melanogaster*, the demographic history of North American flies is still poorly understood, and population genetic analyses of both the ancestral and derived populations are

doi: 10.1534/genetics.112.145912
Manuscript received September 14, 2012; accepted for publication October 30, 2012
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145912/-/DC1.
[1]Corresponding author: Evolutionary Biology, University of Munich,, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany. E-mail: duchen@bio.lmu.de

required to tackle this problem. Begun and Aquadro (1993) and Andolfatto (2001) showed that variation in non-African populations (including North America) is a subset of that found in African populations. They suggested a simple "out-of-Africa" bottleneck scenario. Later, Kauer *et al.* (2002) and Caracristi and Schlötterer (2003) used microsatellite data for 40 X-linked loci to study several populations worldwide. Caracristi and Schlötterer (2003) found that some North American populations present only African alleles, whereas other North American populations present only European alleles. Based on the proportion of shared alleles and F_{ST} values, their study shows that American populations are closer to African populations than to European populations. Baudry *et al.* (2004) and Haddrill *et al.* (2005) analyzed 4 and 10 X-linked loci, respectively, but this time using sequence data. Baudry *et al.* (2004) suggested that rare alleles shared between non-African and African populations might represent immigrants from Africa. This agrees with the hypothesis of admixture between European and African flies suggested by Caracristi and Schlötterer (2003). Furthermore, Haddrill *et al.* (2005) found in their North American sample higher diversity and larger linkage disequilibrium than in their European sample, which is also compatible with an admixture scenario.

To infer the population history of North America, we also revisit the demography of the likely source populations from Africa and Europe. Concerning the demography of African *D. melanogaster* Glinka *et al.* (2003) and Pool and Aquadro (2006) found that African samples have an excess of rare derived mutations when compared to the standard neutral model. This excess can be generated by population expansion or a bottleneck. Li and Stephan (2006) proposed a population expansion model for the African population. However, it is still unclear if Zimbabwe is the center of origin. If Zimbabwe lies outside the center of origin we may expect that a bottleneck model would fit the data of the Zimbabwe population better than expansion, since range expansions are usually associated with bottlenecks and founder effects (Excoffier *et al.* 2009). Therefore, we decided to revisit the expansion scenario proposed by Li and Stephan (2006).

In this study we focus on modeling and inferring the demography of *D. melanogaster* using approximate Bayesian computation (ABC) (Tavaré *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002). First, we revisit the demography of the Zimbabwe population and compare a model of instantaneous population expansion with a population bottleneck. Second, having found the best model for our ancestral population we model the joint demography of Africa, Europe, and North America. Finally, we analyze the observed site-frequency spectrum (SFS) of the Zimbabwe population and compare it to analytical predictions.

Materials and Methods

SNP data

Individuals come from three populations: Zimbabwe in Africa (sample size $n = 12$), The Netherlands in Europe ($n = 12$), and Raleigh in North America ($n = 37$). Sequence data consist of 242 intronic and intergenic X-linked loci from each population. African and European loci were originally target sequenced by Glinka *et al.* (2003), Ometto *et al.* (2005), and Hutter *et al.* (2007), while North American loci were extracted from full-genome sequences (publicly available from the Drosophila Population Genomics Project at http://www.dpgp.org) that were created using Illumina next-generation sequencing (NGS) technology. As a first quality control step for the NGS data, all bases with a Phred quality control score <20 were masked. All 242 orthologous loci extracted from the North American data were then aligned to the European and African sequences using MUSCLE (Edgar 2004) to account for insertion/deletion polymorphism. *Drosophila simulans* has been used as an outgroup sequence. As a second quality control step, the alignments were inspected for singleton polymorphisms private to the North American sample and these positions were removed from further analysis. We believe that a sizable fraction of these singleton polymorphisms are created by sequencing errors. This is reflected by the fact that the average quality score of a base causing a singleton polymorphism is significantly lower than the quality of bases creating variants segregating at higher frequencies (Mann–Whitney *U*-test: $P < 2.2 \times 10^{-16}$) (Supporting Information, Figure S1). From all these loci we computed the mean and the variance of the following summary statistics: the number of segregating sites S_n, Watterson's Θ_W (Watterson 1975), the average number of pairwise differences in all pairwise comparisons of n sequences Π_n, Tajima's D (Tajima 1989), the number of haplotypes K (Depaulis and Veuille 1998), the linkage disequilibrium statistic Z_{nS} (Kelly 1997), and the distance of Nei as a measure of population differentiation (Nei and Li 1979). Summary statistics of the North American population after exclusion of singletons are also reported. Additionally, we computed the joint site-frequency spectrum (JSFS) of all three pairs of populations, namely: Africa–Europe, Africa–North America, and Europe–North America (Figure S2). Each JSFS was summarized in four classes according to the Wakeley–Hey model (Wakeley and Hey 1997). These summaries are W1 (private polymorphisms in population 1), W2 (private polymorphisms in population 2), W3 (fixed differences between populations), and W4 (shared ancestral polymorphisms). This group of summary statistics, plus the summaries of the JSFS, constitutes our "observed vector" or "observed data" (Tables 1 and 2).

Demographic models of Africa

We first analyzed the data from the ancestral population in Africa. We tested whether an instantaneous expansion or a bottleneck fits better the observed data. The instantaneous expansion model had three parameters: ancestral population size, current population size, and time of expansion (Figure S3). The bottleneck model includes the severity as an additional parameter, which is defined as the ratio of the bottleneck duration and the population size during the

Table 1 Mean and variance (in parentheses) of observed summary statistics over all 242 fragments

	Africa (n = 12)	Europe (n = 12)	North America (n = 37)	Africa (no singletons)	Europe (no singletons)	North America (no singletons)
No. of segregating sites S_n	17.55 (81.31)	6.35 (29.31)	13.10 (50.22)	10.70 (42.45)	4.11 (18.30)	7.47 (29.57)
Watterson's Θ_W	5.91 (9.40)	2.11 (3.30)	3.22 (3.12)	3.57 (4.72)	1.36 (2.01)	1.83 (1.79)
Π_n	5.13 (9.06)	2.18 (4.81)	2.52 (3.64)	3.92 (6.35)	1.36 (2.56)	2.05 (3.16)
Tajima's D	−0.67 (0.34)	−0.09 (1.43)	−0.77 (1.05)	0.33 (0.43)	−0.006 (1.56)	0.21 (1.15)
No. of haplotypes K	9.46 (5.26)	3.87 (3.71)	10.31 (23.24)	8.09 (9.47)	2.85 (2.62)	6.98 (19.25)
Kelly's Z_{nS}	0.15 (0.01)	0.43 (0.075)	0.21 (0.055)	0.23 (0.03)	0.53 (0.08)	0.38 (0.16)

bottleneck (Figure S3). We fixed the duration of the bottleneck to 1000 generations (Laurent *et al.* 2011).

Demographic models of North America–Europe–Africa

Based on the best model for the ancestral population we tested five different models that included all three populations (Figure 1 and Table S2). These five models are: model A ("no migration"), which comprises Africa as the ancestral population; the colonization of Europe is followed by exponential growth, and the colonization from Europe to North America with subsequent exponential growth. Model B ("migration") matches model A but adds an equal migration rate between all populations starting at the colonization time of North America (we assumed that migration between continents increased significantly with human dispersal, which started a few centuries ago). Model C ("admixture") equals the previous models until North America is founded through an admixture between Africa and Europe followed by exponential growth in North America. In model C, we estimated the proportion of European and African ancestry in the founding population of North America. Model D ("no migration II") has Africa as the ancestral population with North America and Europe splitting independently from Africa. Finally, model E ("migration II") matches model D but adds an equal migration rate between all populations starting at the colonization time of North America. Models A and D have 10 parameters, and models B, C, and E have 11 parameters each (Figure 1). In all models the time of colonization of North America was given a very small prior around 200 years ago (the time of the reported colonization of North America). We also let migration due to human-associated dispersal start at this same time (for models B and E). Model selection was performed with all models. For further analysis we selected only models A to C because of the biological assumptions that were already presented in the Introduction. A thorough explanation of the reasons why we discarded other models is presented in the *Discussion*. A more detailed description of all analyzed models can be found in the supporting information (Table S2) and in Figure 1.

ABC simulations

We simulated 100,000 data sets for each of the models described above following the protocol of Laurent *et al.* (2011). Each simulated data set consisted of 242 loci with individual per locus sample sizes, as well as mutation and recombination rates identical to the ones found in the observed data set. Mutation and recombination rates per site per generation for each locus were taken from Laurent *et al.* (2011). Our primary tool was the coalescent simulator *ms* by Hudson (2002). Each parameter was chosen from uniform prior distributions (see Table S1). Missing nucleotides (mostly present in the North American population) were also simulated at the same positions as they occur in the observed data. We accomplished this by following two steps: (1) from the observed data set we generated a missing-nucleotide table with the relative positions (beginning and end) of each chunk of missing nucleotides and recorded this information for each line and for each fragment and (2) by a simple manipulation of the *ms* output we masked all simulated polymorphisms that occurred at the same relative positions that were indicated in the missing-nucleotide table. From the *ms* output we also excluded all singletons that occurred in the simulated North American population. Following the same procedure as with the observed data set we calculated the summary statistics, the SFS, and the JSFS from the modified *ms* output, taking into account missing data in all calculations. Handling of priors, simulation of missing data, exclusion of singletons, and calculation of summary statistics was coded by ourselves. The software

Table 2 Comparisons between pairs of populations

	Africa–Europe	Africa–North America	Europe–North America
Distance of Nei (with singletons in North America)	0.78 (0.66)	1.12 (1.38)	0.59 (1.15)
Distance of Nei (without singletons in North America)	0.69 (0.44)	1.01 (0.93)	0.53 (0.72)
W1 (private polymorphisms of population 1)	2278	1961	214
W2 (private polymorphisms of population 2)	363	743	924
W3 (fixed differences between populations)	17	86	89
W4 (shared polymorphisms between poulations)	647	990	809

The first two lines denote mean and variance (in parentheses) of Nei's distance, and lines 3 to 6 the observed classes of the JSFS.

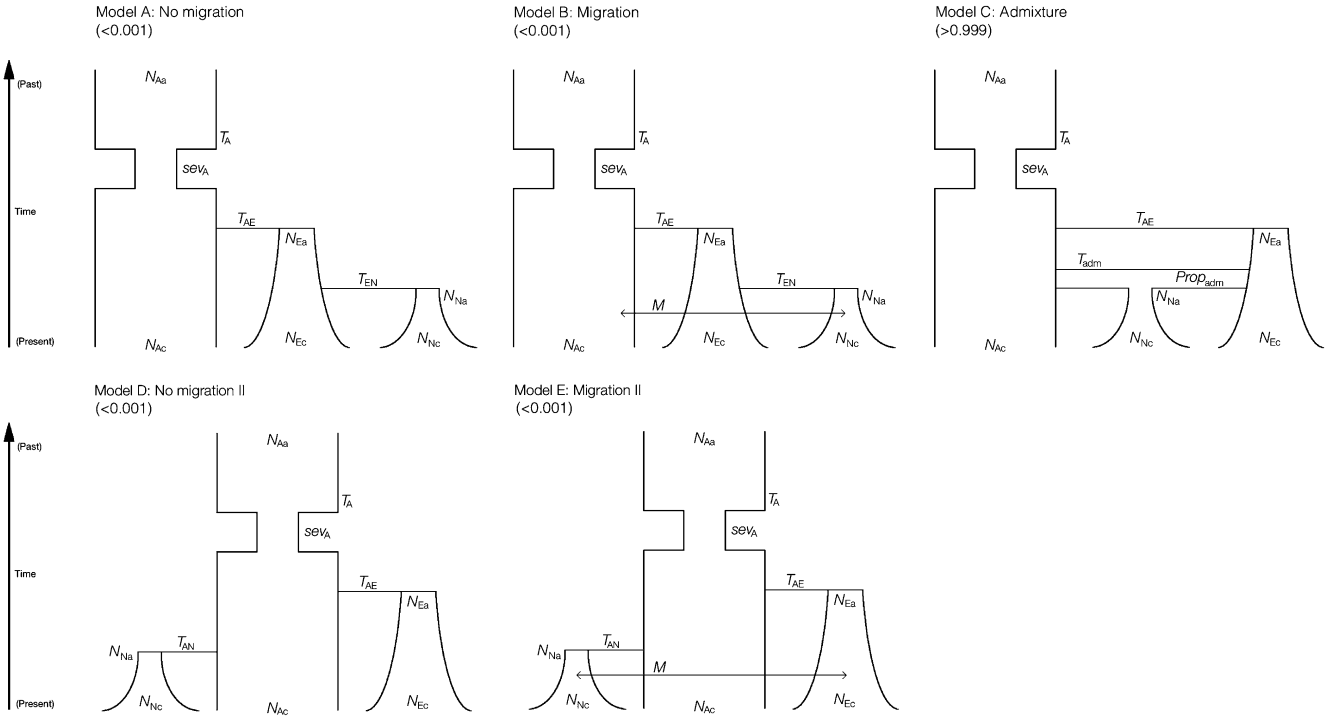

Figure 1 Three-population models. Numbers in parentheses are the posterior probabilities of each model. The symbols are explained in Table 3.

msABC (Pavlidis *et al.* 2010b) is able to perform similar simulations but does not calculate the JSFS. However, we still used msABC to validate our prior distributions. We launched simulations on a 64-bit Linux cluster with 510 nodes (at the Leibniz-Rechenzentrum LRZ, Munich).

Model choice

Model selection was also performed within an ABC framework. Posterior probabilities for each model were calculated according to Fagundes *et al.* (2007). Model selection was done based on the mean and the variance of S_n, mean and variance of Tajima's D and linkage disequilibrium (Z_{nS}). In our analysis (see *Results*) Watterson's Θ_W, Π_n, and K were correlated with S_n and therefore its inclusion did not change the results of the model choice procedure. Model selection was also performed separately using the summaries of the JSFS of all pairs of populations. The model with the highest posterior probability when comparing bottleneck and expansion for the African population as well as the three-population models was chosen as the best fit to the observed data. A validation for using 100,000 simulations for model choice was also performed: we conducted model choice for bottleneck/expansion and between all three-population models A to E for varying numbers of simulations ranging from 10,000 to 200,000 simulations. For the bottleneck *vs.* expansion case we show that starting at 50,000 simulations the posterior probability of the best model does not change significantly when the number of simulations is increased (Figure S4). For the three-population model choice the posterior probability of the best model is always >0.999 if the number of simulations is 10,000 or higher. Therefore a choice of 100,000 simulations for model choice is enough. Model choice performance was assessed by simulating 1000 different pseudo-observed data sets under models A, B, and C (samples for each parameter were taken from the prior distributions as well as from the posterior distributions based on the rejection method). Model choice was performed using the same method as above for each simulated vector of summary statistics. We considered one model to be preferred over the other if the Bayes factor of the models under comparison was above 3.

Parameter estimation

We estimated population parameters of the best African model and of the best three-population model. The number of simulations for parameter estimation was increased to 1,000,000. To validate the use of 1,000,000 simulations for parameter estimation we calculated the mean square error (MSE) of model parameters for varying numbers of simulations, ranging from 100,000 to 1,000,000 simulations (Table S3). Additionally, we also plot the mode and the 95% confidence intervals for varying numbers of simulations (Figure S5). We show that the MSE of each estimate and the estimated mode stay both relatively constant (Table S3 and Figure S5). Therefore, 1,000,000 simulations are enough for parameter estimation. Estimation was based on ABC rejection (Tavaré *et al.* 1997; Pritchard *et al.* 1999) and regression (Beaumont *et al.* 2002) methods. Both methods were performed using Wegmann's ABCtoolbox (Wegmann *et al.* 2009) and checked with Csilléry's abcR (Csilléry *et al.* 2012). First, we pooled all statistics and checked for correlations with

the parameters. We did not keep statistics that did not correlate with any parameter, because keeping them does not provide information for the estimation and would only add noise to the final estimates. All these statistics were transformed using partial least squares (p.l.s.) as implemented in Wegmann *et al.* (2009). This transformation is advantageous because it extracts a small number of orthogonal components from a highly dimensional array of summary statistics. The new set of transformed statistics (with reduced dimensionality) reduces the noise produced by uninformative summary statistics. Moreover, the p.l.s.-transformed statistics are completely uncorrelated with one another ensuring the assumption of singularity, which is required for estimating parameters according to the regression method (Beaumont *et al.* 2002).

Predictive simulations

To check for the quality of our parameter estimates we took two approaches: (1) we sampled parameter values from the posterior distributions (based on the regression method) of each parameter estimate and resimulated data sets, and (2) we plotted the distributions of summary statistics directly from the set of the 5000 simulations closest to the observed data (which represents a sample of the joint posterior distribution based on the rejection method). The resulting distributions of summary statistics were compared to the observed ones for both approaches and plots were generated (see *Results*). Both approaches were performed only under the best model, since this is a test to see how well the best model fits the observed data. The same predicitive simulations were also performed for autosomal data (50 intergenic and intronic loci from chromosome 3R) to check how good our best model can predict autosomal summary statistics. For the sake of computational simplicity we assumed a relative effective population size (N_e) ratio of 0.75 for X-linked *vs.* autosomal loci in our simulations. This assumes a 1:1 male/female ratio in all populations even though we have evidence that actual sex ratios might deviate from these expectations (Hutter *et al.* 2007). However, we expect that this simplification should have only minor effects on our ability to predict the autosomal data since even in extreme cases of sex bias the X/A ratio of N_e can never drop below 0.5625 or exceed 1.125 (Hedrick 2011, Chap. 4).

Prediction of the site-frequency spectrum of Zimbabwe

Our available sequence data not only allow us to summarize genetic diversity with S_n, Θ_W, or Π_n, but also allow us to compute the observed SFS of the African population (Figure 2) and compare it to predictions under a given demographic model. Analytical methods for predicting the SFS of one population for arbitrary deterministic changes in population size have been successfully developed (Griffiths and Tavaré 1998; Živković and Wiehe 2008; Živković and Stephan 2011) and are briefly revisited as follows. Let T_n, . . ., T_2 be the time periods during which the genealogy has n, . . ., 2 lineages, respectively. Furthermore, let $\lambda(t) = N(t)/N$ de-

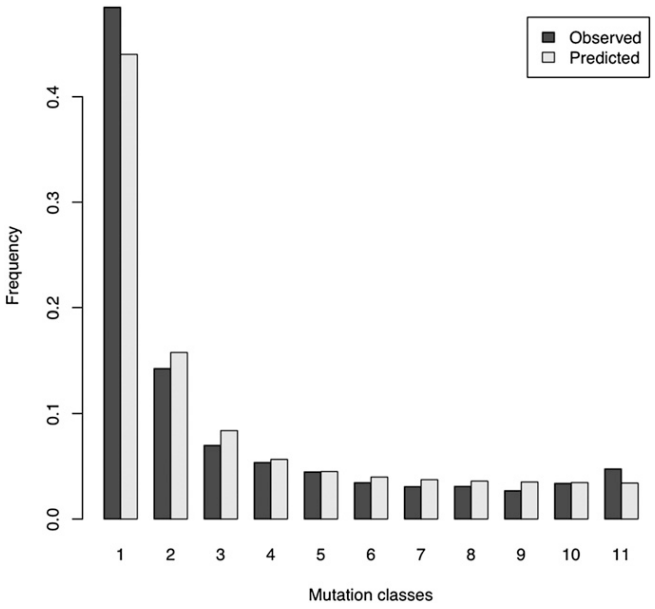

Figure 2 Observed (solid) and predicted (shaded) site-frequency spectrum of the African population. To calculate the frequency classes Equation 4 was used.

note the ratio of the population sizes at time t in the past and the present. The probability $p_{n,k}(i)$ that a randomly chosen line of waiting time T_k, $k = n$, ..., 2, has i descendants, $i = 1$, ..., $n - 1$, during time T_n (Fu 1995; Griffiths and Tavaré 1998) is

$$p_{n,k}(i) = \binom{n-i-1}{k-2} \Big/ \binom{n-1}{k-1}. \tag{1}$$

The mean waiting times are given by

$$E(T_k) = \sum_{j=k}^{n} (-1)^{j+k} \alpha_{n,j,k} \int_0^\infty \exp\left\{ -\binom{j}{2} \int_0^t \frac{1}{\lambda(u)} du \right\} dt, \tag{2}$$

where

$$\alpha_{n,j,k} = \frac{(2j-1)n!(n-1)!(k+j-2)!}{(j-k)!k!(k-1)!(n-j)!(n+j-1)!}.$$

The integral in (2) can be solved explicitly for models that consist of multiple instantaneous changes in population size and be evaluated numerically for models that include phases of exponential growth. Let L_i be the total length of branches leading to i descendants, where i represents singletons, doubletons, etc. Then,

$$E(L_i) = \sum_{k=2}^{n-i+1} k p_{n,k}(i) E(T_k). \tag{3}$$

Assuming an infinitely many sites mutation model (Kimura 1969), the expected unfolded site frequency ξ_i for each class i is given by

$$E(\xi_i) = \frac{E(L_i)}{\sum_{k=2}^{n} k\, E(T_k)}. \quad (4)$$

We use the ABC parameter estimates obtained for Zimbabwe as an input to the equations shown above, calculate the SFS based on Equation 4, and compare it to the observed SFS.

Results

Observed data

A first examination of the observed summary statistics (Table 1) shows that Africa is the most diverse population (based on the number of segregating sites), followed by North America and Europe. Watterson's Θ_W and Π_n follow the same pattern. Tajima's D is most negative in North America (−0.77), followed by Africa (−0.67) and Europe (−0.09). Linkage disequilibrium (Z_{nS}) is highest in Europe (0.43) compared to North America (0.21) and Africa (0.15). Population differentiation (Table 2) is highest between Africa and North America (distance of Nei = 1.12), followed by Africa–Europe (0.78) and North America–Europe (0.59). All these comparisons are based on the observed data set that included singletons in North America. The resulting statistics of North America after excluding singletons can also be found in Table 1.

The SFS of the African population is shown in Figure 2. Regarding the JSFS (Table 2) we observe an excess of private polymorphisms in Africa when compared to private polymorphisms in Europe (2278 *vs.* 363) and North America (1961 *vs.* 743) (W1 *vs.* W2). We must keep in mind that singletons were excluded from the North American population, and these singletons are mostly private to North America. The opposite pattern is seen when comparing private polymorphisms in Europe to private polymorphisms in North America (214 *vs.* 924). Shared polymorphism (W4) has its lowest value between Africa and Europe (647) when compared to Africa–North America (990) and Europe–North America (809). The number of fixed differences between populations is small in all pairwise comparisons (W3).

African demography

Model choice results show that a population bottleneck in Africa (P = 0.987) fits the observed data better than an expansion (P = 0.013). We used the following statistics for parameter estimation of the best model: mean and variance of S_n, mean and variance of Tajima's D, and mean Z_{nS}. We estimated these parameters (Table 3) using the priors listed in Table 4. After the reduction of dimensionality using partial least squares (see *Materials and Methods*) we kept three components from the original five statistics used. The estimated ancestral and current N_e are 4.9 million and 5.2 million individuals, respectively. The bottleneck severity (Log_{10} scale) was estimated as 0.21, which corresponds to ~620 individuals for a fixed bottleneck duration of 1000 generations. The estimated time of the bottleneck is ~200,000 years ago, assuming 10 generations per year (Table 4 and Figure S6). Predicted distributions of summary statistics for the bottleneck and the expansion models overlap significantly. However, observed Tajima's D as well as the mean and the variance of S_n are reproduced more often by the bottleneck model than by the expansion model (Figure S7). Estimations of the African parameters were also performed using the classes of the folded SFS of Zimbabwe but the results do not vary significantly (data not shown).

Table 3 Parameters used in models A, B, C, D, and E

Abbreviation of parameter	Explanation
N_{Aa}	Ancient population size of Africa
sev_A	Severity of the bottleneck in Africa
T_A	Time of the bottleneck in Africa
N_{Ac}	Current population size of Africa
T_{AE}	Time of split between Africa and Europe
T_{AN}	Time of split between Africa and North America
N_{Ea}	Starting population size of Europe
N_{Ec}	Current population size of Europe
T_{EN}	Time of split between Europe and North America
N_{Na}	Starting population size of North America
N_{Nc}	Current population size of North America
M	Migration rate between all populations
T_{adm}	Time of admixture between Africa and Europe
$Prop_{adm}$	Proportion of European admixture in North America

Site-frequency spectrum

The SFS of the observed African data has an excess of high-frequency-derived variants (Figure 2, solid bars), while the predicted SFS under a bottleneck does not show such a large excess (Figure 2, shaded bars). Predicted values were calculated using the modes of the parameter estimates under the bottleneck scenario (Table 4) and applying Equation 4. Predicted values fit the observed SFS better than the expansion model of Li and Stephan (2006) for the intermediate-frequency classes, but not for the low-frequency variants. The largest relative discrepancies are found for both models for the high-frequency variants that make the SFS slightly U shaped.

Table 4 Parameter estimates of the African population

Parameter	Prior	Mode	95% quantiles
N_{Ac}	unif(1×10^5, 1×10^7)	4,975,360 individuals	(2.40×10^6, 9.13×10^6)
T_A (in years)	unif(1×10^2, 4×10^5)	237,227 years ago	(0.82×10^5, 3.45×10^5)
N_{Aa}	unif(1×10^5, 1×10^7)	5,224,100 individuals	(1.98×10^6, 9.55×10^6)
sev_A (Log_{10})	unif(−2,2)	0.21	(−0.15, 0.57)

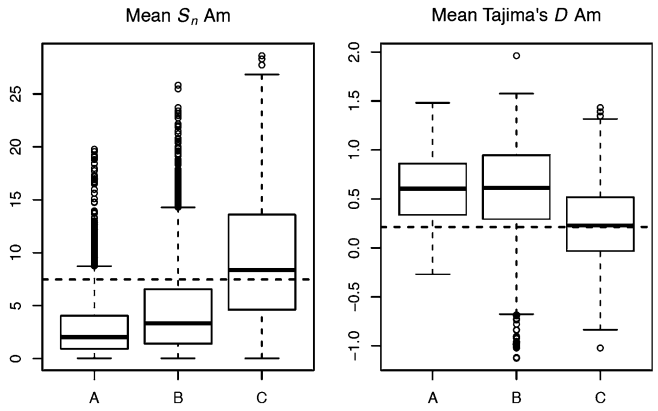

Figure 3 Predicted summary statistics under models A, B, and C for the North American population based on the rejection method. The horizontal dashed line represents the observed value.

North American demography

The model with the highest posterior probability is the admixture model C with $P > 0.999$. Model choice yields the same results when using summary statistics and also when using the JSFS (in both cases the posterior probability of model C is >0.999). Parameters of this model are explained in Table 3. Predictive simulations based on both the regression and rejection methods show that admixture is the only model that can explain the diversity observed in North America (Figure 3 and Figure S8). Admixture can also explain better the observed Tajima's *D* in America (Figure 3). It is important to remember at this point that diversity in North America is higher than in Europe, although the colonization of North America has been much more recent than the one inferred for Europe. It is thus reasonable to believe that admixture is playing an important role in this case.

Other parameters, such as diversity in Africa and Europe can be explained by both admixture and nonadmixture models (Figure S8 and Figure S9). The accuracy of the model choice procedure shows that the simulated model could be correctly identified in 90% of the cases. The cases in which model C is not preferred occur when one or a combination of the following events happen: (a) the time of split between African and European populations is very young (about 1000 to 2000 years ago), (b) the proportion of European ancestry in the North American population is very high (above 90%), and (c) the founding population of Europe is large (in the order of 100,000 individuals). The results of model choice performance when sampling from the posterior distributions of each parameter do not vary significantly with the ones we provide here (see *Materials and Methods*).

For estimating the parameters of model C, we used the following statistics: mean and variance of S_n in Africa, mean and variance of Tajima's *D* in Africa, mean *K* in Africa, mean and variance of Tajima's *D* in Europe, mean and variance of *K* in Europe, mean Z_{nS} in Europe, mean S_n in North America, mean and variance Tajima's *D* in North America, mean and variance of *K* in North America, mean Z_{nS} in North America, mean distance of Nei Africa–Europe, mean distance of Nei Africa–North America, mean distance of Nei Europe–North America, W1 Africa–North America, W2 Africa–North America, W4 Africa–North America, W1 Africa–Europe, W2 Africa-Europe, and W2 Europe–North America. The abovementioned statistics were chosen after pooling all statistics and checking for correlations between statistics and parameters (see *Materials and Methods*). After dimensionality reduction using partial least squares we kept six components. Parameter estimates (Table 5 and Figure S10) imply that African and European populations split around 19,000 years ago and Europe was founded with around 17,000 individuals. These estimates are in agreement with previous studies (Li and Stephan 2006; Laurent *et al.* 2011). The North American population was founded by ~2500 individuals from which ~85% are of European ancestry and the remaining of African ancestry (Figure 4). The current population sizes of Europe and North America cannot be estimated accurately.

Predictive simulations of model C (Figure S11 and Figure S12) were generated by sampling parameters from the posterior distributions (based on the regression method). These parameters were used to simulate data sets and calculate summary statistics and JSFS statistics (see *Materials and Methods*). The resulting distributions show that all summary statistics can be well predicted by the admixture model (Figure S11 and Figure S12). The only statistics that are overestimated are the number of fixed differences (W3) between Africa and North America or Europe and North America and the distance of Nei between Europe and America. W3 and distance of Nei are related to each other, and an increase in one involves always an increase in the other. An improvement of the model in this aspect is discussed below

Table 5 Joint parameter estimates of the European and North American populations

Parameter	Prior	Mode	95% quantiles
T_{AE} (decimal log generations)	unif(4,7)	5.29 (~19,000 years ago)	(4.69, 5.86)
T_{adm} (decimal log generations)	unif(2,4)	3.16	(2.08, 3.82)
N_{Ec}	unif($1x10^4$,1×10^7)	3,122,470 individuals	(0.39×10^6, 9.55×10^6)
N_{Ea} (decimal log)	unif(2,5)	4.23 (~17,000 individuals)	(3.58, 4.83)
N_{Nc}	unif($1x10^4$,3×10^7)	15,984,500 individuals	(1.11×10^6, 28.8×10^6)
N_{Na} (decimal log)	unif(2,5)	3.40 (~2500 individuals)	(2.20, 4.79)
$Prop_{adm}$	unif(0.01,0.99)	0.85	(0.64, 0.97)

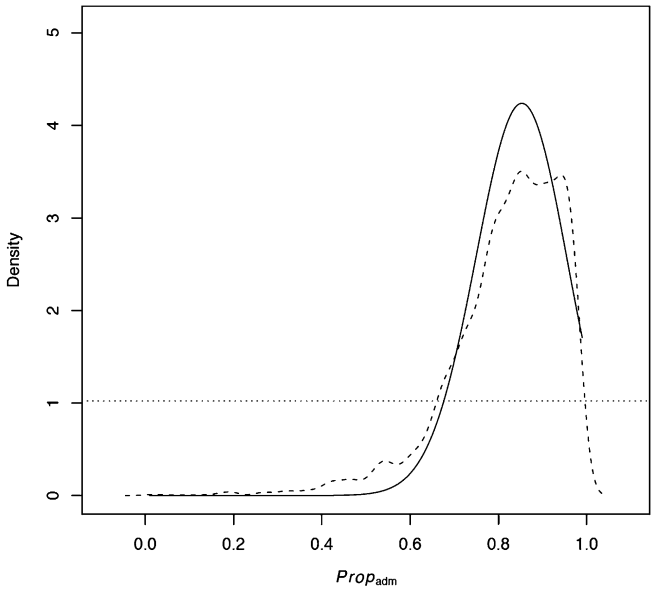

Figure 4 Probability density of the proportion of European admixture based on the regression method (solid line) and rejection method (dashed line). The horizontal dotted line represents the uniform prior distribution.

(see *Discussion*). Model C was also able to predict autosomal summary statistics quite accurately (Figure S13 and Figure S14) even under the simplified assumption of equal sex ratios in all populations (see *Material and Methods*).

Discussion

The demography of the Zimbabwe population was modeled in several studies as a simple expansion process (Glinka *et al.* 2003; Ometto *et al.* 2005; Laurent *et al.* 2011). However, it is still unclear if the Zimbabwe population is the source from which all other *D. melanogaster* populations derive. Based on this scenario we may expect that a bottleneck model would fit the data of the Zimbabwe population better than expansion, since range expansions are usually associated with bottlenecks and founder effects (Excoffier *et al.* 2009). Indeed, what we find here is exactly that pattern: the bottleneck model is significantly preferred over the expansion model.

The predictive simulations of models A, B, and C show that all models are able to explain the diversity observed in Africa and Europe (Figure S8 and Figure S9). However, only model C (including admixture) is able to fully explain the diversity observed in North America. Model A involves a recent foundation of North America from Europe, but North America shows currently greater diversity. This is hard to explain without considering an input from the ancestral population. Model B provides this input from Africa through migration, but to be able to reach the levels of diversity observed in North America we would need unrealistically high rates of migration. However, this would not be compatible with the observed values of population differentiation. Model C is in accordance with the observed data in this aspect. Another aspect that favors the admixture model over the others is that the values of Tajima's *D* in North America and Europe can also be better explained. We do not have an intuitive explanation why a recent admixture event has an influence on Tajima's *D* in one of the parental populations (*i.e.*, the European one).

Among all tested models (Table S2), we selected models A, B, and C for two main reasons. First, there is evidence that North American *D. melanogaster* has been introduced from Europe (see Introduction) and we have strong biological reasons to believe that North American diversity was generated through admixture and/or migration with African populations. Second, we wanted to keep the models as simple as possible. When we examined the data we observed that the North American population shares polymorphisms mostly with the European population and, to a lesser extent, with the African population. This observation fits the hypothesis of a European contribution. A model in which North America is derived from the African population without any European contribution would not be able to explain the shared polymorphism between North America and Europe in the observed data.

In addition to the three main demographic models (*i.e.*, models A, B, and C), we examined two more models in which the North American population derives directly from the African one. This alternative topology of the population's genealogy was tested without migration (model D, Table S2) and considering a simple migration process, identical to the one used in model B (model E, Table S2). These models represent possible alternative explanations for the high diversity harbored by the North American population. However, when compared to model C, models D and E are less supported by the data set as indicated by their associated posterior probabilities (>0.999, <0.001, and <0.001, respectively).

We note here that our modeling of the dispersal patterns between worldwide populations of *D. melanogaster* is a crude simplification of the real, but unknown migratory processes characterizing this species. It is well possible that more complex demographic models allowing specific, and potentially asymmetric, migration rates between all pairs of populations might be a more accurate representation of reality. However, in our case, these more sophisticated models have the property of having divergence time and specific migration rates as free parameters for several pairs of populations. A recent simulation study showed that the joint estimation of these two parameters in an ABC framework does not yield satisfying results (Tellier *et al.* 2011). Indeed, it is not clear at the present time which summarization of the raw data set would allow for an accurate joint estimation of divergence times and migration rates within an ABC framework. Although more work is needed to develop methods that allow for the estimation of more complex models, the analysis presented in this study shows that the history of the North American population is well characterized by an admixture of alleles coming from European and African populations.

The admixture model C can predict most of the observed summary statistics and JSFS equally well or better than the other models, except for the observed population differentiation (distance of Nei) between Africa and North America, which is better explained by model A or B (Figure S8). This higher simulated population differentiation in model A or B is associated with lower values of diversity in North America than the observed one, which is still a drawback for these models. We investigated this fact by adding more parameters to the model. We tested three variations of model C: model C1 has an extra bottleneck during the colonization of North America from Africa, model C2 has an extra bottleneck during the colonization of North America from Europe, and model C3 has both bottlenecks (Figure S15). While including the additional bottlenecks can account for the observed population differentiation they also reduce diversity below the observed values. Therefore, when compared to the original admixture model, models C1, C2, and C3 were not favored.

Another possible model in which higher values of population differentiation could be expected is a scenario in which samples are considered to be taken from demes in a metapopulation. If we have samples from different demes from different populations we may not expect migration or admixture to take place equally between all sampled demes, which may lead to higher values of population differentiation. Even though population differentiation in African populations is minimal (Yukilevich *et al.* 2010) this hypothesis still needs to be investigated further, with additional analyses of populations from Africa, Europe or North America, which is beyond the scope of this study.

To obtain further insight into the demography of the Zimbabwe population, we compared the SFS of this population with that predicted under a bottleneck. Regarding the input parameters for this prediction we used the modes (as point estimates) of the posterior distributions that were generated by the ABC regression step (see Table 4). Figure 2 shows the observed SFS compared to the predicted SFS under the conditions described above. Li and Stephan (2006) fitted a population expansion model to this same observed SFS (Figure 3 of Li and Stephan 2006). The bottleneck model in our study fits the intermediate-frequency classes better, whereas the population expansion model is more compatible with the classes of the singletons and doubletons. However, for the high-frequency variants both models show relatively large discrepancies. According to Li and Stephan (2006), this may indicate evidence for positive selection, a hypothesis that needs to be further tested. An alternative explanation for the excess of high-frequency variants may be ancestral state misidentification (Hernandez *et al.* 2007). Note that ancestral misidentification does not change our main ABC results, since the summary statistics used (including the folded SFS) are unaffected by polarization.

Although our modeling approach takes into account the combined effects of mutation, genetic drift, and migration we point out that we did not consider any form of natural selection in this analysis. This omission does not reflect that we believe that the impact of selection is minimal in our data set but rather the lack of available methods to estimate demographic and selective forces simultaneously. We think that such methods would greatly improve the interpretation of data sets like the one we present here, since several studies recently reported evidence that, contrary to previous beliefs, negative and positive selection have a substantial impact on the genetic variation harbored by natural populations of *D. melanogaster* (Macpherson *et al.* 2007; Jensen *et al.* 2008). Until such methods are available it is hard to predict to what extent the results presented in this study are affected by a reduction of the evolutionary history of *D. melanogaster* to a strictly neutral nonequilibrium model.

Nonetheless, we stress that the main result of this study, which is the identification of a substantial contribution of the African gene pool to the North American population, cannot be invalidated by including selection in our analysis. The reason for this is that the above-mentioned result relies on the observation that the level of genetic diversity found in the North American population is too high compared to expectations under a model in which the North American population would derive exclusively from the European one.

In conclusion, this study generated the first joint demographic analysis of African, European and North American populations of *D. melanogaster*. We analyzed the African population and found that a bottleneck fits the observed data better than an instantaneous population expansion. Regarding the North American population, we found that an admixture model fits the observed data significantly better than models involving colonization only from Europe or migration. We estimated the population parameters of all populations, from which we highlight the time of split between Africa and Europe (~19,000 years ago) and the proportion of European and African ancestry in the North American population (85% and 15%, respectively). The time of colonization of North America was given a very small prior because we know it took place ~200 years ago. In general, having described such a demographic model for North America, Africa, and Europe will be of valuable importance when looking for signatures of adaptation in any of these populations.

Acknowledgments

We thank Laurent Excoffier for his scientific advice on ABC and the modeling of population range expansions. We are also grateful to two reviewers for their valuable comments. This project has been funded by the Research Unit 1078 of the German Research Foundation (grant STE 325/12 to W. S. and HU 1776/2 to S.H.) and the Volkswagen Foundation (grant I/84232 to D.Z.).

Literature Cited

Andolfatto, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18: 279–290.

Baudry, E., B. Viginier, and M. Veuille, 2004 Non-African populations of *Drosophila melanogaster* have a unique origin. Mol. Biol. Evol. 21: 1482–1491.

Beaumont, M. A., W. Zhang, and D. J. Balding, 2002 Approximate Bayesian Computation in population genetics. Genetics 162: 2025–2035.

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365: 548–550.

Caracristi, G., and C. Schlötterer, 2003 Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol. Biol. Evol. 20: 792–799.

Csilléry, K., M. G. B. Blum, and O. François, 2012 abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3: 475–479.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. 4: 106–111.

Depaulis, F., and M. Veuille, 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15: 1788–1790.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Excoffier, L., M. Foll, and R. J. Petit, 2009 Genetic consequences of range expansions. Annu. Rev. Ecol. Evol. Syst. 40: 481–501.

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano *et al.*, 2007 Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. USA 104: 17614–17619.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. 48: 172–197.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165: 1269–1278.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Stoch. Models 14: 273–295.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto, 2005 Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15: 790–799.

Hedrick, P. W., 2011 *Genetics of Populations*. Jones & Bartlett, Boston.

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1792–1800.

Howard, L. O., 1900 A contribution to the study of the insect fauna of human excrement. Proc. Washington Acad. Sci. 2: 541–604.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Hutter, S., H. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan, 2007 Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177: 469–480.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–1410.

Jensen, J. D., K. R. Thornton, and P. Andolfatto, 2008 An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet. 4: e1000198.

Johnson, C. W., 1913 The distribution of some species of Drosophila. Psyche 20: 202–205.

Kauer, M., B. Zangerl, D. Dieringer, and C. Schlötterer, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics 160: 247–256.

Keller, A., 2007 *Drosophila melanogaster*'s history as a human commensal. Curr. Biol. 17: R77–R81.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics 146: 1197–1206.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22: 159–225.

Laurent, S. J., A. Werzner, L. Excoffier, and W. Stephan, 2011 Approximate Bayesian analysis of *Drosophila melanogaster* polymorphism data reveals a recent colonization of Southeast Asia. Mol. Biol. Evol. 28: 2041–2051.

Li, H., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet. 2: 1580–1589.

Lintner, J. A., 1882 *First Annual Report on the Injurious and Other Insects of the State of New York*. Weed, Parsons, Albany, NY.

Macpherson, J. M., G. Sella, J. C. Davis, and D. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. Genetics 117: 2083–2099.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269–5273.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. 15: 1566–1575.

Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. 22: 2119–2130.

Pavlidis, P., S. Hutter, and W. Stephan, 2008 A population genomic approach to map recent positive selection in model species. Mol. Ecol. 17: 3585–3598.

Pavlidis, P., J. D. Jensen, and W. Stephan, 2010a Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907–922.

Pavlidis, P., S. Laurent, and W. Stephan, 2010b msABC: a modification of Hudson's *ms* to facilitate multi-locus ABC analysis. Mol. Ecol. Res. 10: 723–727.

Pool, J. E., and C. F. Aquadro, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics 174: 915–929.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999 Population growth of human Y chromosome: a study of Y chromosome microsatellites. Mol. Biol. Evol. 16: 1791–1798.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science 312: 1614–1620.

Stephan, W., and H. Li, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98: 65–68.

Sturtevant, A. H., 1920 Genetic studies on *Drosophila simulans*. I. Introduction: hybrids with *Drosophila melanogaster*. Genetics 5: 488–500.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.

Tellier, A., P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. E. Rose *et al.*, 2011 Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PLoS ONE 6: e18155.

Wakeley, J., and J. Hey, 1997 Estimating ancestral population parameters. Genetics 145: 847–855.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wegmann, D., C. Leuenberger, and L. Excoffier, 2009 Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. Genetics 182: 1207–1218.

Yukilevich, R., T. L. Turner, F. Aoki, S. V. Nuzhdin, and J. R. True, 2010 Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. Genetics 186: 219–239.

Zayed, A., and C. W. Whitfield, 2008 A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. Proc. Natl. Acad. Sci. USA 105: 3421–3426.

Živković, D., and T. Wiehe, 2008 Second-order moments of segregating sites under variable population size. Genetics 180: 341–357.

Živković, D., and W. Stephan, 2011 Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. Theor. Popul. Biol. 79: 184–191.

Communicating editor: D. Begun

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145912/-/DC1

Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population

Pablo Duchen, Daniel Živković, Stephan Hutter, Wolfgang Stephan, and Stefan Laurent

DOI: 10.1534/genetics.112.145912

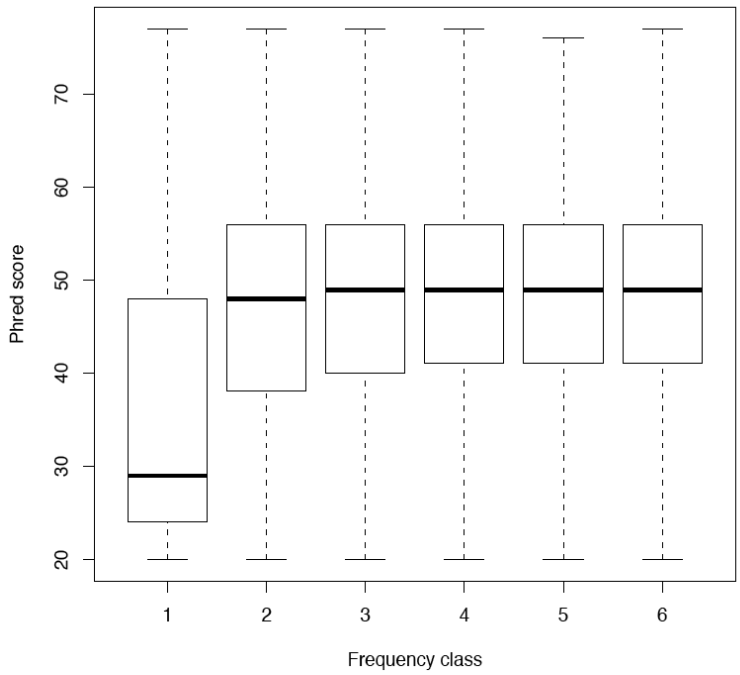

Figure S1 Phred quality scores of individual base calls belonging to the first six classes of the site frequency spectrum in the North American population (calculated from the DPGP1 raw fastq files). The middle bands indicate the median values, the boxes the upper and lower quartiles and the whiskers the minimum and maximum values.

<table>
<tr><td></td><td>0</td><td>1</td><td>2</td><td>...</td><td>n_2-2</td><td>n_2-1</td><td>n_2</td></tr>
<tr><td>0</td><td>X</td><td colspan="5">W2</td><td>W3</td></tr>
<tr><td>1</td><td rowspan="5">W1</td><td colspan="5" rowspan="5">W4</td><td rowspan="5">W1</td></tr>
<tr><td>2</td></tr>
<tr><td>...</td></tr>
<tr><td>n_1-2</td></tr>
<tr><td>n_1-1</td></tr>
<tr><td>n_1</td><td>W3</td><td colspan="5">W2</td><td>X</td></tr>
</table>

Figure S2 Joint Site Frequency Spectrum (JSFS) classes, according to the Wakeley-Hey model. On left most column we have the sample size n_1 of population 1. On the upper most row we have the sample size n_2 of population 2. The summary statistics proposed by Wakeley-Hey (1997) are represented by the letters W1 to W4.

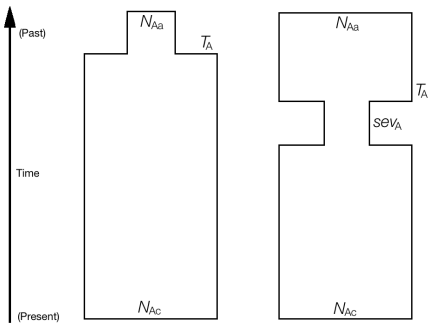

Figure S3 Population expansion (left) versus Bottleneck (right) model in Africa. The posterior probability of the Expansion model is 0.013. The posterior probability of the Bottleneck model is 0.987. Parameters are explained in Table 3.

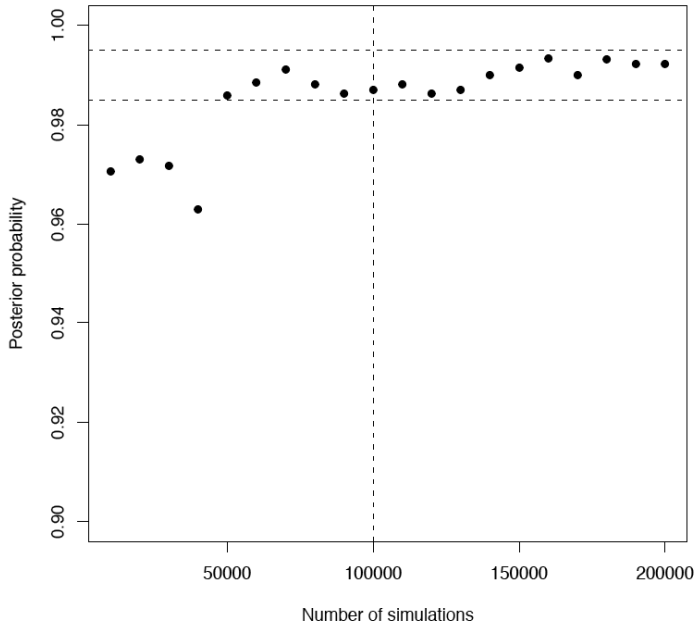

Figure S4 Behavior of the posterior probabilities of the Bottleneck model for different numbers of simulations. In the case of the Admixture model (model C) the posterior probability is always above 0.999 for different numbers of simulations.

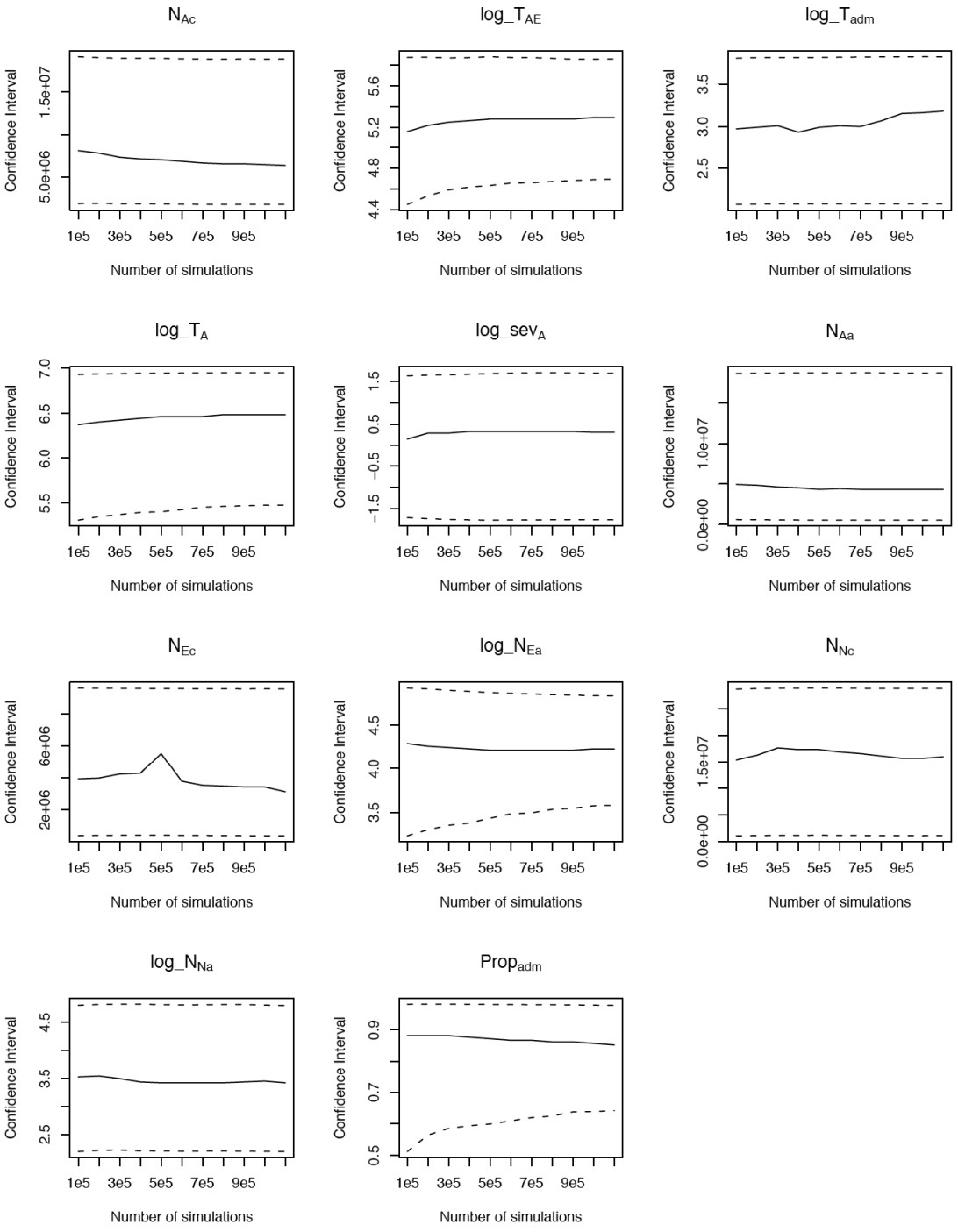

Figure S5 Behavior of the modes and 95% confidence intervals of the estimates of the parameters of the Admixture model (model C) for different numbers of simulations. Solid line: mode, dashed lines: upper and lower confidence intervals.

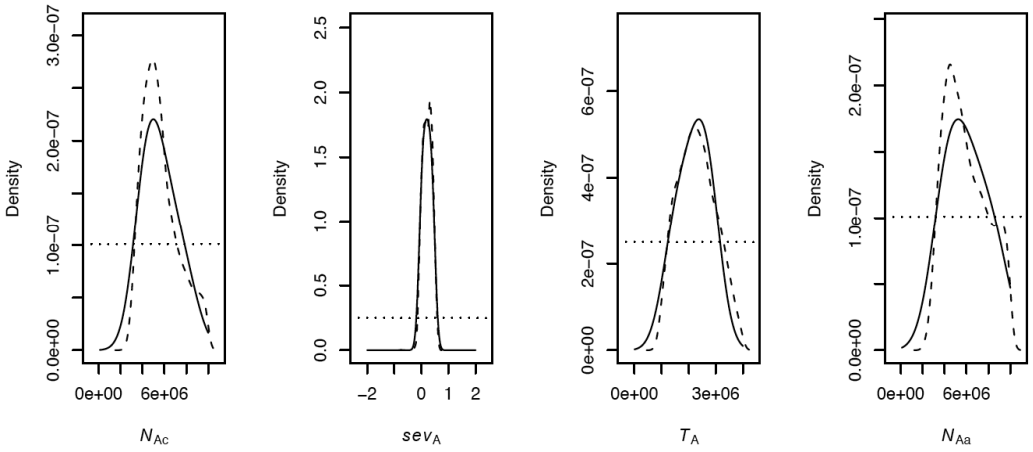

Figure S6 Posteriors of the Bottleneck model in Africa. Posteriors are represented by the rejection method (dashed line) and the regression method (solid line). Parameter abbreviations are explained in Table 3. Mode and confidence interval for each parameter are shown in Table 4.

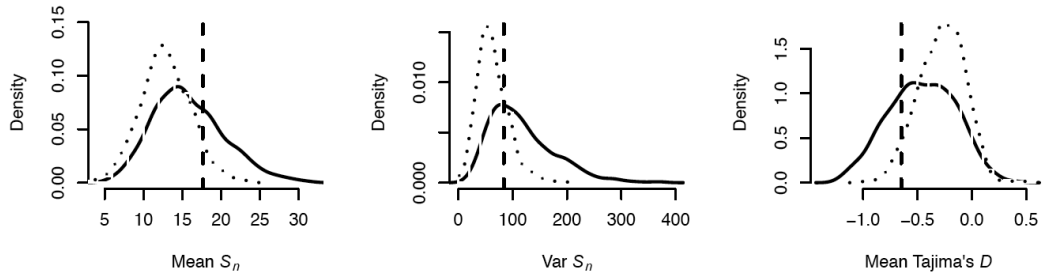

Figure S7 Predictions of the Bottleneck versus Population Expansion in Africa. Solid line: Bottleneck, dotted line: Population expansion, vertical dashed line: observed value. Parameters for predictive simulations are drawn from the posterior distributions generated by the regression method (see Materials and Methods).

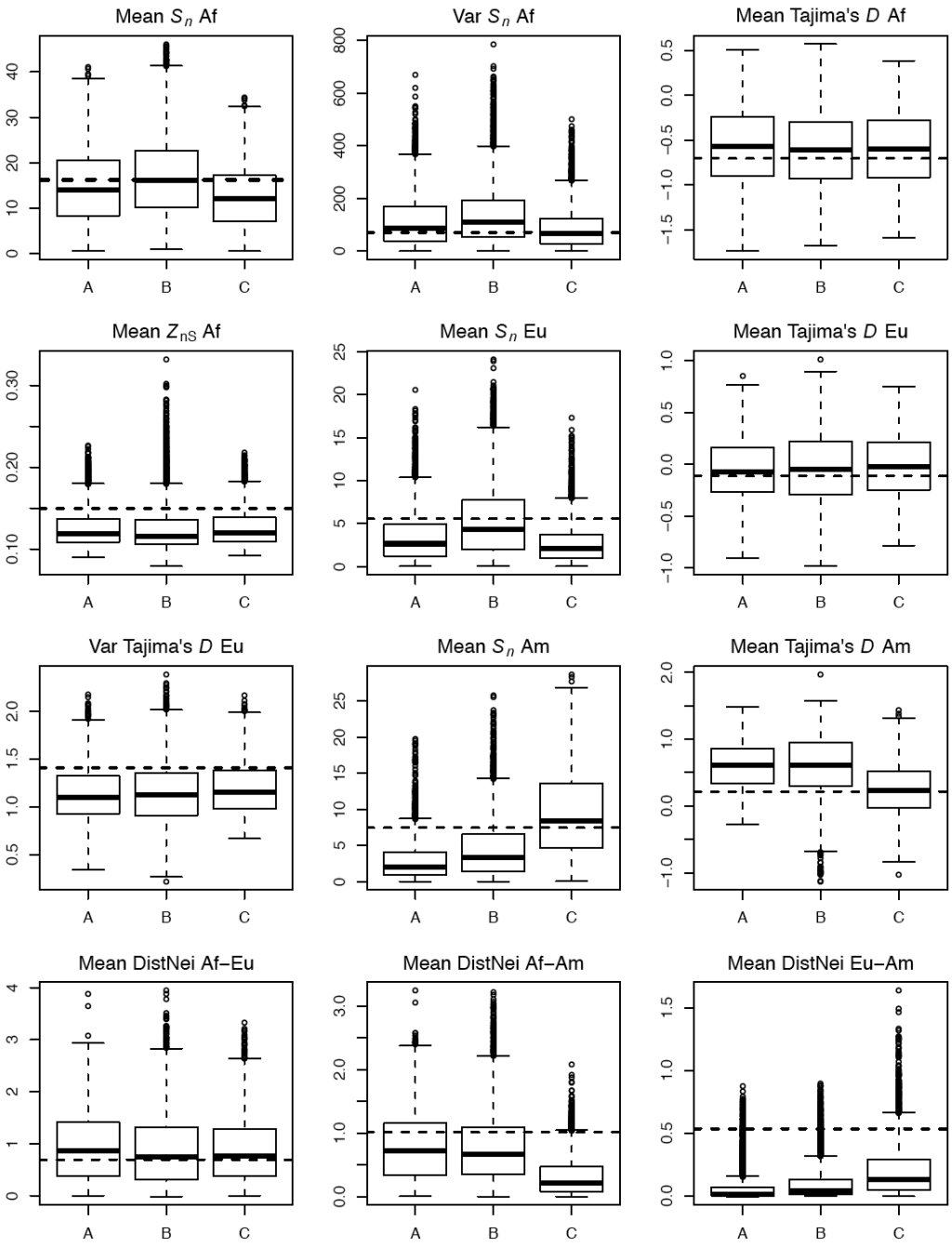

Figure S8 Predictions of summary statistics for models A, B and C based on the rejection method. The horizontal dashed line represents the observed value.

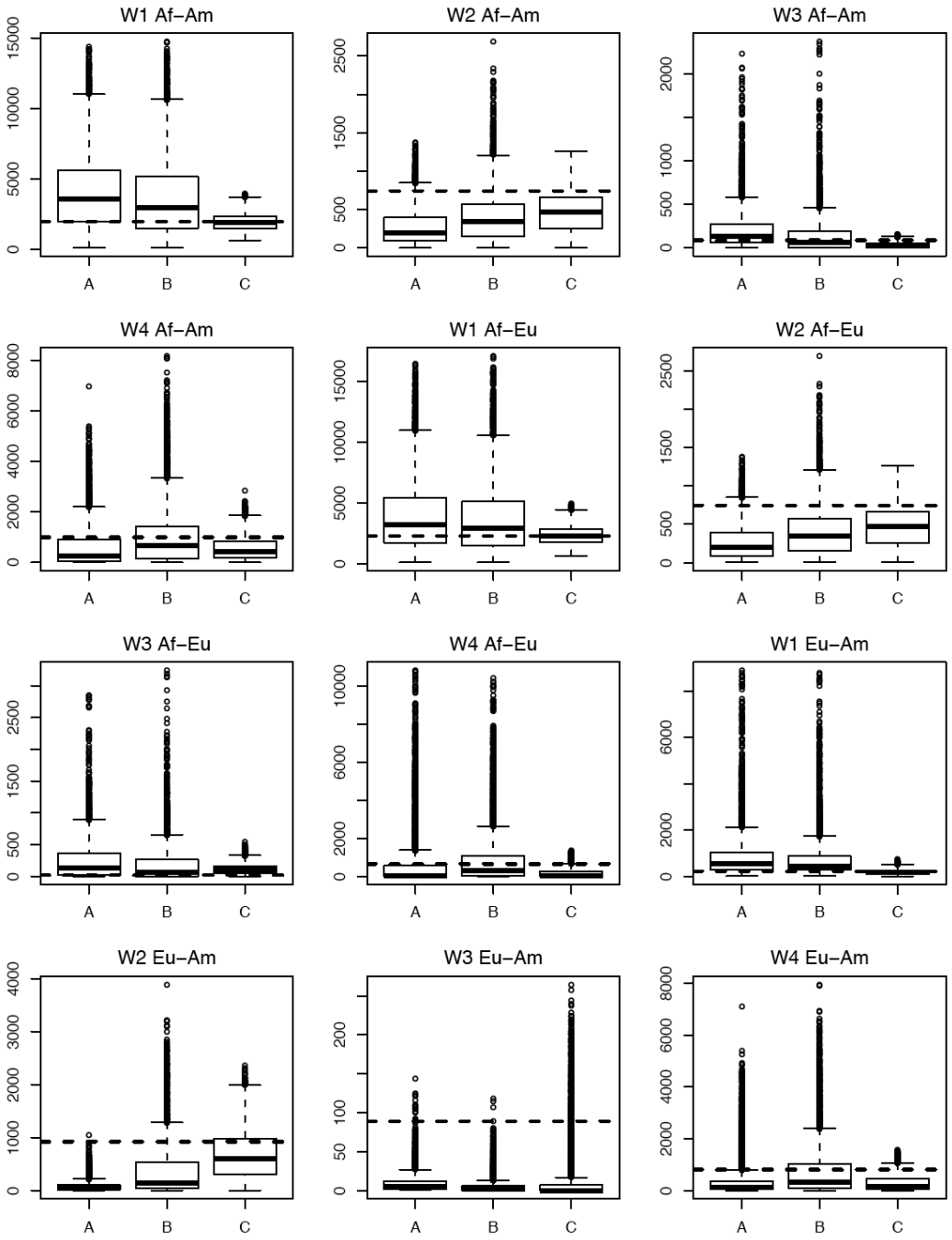

Figure S9 Predictions of the JSFS for models A, B and C based on the rejection method. The horizontal dashed line represents the observed value.

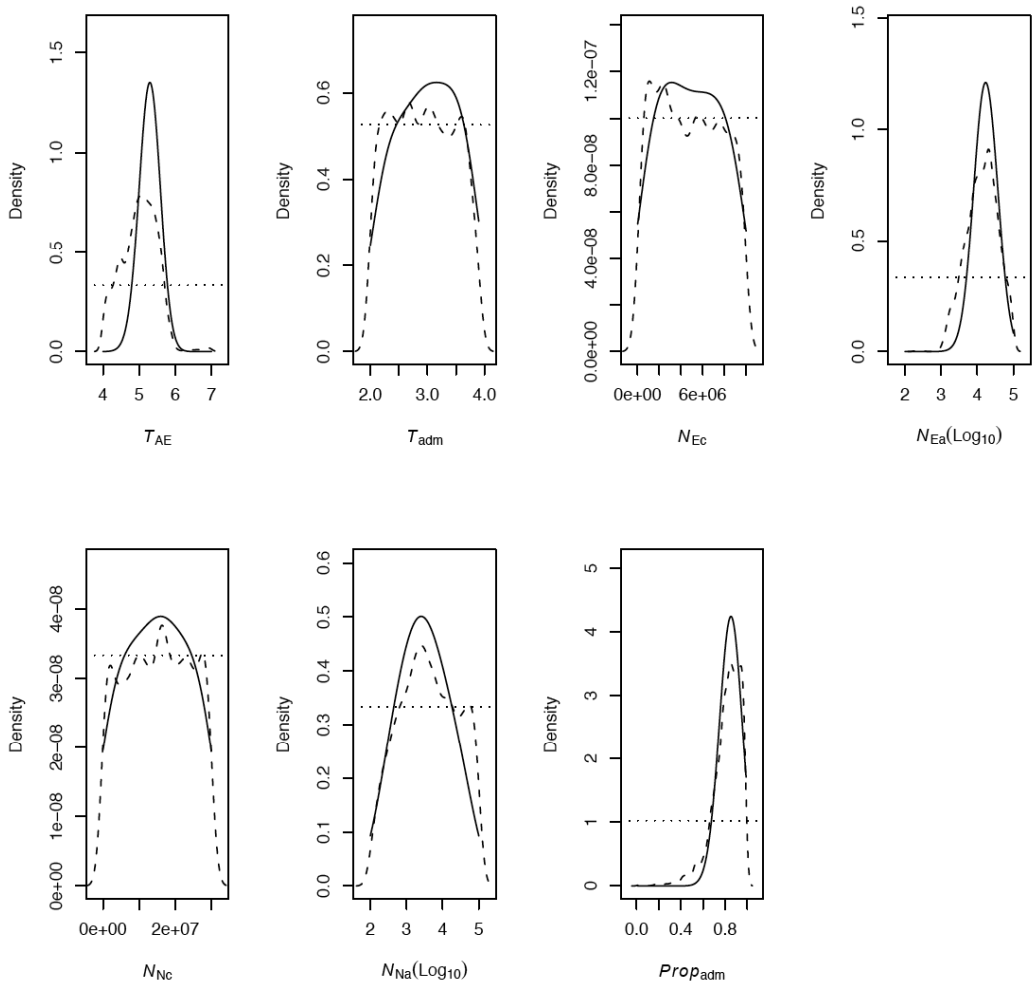

Figure S10 Posteriors of the Admixture model C. Posteriors are represented by the rejection method (dashed line) and the regression method (solid line). Parameter abbreviations are explained in Table 3. Mode and confidence interval for each parameter are shown in Table 5.

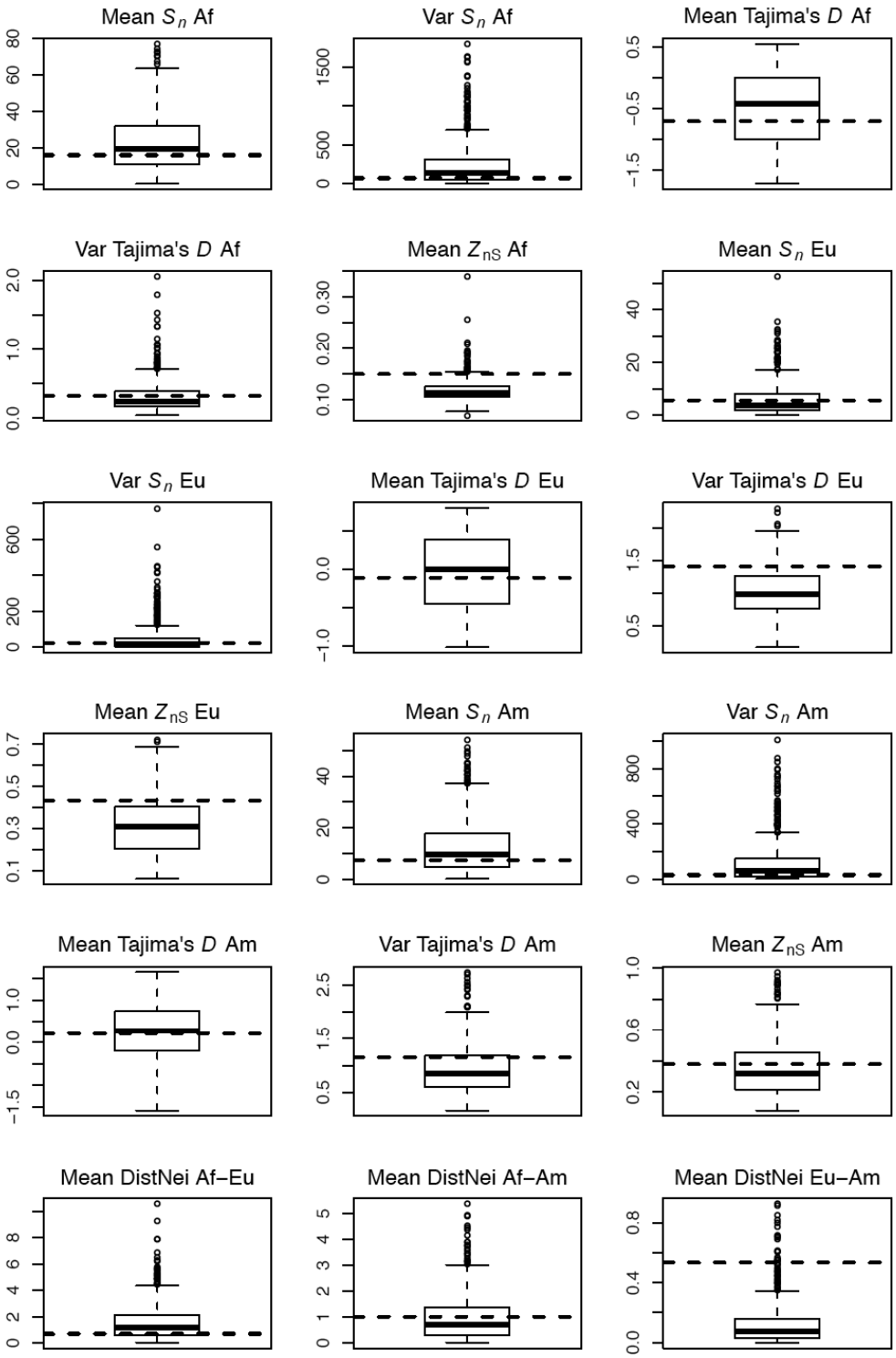

Figure S11 Predicted statistics of model C. Predictions of the mean and variance of S_n, mean and variance of Tajima's D and mean Z_{nS} are shown for each population. Predicted mean Distance of Nei for all pairs of populations are shown as well. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

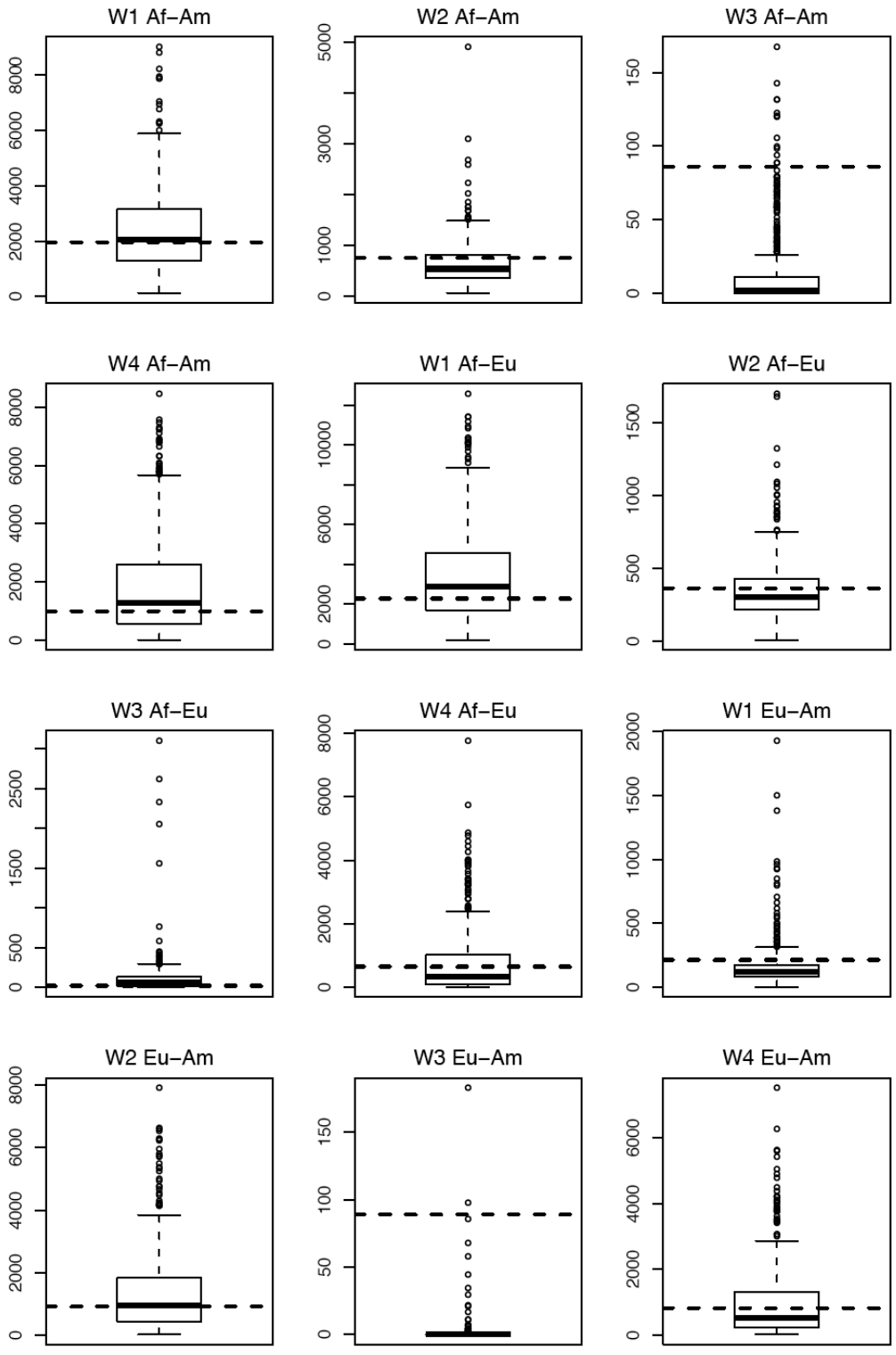

Figure S12 Predicted JSFS of model C. Predictions of each Wakeley-Hey (1997) class are shown. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

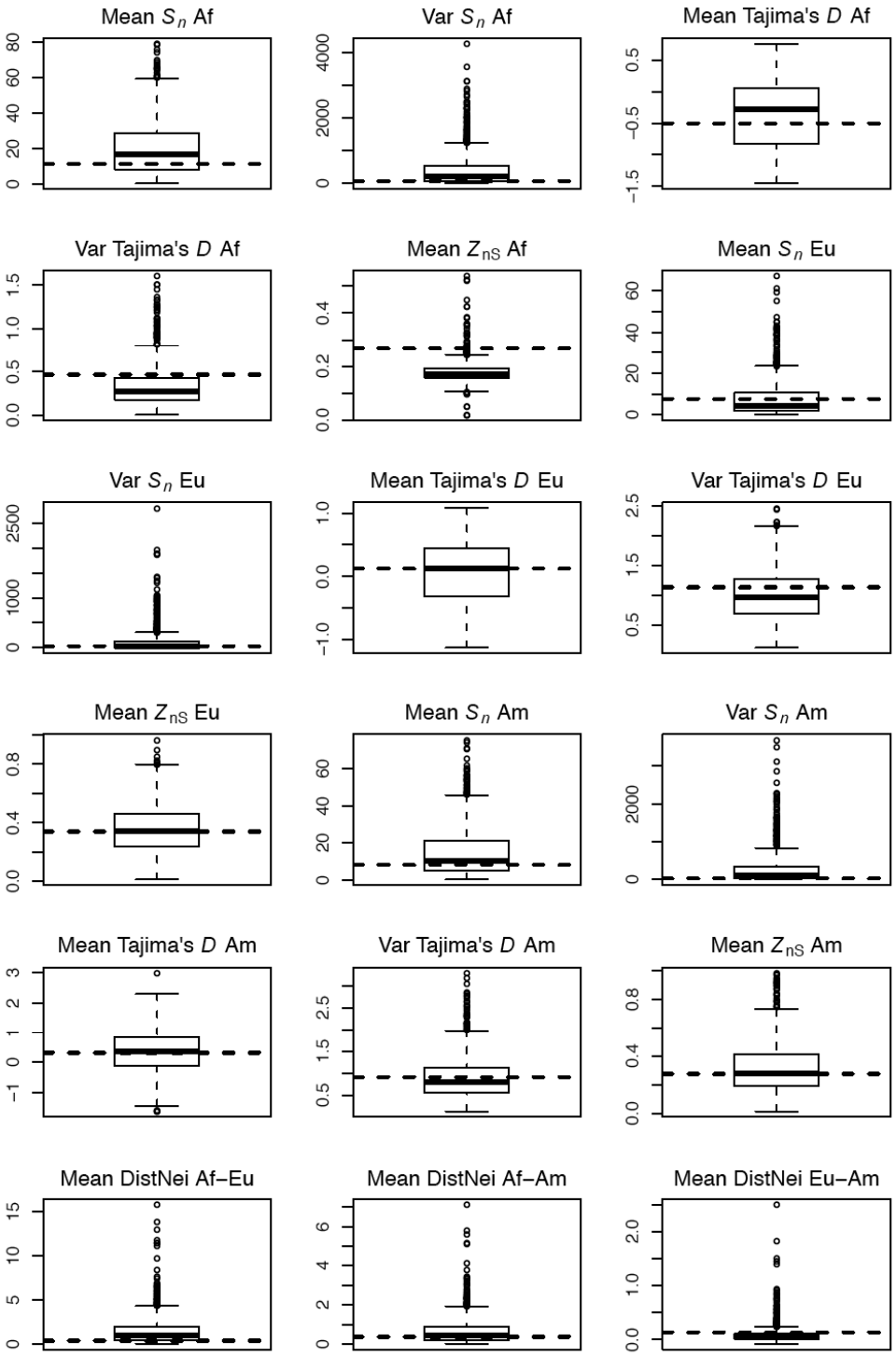

Figure S13 Predicted statistics of model C for autosomal loci (chromosome 3). Predictions of the mean and variance of S_n, mean and variance of Tajima's D and mean Z_{nS} are shown for each population. Predicted mean Distance of Nei for all pairs of populations are shown as well. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

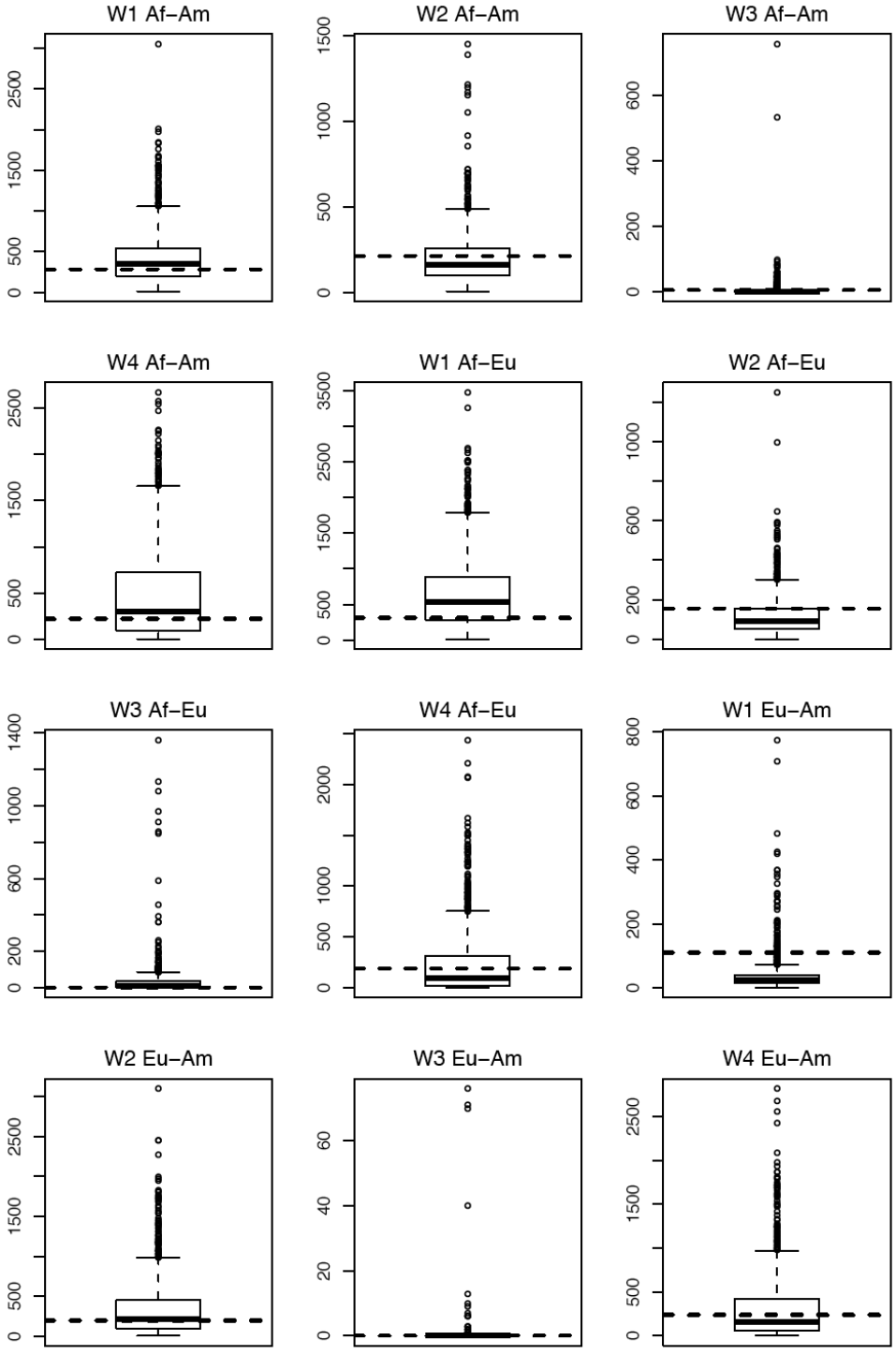

Figure S14 Predicted JSFS of model C for autosomal data (chromosome 3). Predictions of each Wakeley-Hey (1997) class are shown. Statistics are predicted by sampling parameters from the posterior distributions based on the regression method (see main text for details).

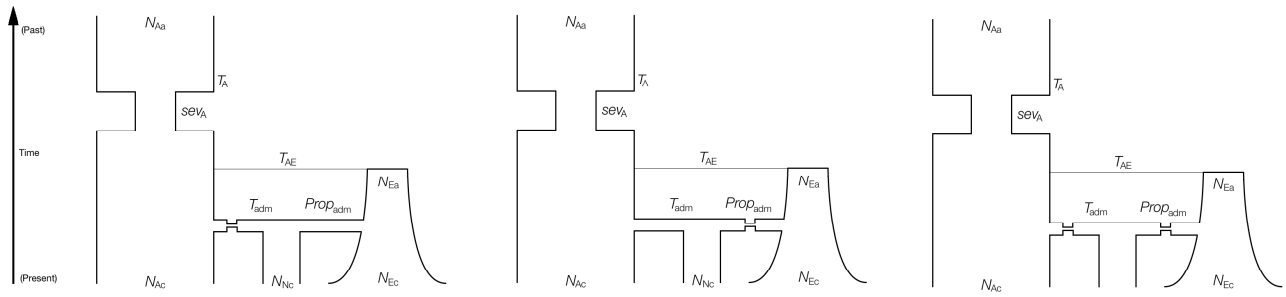

Figure S15 Models C1 (left), C2 (middle) and C3 (right).

Table S1 Parameters and priors used in the one-population models and in models A, B, C, D and E.

Parameter	Prior	Model
Current size Africa: N_{Ac}	unif($1x10^5$,$1x10^7$)	Bottleneck and Expansion
Time of bottleneck Africa: T_A	unif($1x10^2$,$4x10^5$)	Bottleneck and Expansion
Ancient size Africa: N_{Aa}	unif($1x10^5$,$1x10^7$)	Bottleneck and Expansion
Severity of bottleneck Africa: sev_A (decimal log)	unif(-2,2)	Bottleneck
Time of split Africa-Europe (decimal log): T_{AE}	unif(4,7)	Model A,B,C,D,E
Time of split Europe-North America (decimal log): T_{EN}	unif(4,7)	Model A,B
Time of split Africa-North America (decimal log): T_{AN}	unif(4,7)	Model D,E
Time of admixture (decimal log): T_{adm}	unif(2,4)	Model C
Current size Europe: N_{Ec}	unif($1x10^4$,$1x10^7$)	Model A,B,C,D,E
Ancient size Europe (decimal log): N_{Ea}	unif(2,5)	Model A,B,C,D,E
Current size North America: N_{Nc}	unif($1x10^4$,$3x10^7$)	Model A,B,C,D,E
Ancient size North America (decimal log): N_{Na}	unif(2,5)	Model A,B,C,D,E
Proportion of European admixture: $Prop_{adm}$	unif(0.01,0.99)	Model C
Migration rate (decimal log): M	unif(-10,-2)	Model B,E

Table S2 Three-population models covered in this study.

Model	Description	Posterior Probability
A	"No migration" model. Comprises Africa as the ancestral population, colonization of Europe followed by exponential growth, and the colonization from Europe to North America with subsequent exponential growth.	< 0.001
B	"Migration" model, matches Model A but adds an equal migration rate between all populations starting at the colonization time of North America.	< 0.001
C	"Admixture" model, equals the previous models until the North American population is founded through an admixture between Africa and Europe followed by exponential growth in North America.	> 0.999
D	"No migration II" model, North America and Europe split independently from Africa, no migration.	< 0.001
E	"Migration II" model, same as model D plus one single rate of migration starting when the North American population is founded.	< 0.001

Table S3 Mean squared error (MSE) of the (log_{10}) parameter estimates of model C for varying numbers of simulations.

	100000	200000	300000	400000	500000	600000	700000	800000	900000	1000000
N_{Ac}	0.019	0.0101	0.00718	0.00574	0.00443	0.00296	0.00201	0.00149	0.00154	0.00125
T_{adm}	0.242	0.293	0.275	0.306	0.335	0.291	0.309	0.305	0.326	0.322
T_{AE}	0.0693	0.0388	0.0271	0.0214	0.018	0.0147	0.0128	0.011	0.00996	0.00927
T_{A}	0.0447	0.0498	0.043	0.0407	0.0352	0.0318	0.0317	0.025	0.0233	0.0203
sev_{A}	0.0178	0.03	0.03	0.0307	0.03	0.0311	0.0298	0.0291	0.03	0.0326
N_{Aa}	0.00114	0.00243	0.00422	0.00464	0.00661	0.00688	0.00767	0.00835	0.00869	0.00871
N_{Ec}	0.0221	0.0839	0.0831	0.111	0.0804	0.0818	0.069	0.0658	0.0434	0.0366
N_{Nc}	0.000554	0.000369	0.00059	0.000937	0.00103	0.000636	0.00054	0.000316	0.000336	0.000402
N_{Ea}	0.00605	0.00624	0.0075	0.00801	0.0086	0.00942	0.0104	0.011	0.0118	0.0123
N_{Na}	0.471	0.534	0.514	0.505	0.457	0.444	0.443	0.467	0.517	0.509
$Prop_{adm}$	0.00148	0.00149	0.00169	0.00196	0.00213	0.00222	0.00222	0.00221	0.00219	0.00214

Chapter 2: Estimates of divergence time and migration rate between African and European populations of *Drosophila melanogaster*: an approach based on Approximate Bayesian Computation

Estimates of divergence time and migration rate between African and European populations of *Drosophila melanogaster*: an approach based on Approximate Bayesian Computation

Pablo Duchén, Stefan Laurent, Wolfgang Stephan

Abstract

Populations differentiate from each other in the presence of natural selection, genetic drift and gene flow, but these forces do not contribute equally to differentiation. Genetic drift is stronger in smaller populations and selection is stronger in bigger ones. Gene flow, however, may be able to overcome the effects of selection and population size, which highlights the importance of migration in population differentiation. *Drosophila melanogaster* is a perfect study system for migration given its worldwide distribution. Interestingly, very little is known about the actual amount of gene flow among *D. melanogaster* populations, although the existence of migration in this species is well acknowledged. In this study we use Next Generation Sequencing (NGS) data together with Approximate Bayesian Computation (ABC) methods to estimate migration rates and divergence times in African (Rwanda) and European (France) populations of *D. melanogaster*. We compared three models: no migration, symmetrical migration, and asymmetrical migration, the last one showing the highest posterior probability. We found that the split between these two populations is similar to previous reports on other African samples, ranging between 10,000 to 30,000 years ago. We also found that the migration rate from Africa to Europe is slightly lower than the migration rate from Europe to Africa, and that these migra-

tion rates ($Nm \sim 10$) are higher than previous reports ($Nm = 2$). Overall, there is evidence of an overall increase of gene flow in the last 30 years, possibly associated with an increase in human migration in this period of time.

Introduction

Gene flow or migration, defined as the movement of genes from one population to another, affects significantly the differentiation between populations. While natural selection and genetic drift may increase differentiation, migration reduces this effect by bringing gene pools back together. Even when these evolutionary forces act together the contribution of each force will vary with population size. For instance, genetic drift plays a major role in small populations while selection becomes more effective in larger populations. Gene flow, on the other hand, may be able to overcome the effects of both population size and selection strength. Haldane (1930) showed that migration exceeds the effect of selection if the fraction m/s is bigger than 1 (where m and s are the migration rate and selection coefficient, respectively). Conversely, Wright (1931) showed that two populations will not diverge if the product Nm of population size and migration rate is bigger than 1. These two examples show how gene flow significantly affects other evolutionary forces and highlights the importance of studying and quantifying migration patterns in several species.

Among the species capable of migrating the fruit fly *Drosophila melanogaster* is one of the most successful colonizers. A proof of this is the worldwide distribution of this species, with latitudes ranging from Tasmania (Agis and Schlötterer, 2001) to Finland (Hackman, 1954) or Sweden (Bächli et al., 2005), and altitudes ranging from sea level up to more than 3000 m (personal observation). It becomes clear that migration has been a key factor to explain the current distribution of *D. melanogaster* from its origin in sub-Saharan Africa (Lachaise et al., 1988) to its current range. After the origin of this species several population splits took place

not only in Africa but also out of Africa (Stephan and Li, 2007), starting with the split between Africa and Europe some 19,000 years ago (Duchen et al., 2013), Europe and Asia some 2,500 years ago (Laurent et al., 2011), or recent colonization of North America around 200 years ago (Johnson, 1913; Sturtevant, 1920; Keller, 2007). At the beginning, migration was most likely limited to the fly's intrinsic capabilities of moving around, but then it increased when *D. melanogaster* gradually became a human commensal (Lachaise and Silvain, 2004). At some point, human-mediated migration became significant at a large scale when, a few hundred years ago, agriculture-associated trade between Africa, Europe, Asia and America became frequent and well established.

Although the existence of migration in *D. melanogaster* is nowadays well acknowledged (David and Capy, 1988; Begun and Aquadro, 1993; Glinka et al., 2003; Haddrill et al., 2005; Ometto et al., 2005) there are very few studies that quantified the actual amount of gene flow in this species. The first study that attempted to measure migration dates back to 1966 when Wallace (1966, 1970) studied the dispersal of *D. melanogaster* in a tropical population from Colombia. Coyne and Milstead (1987) studied the dispersal capabilities of *D. melanogaster* in a Maryland orchard by release/capture and found that it alone can disperse several kilometers per day. All in all, these studies analyzed migration only locally, that is, migration in a small area surrounding a location of release. So far, the only study that quantified migration rates in a larger scale is the one by Singh and Rhomberg (1987). They used enzyme assays to calculate heterozygosity and used information on rare alleles to estimate migration rates between several populations distributed worldwide. They found that all populations had significant levels of migration ($Nm > 1$). Kennington et al. (2003) used microsatellite data to study asymmetrical migration along a cline in eastern Australia. They developed a method based on the proportion of private alleles to examine departures from symmetrical migration. They found

significant levels of gene flow among these populations but asymmetrical migration was uncommon.

The main goal of the present study is to estimate migration rates in *D. melanogaster* at a large scale. We think that this species constitutes a perfect model system for this purpose given its rich history of dispersal and colonization and the large amount of full-genome sequences available for populations in Africa, Europe and North America. The great advantage of having next-generation sequencing (NGS) data at hand is the type of information that can be extracted from it, including genes, intergenic loci, introns, silent sites, etc. Population parameter estimation then follows with methods such as Approximate Bayesian Computation (ABC), which can be readily used with NGS data. Other advantages of ABC include: 1) it can be used with demographic models with any degree of complexity, 2) it generates confidence intervals for each population parameter estimate (as expected from Bayesian methods), and 3) it allows for the inclusion of recombination. Here, we will model the demography of African and European populations and use ABC to estimate split times and migration rates, among other population parameters. With this study we want to contribute to the current knowledge of gene flow and history of *D. melanogaster* by making use of state of the art sequencing technology and parameter estimation methods.

Materials and Methods

The analysis presented here comprises three main parts. First, we designed demographic models suitable for one African and one European population allowing for migration between them (Figure 1). Second, we analyzed the performance of ABC when estimating migration rates and split times jointly. For this purpose we used simulated data sets with known population parameters. Finally, we applied ABC to jointly estimate migration rate and split time between actual sequences from Africa

and Europe. We used available sequences from Rwanda (Gikongoro) and France (Lyon).

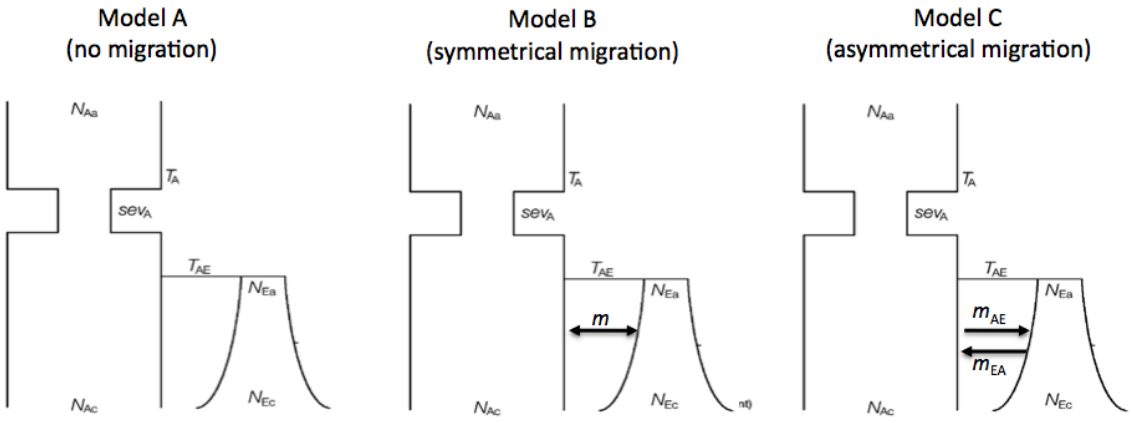

Figure 1: Two-population demographic models. Descriptions of each model are presented in Table 2.

Demographic model

The basic demographic model includes Africa as the ancestral population and the colonization of Europe followed by exponential growth (Figure 1). We modeled the ancestral population as a bottleneck in order to reflect the founding event produced when this population was colonized from the *D. melanogaster* center of origin (Duchen et al., 2013). Three variations of this model were then analyzed: a model without migration (model A), a model with symmetrical migration (model B), and a model with asymmetrical migration between the two populations (model C). Models A, B, and C had 7, 8, and 9 population parameters respectively. The onset of migration was given a prior between 1000 and 10000 generations in the past. Gene flow before that time was existent but not entirely human-mediated, and it was not as significant as it is nowadays.

Table 1: Parameters and priors used for each model.

Parameter	**Prior**	**Model**
Current size Africa: N_{Ac}	unif(1x10^6,1x10^7)	A,B,C
Time of bottleneck Africa: T_A ($\log_{10}$ generations)	unif(5,7)	A,B,C
Ancient size Africa: N_{Ac}	unif(1x10^4,1x10^7)	A,B,C
Severity of Bottleneck Africa: sev_A ($\log_{10}$)	unif(-2,2)	A,B,C
Time of split Africa-Europe: T_{AE} ($\log_{10}$ generations)	unif(4,7)	A,B,C
Current size Europe: N_{Ec}	unif(1x10^4,1x10^7)	A,B,C
Ancient size Europe: N_{Ea} ($\log_{10}$)	unif(1x10^4,1x10^7)	A,B,C
Migration rate (general): m	unif(-10,-2)	B
Migration rate Africa-Europe: m_{AE} ($\log_{10}$)	unif(-10,-2)	C
Migration rate Europe-Africa: m_{EA} ($\log_{10}$)	unif(-10,-2)	C

ABC simulations

We simulated 100,000 data sets for each of the models described above. Each simulated data set consisted of 88 loci with individual per locus samples sizes, as well as mutation and recombination rates identical to the ones found in the observed data set (see section SNP data). Mutation rates per locus were calculated taking into account the divergence from *Drosophila simulans*. Recombination rates per locus were calculated according to Fiston-Lavier et al. (2010). Our primary tool was the coalescent simulator ms by Hudson (2002). Each parameter was chosen from uniform prior distributions (Table 1). Missing nucleotides were also simulated at the same positions as they occur in the observed data. We accomplished this following the procedure developed by Duchen et al. (2013). From all these simulated loci we computed the mean and variance of the following summary statistics: the number of segregating sites S_n, Wattersons Θ_W (Watterson, 1975), the average number of pairwise differences in all pairwise comparisons of n sequences Π_n, Tajima's D (Tajima, 1989), the number of haplotypes K (Depaulis et al., 1998), the linkage disequilibrium statistic Z_{nS} (Kelly, 1997), and the distance of Nei as a measure of population differentiation (Nei and Li, 1979). We also computed the joint site

frequency spectrum (JSFS) of the simulated ancestral and derived populations.

Table 2: Description of the demographic models covered in this study.

Model	Description	Posterior Probability
A	"No migration" model. Comprises Africa as the ancestral population, colonization of Europe followed by exponential growth.	-
B	"Symmetric migration" model, matches Model A but adds and equal migration rate m between all populations starting at the colonization of Europe.	0.28
C	"Asymmetric migration" model, matches Model B but considers two different migration rates: m_{AE} from Africa to Europe and m_{EA} from Europe to Africa.	0.71

Performance analysis

In order to study how well we can predict divergence time and migration rates jointly we simulated data sets with known divergence (τ) and migration (m) parameters. Other parameters like population sizes, as well as mutation and recombination rates were fixed to known *Drosophila*-like values (Duchen et al., 2013). For model B we arbitrarily chose 3 parameter values for $\log_{10}(\tau)$: 4.5, 5.25, and 6.0, representing an early, intermediate, and old split, respectively. Regarding $\log_{10}(m)$ we also chose three parameter values: -4, -6, and -8, representing little, intermediate, and extensive migration rates, respectively. Combining the 3 values for τ, and the 3 values for m we had a total of 9 different parameter combinations. For each of these 9 parameter combinations we simulated 100 data sets using Hudson's *ms* (Hudson, 2002). For each data set we calculated summary statistics and the JSFS and re-estimated the parameter values. We reduced dimensionality using partial least squares as implemented in ABCtoolbox (Wegmann et al., 2010). We used a total of 25 linear combinations. We reported parameter estimation results by means of the

root mean square error (RMSE) and the relative bias. For model C we replaced the symmetrical migration rate m with asymmetrical migration rates m_{AE} and m_{EA}, representing migration from Africa to Europe and from Europe to Africa, respectively. With this extra migration rate we had a total of 27 parameter combinations, which were subject to the same procedure as in model B.

Single Nucleotide Polymorphism (SNP) data

Individuals come from two populations: Rwanda in Africa (sample size $n = 23$) and Lyon in France ($n = 8$). Sequence data consist of 89 intergenic X-linked loci from each population. These loci were extracted from full-genome sequences (Pool et al., 2012) (publicly available from the Drosophila Population Genomics Project at `http://www.dpgp.org`) that were created using Illumina next-generation sequencing (NGS) technology. Criteria for loci selection included: a) loci should be separated at least 50 kb from each other (to ensure independence); b) loci should be at least 1 kb away from any annotated gene, including UTR regions (to minimize the effect of linked selective pressure); c) loci should be at least 1kb long (to ensure a sufficient number of SNPs); and d) for every SNP the presence of missing nucleotides (N's) should not be greater than 20%. Additional quality control steps performed in Pool et al. (2012) were kept in the present analysis, including: e) masking of all bases with a Phred score lower than 31; f) masking of regions with identity by descent (IBD); and g) masking of admixed tracks (for details see Pool et al. (2012)). *Drosophila simulans* was used as an outgroup sequence.

Summary statistics and Joint Site Frequency Spectrum (JSFS)

From all these loci we computed the mean and variance of the following summary statistics: the number of segregating sites S_n, Wattersons Θ_W (Watterson, 1975), the average number of pairwise differences in all pairwise comparisons of n sequences

Π_n, Tajima's D (Tajima, 1989), the number of haplotypes K (Depaulis et al., 1998), the linkage disequilibrium statistic Z_{nS} (Kelly, 1997), and the distance of Nei as a measure of population differentiation (Nei and Li, 1979). We also computed the JSFS of the simulated ancestral and derived populations. This group of summary statistics, plus each class of the JSFS, constitutes our observed vector.

Results

Performance analysis

To determine if our ABC implementation is able to jointly estimate m and τ we performed simulations for several combinations of m and τ, for both models B and C (Tables 3 and 5, respectively). After visually inspecting the decay of RMSE we chose 25 pls components for all subsequent analyses. These 25 components explained most of the variance and minimized the noise. We found that migration and divergence can be estimated jointly but the accuracy of the estimation improves when divergence is older. This observation applies to both models B and C. By looking at Tables 3 and 4 it is noticeable how values of relative bias and RMSE become smaller when τ gets older. Tellier et al. (2011) made a similar observation when estimating m and τ for seed banks. In our performance analysis for models B and C divergence estimates are more accurate than migration estimates. Divergence times can be estimated very accurately even when divergence is young and these estimates will still improve when divergence gets older. Asymmetrical migration rates are accurately estimated only when $\log_{10}(\tau)$ is greater than 5.25. When divergence is too young accurate migration rates are difficult to obtain.

Table 3: Performance results for model B.

$\log_{10}\tau$	$\log_{10} m$	**Bias** $\log_{10}\tau$	**Bias** $\log_{10} m$	**RMSE** $\log_{10}\tau$	**RMSE** $\log_{10} m$
4.5	-4	-0.028797556	0.1801004	0.189792319	0.961388358
4.5	-6	-0.005146022	0.23363325	0.128939089	1.709185379
4.5	-8	-0.003035533	-0.075679688	0.134782238	1.18825125
5.25	-4	-0.009235486	0.0364829	0.170660393	0.364550418
5.25	-6	-0.023736495	0.254403583	0.178066797	1.788454849
5.25	-8	-0.022272019	-0.035579262	0.169645205	0.992506569
6.0	-4	0.007521433	0.01708605	0.112456019	0.25790577
6.0	-6	0.003526483	0.18807225	0.050214945	1.450178312
6.0	-8	0.0039284	0.0168325	0.050320174	0.444724435

Migration and divergence between Africa and Europe

After studying how well our ABC implementation performs we analyzed a real data set coming from Africa (Gikongoro, Rwanda) and Europe (Lyon, France). Model choice favors the model with asymmetrical migration rates (model C, Table 2). Still, since we made a performance analysis for models B and C we report the parameter estimates of these two models (Table 4). Divergence estimates between the populations of Rwanda and France are similar in models B and C, both yielding $\log_{10}(\tau) = 5.32$ and $\log_{10}(\tau) = 5.54$, respectively. These estimates are very similar to $\log_{10}(\tau) = 5.29$, the equivalent one reported by Duchen et al. (2013). Migration estimates vary between models B and C. Model B (symmetrical migration) estimates $log_{10}(m) = -5.18$, which corresponds to $m = 6.61 \times 10^{-6}$. If migration is rather asymmetrical, then we find that migration from Africa to Europe is less than migration from Europe to Africa: $\log_{10}(m_{AE}) = -5.98$, and $\log_{10}(m_{EA}) = -5.50$ (corresponding to $m_{AE} = 1.05 \times 10^{-6}$ versus $m_{EA} = 3.16 \times 10^{-6}$), but this difference is not significant.

Observed data

A first look at Table 6 indicates that the Rwandan population is more diverse than the French one ($\Pi_n = 7.73$ vs $\Pi_n = 2.51$). Because of the different sample sizes we use Π_n, the average number of pairwise differences, to make comparisons between populations. The Rwanda population has a larger excess of singletons than the French population ($D = -1.01$ vs $D = -0.35$). This excess indicates that both populations do not behave neutrally and have been subject to demographic and selective effects. Regarding linkage disequilibrium the European population shows a greater value of $Z_{nS} = 0.37$ compared to $Z_{nS} = 0.08$ in Rwanda. Differentiation is high (Distance of Nei = 1.22), which can be reflected by the total amount of private polymorphisms in Rwanda (W1 = 3153) and France (W2 = 219) when compared to the number of shared polymorphisms (W4 = 408). The number of fixed differences between populations is small (W3 = 10) (Table 7).

Table 4: Parameter estimates of migration and divergence for models B and C.

Parameter	**Prior**	**Mode (95% confidence intervals)**	**Model**
T_{AE} ($\log_{10}$ generations)	unif(4,7)	5.32 (4.95,5.70)	B
$m(\log_{10})$	unif(-10,-2)	-5.18 (-8.50,-2.85), $Nm \sim 33$	B
T_{AE} ($\log_{10}$ generations)	unif(4,7)	5.54 (5.16,5.92)	C
$m_{AE}(\log_{10})$	unif(-10,-2)	-5.98 (-9.51,-3.11), $Nm \sim 5$	C
$m_{EA}(\log_{10})$	unif(-10,-2)	-5.50 (-9.35,-2.93), $Nm \sim 13$	C

Table 5: Performance results for model C.

$\log_{10}\tau$	$\log_{10} m_{AE}$	$\log_{10} m_{EA}$	**Bias** $\log_{10}\tau$	**Bias** $\log_{10} m_{AE}$	**Bias** $\log_{10} m_{EA}$	**RMSE** $\log_{10}\tau$	**RMSE** $\log_{10} m_{AE}$	**RMSE** $\log_{10} m_{EA}$
4.5	-4	-4	-0.027829	0.169555	0.393976	0.203714	0.997351	2.040190
4.5	-4	-6	-0.032854	0.132168	0.136618	0.219388	0.842158	1.571772
4.5	-4	-8	-0.033155	0.108650	-0.167938	0.206506	0.755752	1.884074
4.5	-6	-4	0.008486	0.136016	0.366639	0.147034	1.371578	1.856119
4.5	-6	-6	-0.004546	0.164559	0.125965	0.136607	1.419205	1.423708
4.5	-6	-8	-0.009705	0.151627	-0.141506	0.147622	1.260827	1.631664
4.5	-8	-4	0.002490	-0.144219	0.379102	0.160574	1.512311	1.973454
4.5	-8	-6	-0.007025	-0.118541	0.140370	0.145679	1.316526	1.463076
4.5	-8	-8	-0.000023	-0.123415	-0.158692	0.129203	1.396627	1.804983
5.25	-4	-4	-0.016905	0.007646	0.032271	0.155245	0.351082	0.372521
5.25	-4	-6	-0.015814	0.001716	0.227941	0.132449	0.379174	1.723619
5.25	-4	-8	-0.020782	-0.003711	-0.055728	0.154678	0.372746	1.100955
5.25	-6	-4	0.001128	0.078395	0.013678	0.136369	1.024925	0.327457
5.25	-6	-6	-0.008894	0.059300	0.211525	0.109085	1.030991	1.576377
5.25	-6	-8	-0.007774	0.126502	-0.069899	0.116904	1.304842	1.120604
5.25	-8	-4	0.001300	-0.146983	0.016191	0.137589	1.522591	0.317014
5.25	-8	-6	-0.004414	-0.157988	0.217288	0.102133	1.669312	1.630134
5.25	-8	-8	-0.005879	-0.177837	-0.066934	0.096628	1.846462	1.175981
6.0	-4	-4	0.001212	-0.069841	-0.035869	0.143620	0.384569	0.391674
6.0	-4	-6	-0.002230	-0.051751	0.186735	0.066433	0.339869	1.484811
6.0	-4	-8	-0.000999	-0.082304	-0.011809	0.058181	0.394317	0.472546
6.0	-6	-4	0.015559	0.270689	-0.078482	0.130195	1.841790	0.447217
6.0	-6	-6	0.000459	0.241677	0.213670	0.049268	1.761779	1.633798
6.0	-6	-8	0.000760	0.264056	0.009749	0.039428	1.814339	0.312784
6.0	-8	-4	0.015509	-0.022511	-0.084512	0.135000	0.803076	0.464689
6.0	-8	-6	0.001162	0.017690	0.212129	0.041850	0.693145	1.684660
6.0	-8	-8	0.000433	0.007790	0.015678	0.044984	0.682992	0.343068

Discussion

In this study we quantified gene flow between African and European populations of *D. melanogaster*. Previous results (Singh and Rhomberg, 1987) showed that the product Nm of population size and migration rate between African and European populations was in the order of 2. Our results show that Nm is around 10, which may represent a significant increase of migration rate in the last 25 years. Since *D. melanogaster* is a human commensal we think that this increase in migration rate is correlated with an increase in agricultural trade in the last decades.

Gene flow rates between Africa an Europe are not symmetrical, which is supported by model C being preferred over the others (Table 2). Although the difference between m_{AE} and m_{EA} does not seem to be significant (Table 4) there appears to be more migration from Europe to Africa ($Nm \sim 13$) compared to migration from Africa to Europe ($Nm \sim 5$). We might expect such a difference if European flies are more successful when reintroduced in Africa, which is a common pattern for invasive species (Blossey and Notzold, 1995; Daehler, 2003; Short and Petren, 2012). There is also actual evidence of non-African admixture in African populations (Pool et al., 2012). However, we have to keep in mind that not all invasive species behave the same way and that African flies might be actually less successful in temperate regions, but it is known that *D. melanogaster* is particularly invasive and spreads rapidly in new environments (Sturtevant, 1920; Keller, 2007). Alternatively, it is also possible that there is simply more movement from Europe to Africa than the other way around.

Estimates of population size in Rwanda are different from that of Zimbabwe, and the same applies to the population of France compared to The Netherlands. Although the confidence intervals of these estimates overlap we do not expect different populations have similar population sizes even if they are close to each other, since they could still have different histories. Divergence time between Rwanda

Table 6: Mean and variance (in parenthesis) of observed summary statistics over all 88 loci.

	Africa ($n = 23$)	**Europe** ($n = 8$)
No. of segregating sites S_n	39.9 (245.4)	6.86 (31.98)
Wattersons Θ_W	10.88 (18.14)	2.65 (4.78)
Π_n	7.73 (10.0)	2.51 (5.04)
Tajimas D	-1.10 (0.12)	-0.35 (0.74)
No. of haplotypes K	19.78 (17.71)	4.19 (3.67)
Kellys Z_{nS}	0.08 (0.01)	0.37 (0.07)

and France does not seem to be significantly different to the one reported between Zimbabwe and The Netherlands. We think this might be the case if the founding population of Europe had representatives of both Rwanda and Zimbabwe in similar proportions. Finally, by looking at Tajima's D and the SFS of Rwanda using neutral loci we find footprints of a bottlenecked and an expanding population. This tells us either that Rwanda (or Zimbabwe) is not at the center of origin of *D. melanogaster*, or that selection is affecting the loci that we are studying. We think both of these cases are taking place simultaneously.

Table 7: Comparison between Africa and Europe: population differentiation and summaries of the JSFS. The first line denotes mean and variance (in parenthesis) of distance of Nei as a measure of population differentiation, and lines 2 to 5 represent the classes of the JSFS according to Wakeley and Hey (1997).

	Africa-Europe
Distance of Nei	1.22 (0.48)
W1 (private polymorphisms of Africa)	3153
W2 (private polymorphisms of Europe)	219
W3 (fixed differences between populations)	10
W4 (shared polymorphisms between populations)	408

Acknowledgements

We would like to thank Stephan Hutter for his help in extracting the loci here analyzed.

References

Agis, M. and C. Schlötterer. 2001. Microsatellite variation in natural drosophila melanogaster populations from new south wales (australia) and tasmania. Molecular ecology 10:1197–1205.

Bächli, G., C. R. Vilela, S. A. Escher, and A. Saura. 2005. The Drosophilidae (Diptera) of Fennoscandia and Denmark. *in* Fauna Entomologica Scandinavica vol. 39. Brill Academic Publishers.

Begun, D. J. and C. F. Aquadro. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548–550.

Blossey, B. and R. Notzold. 1995. Evolution of increased competitive ability in invasive nonindigenous plants: a hypothesis. Journal of Ecology 83:887–889.

Coyne, J. A. and B. Milstead. 1987. Long-distance migration of *Drosophila*. 3. Dispersal of *D. melanogaster* alleles from a Maryland orchard. The American Naturalist 130:70–82.

Daehler, C. C. 2003. Performance comparisons of co-occurring native and alien invasive plants: implications for conservation and restoration. Annual Review of Ecology, Evolution, and Systematics Pages 183–211.

David, J. R. and P. Capy. 1988. Genetic variation of *Drosophila melanogaster* natural populations. TRENDS in Genetics 4:106–111.

Depaulis, F., M. Veuille, et al. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution 15:1788–1790.

Duchen, P., D. Živković, S. Hutter, W. Stephan, and S. Laurent. 2013. Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population. Genetics 193:291–301.

Fiston-Lavier, A.-S., N. D. Singh, M. Lipatov, and D. A. Petrov. 2010. *Drosophila melanogaster* Recombination Rate Calculator. Gene 463:18–20.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De-Lorenzo. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165:1269–1278.

Hackman, W. 1954. Die Drosophila-Arten Finnlands. Notulae Entomologicae 34:130–139.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Research 15:790–799.

Haldane, J. B. S. 1930. A mathematical theory of natural and artificial selection.(part vi, isolation.). Pages 220–230 *in* Mathematical Proceedings of the Cambridge Philosophical Society vol. 26 Cambridge Univ Press.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Johnson, C. W. 1913. The distribution of some species of *Drosophila*. Psyche 20:202–205.

Keller, A. 2007. *Drosophila melanogaster*'s history as a human commensal. Current Biology 17:R77–R81.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics 146:1197–1206.

Kennington, W. J., J. Gockel, and L. Partridge. 2003. Testing for asymmetrical gene flow in a drosophila melanogaster body-size cline. Genetics 165:667–673.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evolutionary Biology 22:159–225.

Lachaise, D. and J.-F. Silvain. 2004. How two afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. Genetica 120:17–39.

Laurent, S. J., A. Werzner, L. Excoffier, and W. Stephan. 2011. Approximate bayesian analysis of drosophila melanogaster polymorphism data reveals a recent colonization of southeast asia. Molecular biology and evolution 28:2041–2051.

Nei, M. and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences 76:5269–5273.

Ometto, L., S. Glinka, D. De-Lorenzo, and W. Stephan. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Molecular Biology and Evolution 22:2119–2130.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. Emerson, P. Saelao, D. J. Begun, et al. 2012.

Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genetics 8:e1003080.

Short, K. H. and K. Petren. 2012. Rapid species displacement during the invasion of Florida by the tropical house gecko *Hemidactylus mabouia*. Biological Invasions 14:1177–1186.

Singh, R. S. and L. R. Rhomberg. 1987. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. Genetics 115:313–322.

Stephan, W. and H. Li. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Sturtevant, A. H. 1920. Genetic studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. Genetics 5:488–500.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tellier, A., P. Pfaffelhuber, B. Haubold, L. Naduvilezhath, L. E. Rose, T. Städler, W. Stephan, and D. Metzler. 2011. Estimating parameters of speciation models based on refined summaries of the joint site-frequency spectrum. PloS One 6:e18155.

Wakeley, J. and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847–855.

Wallace, B. 1966. On the dispersal of *Drosophila*. American Naturalist 100:551–563.

Wallace, B. 1970. Observations on the microdispersion of drosophila melanogaster. Pages 381–399 *in* Essays in Evolution and Genetics in Honor of Theodosius Dobzhansky. Springer.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256–276.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier. 2010. Abctoolbox: a versatile toolkit for approximate bayesian computations. BMC bioinformatics 11:116.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97–159.

Chapter 3: Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture

OPEN ACCESS Freely available online

PLOS | GENETICS

Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture

John E. Pool[1]*, Russell B. Corbett-Detig[2], Ryuichi P. Sugino[1], Kristian A. Stevens[3], Charis M. Cardeno[3], Marc W. Crepeau[3], Pablo Duchen[4], J. J. Emerson[5], Perot Saelao[3], David J. Begun[3], Charles H. Langley[3]

1 Laboratory of Genetics, University of Wisconsin–Madison, Madison, Wisconsin, United States of America, **2** Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Evolution and Ecology, University of California Davis, Davis, California, United States of America, **4** Section of Evolutionary Biology, University of Munich, Planegg-Martinsried, Germany, **5** Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America

Abstract

Drosophila melanogaster has played a pivotal role in the development of modern population genetics. However, many basic questions regarding the demographic and adaptive history of this species remain unresolved. We report the genome sequencing of 139 wild-derived strains of *D. melanogaster*, representing 22 population samples from the sub-Saharan ancestral range of this species, along with one European population. Most genomes were sequenced above 25X depth from haploid embryos. Results indicated a pervasive influence of non-African admixture in many African populations, motivating the development and application of a novel admixture detection method. Admixture proportions varied among populations, with greater admixture in urban locations. Admixture levels also varied across the genome, with localized peaks and valleys suggestive of a non-neutral introgression process. Genomes from the same location differed starkly in ancestry, suggesting that isolation mechanisms may exist within African populations. After removing putatively admixed genomic segments, the greatest genetic diversity was observed in southern Africa (*e.g.* Zambia), while diversity in other populations was largely consistent with a geographic expansion from this potentially ancestral region. The European population showed different levels of diversity reduction on each chromosome arm, and some African populations displayed chromosome arm-specific diversity reductions. Inversions in the European sample were associated with strong elevations in diversity across chromosome arms. Genomic scans were conducted to identify loci that may represent targets of positive selection within an African population, between African populations, and between European and African populations. A disproportionate number of candidate selective sweep regions were located near genes with varied roles in gene regulation. Outliers for Europe-Africa F_{ST} were found to be enriched in genomic regions of locally elevated cosmopolitan admixture, possibly reflecting a role for some of these loci in driving the introgression of non-African alleles into African populations.

Citation: Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, et al. (2012) Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. PLoS Genet 8(12): e1003080. doi:10.1371/journal.pgen.1003080

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received May 4, 2012; **Accepted** September 27, 2012; **Published** December 20, 2012

Funding: This work was supported by NIH grant HG02942 to CHL and DJB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jpool@wisc.edu

Introduction

Drosophila melanogaster has a well known history and ongoing role as a model organism in classical and molecular genetics. Its well-annotated genome [1,2] and genetic toolkit have also made it an important model organism in the field of population genetics, in many cases motivating the development of broadly applicable theoretical models and statistical methods. Prior to the advent of DNA sequencing, studies of inversions and allozymes in *D. pseudoobscura* [3,4], and later *D. melanogaster* [5,6], provided some of the field's first glimpses of genetic polymorphisms within and between populations, often providing evidence for geographic clines consistent with local adaptation.

The analysis of DNA sequence data from the *Drosophila Adh* gene motivated the development of methods that compare polymorphism and divergence at different gene regions [7] or functional categories of sites [8], and offered examples of non-neutral evolution. Sequence polymorphism data from additional *D. melanogaster* genes revealed that recombination rate is strongly correlated with nucleotide diversity but not between-species divergence in *D. melanogaster* [9]. This result suggested that genetic hitchhiking [10] could be an important force in molding diversity across the *Drosophila* genome, but it also motivated the suggestion that background selection against linked deleterious variants [11] should likewise reduce diversity in low recombination regions of the genome.

Larger multi-locus data sets initially came from studies of microsatellites and short sequenced loci. Several of these studies compared variation between ancestral range populations from sub-Saharan Africa and more recently founded temperate populations from Europe, finding that non-African variation is

Author Summary

Improvements in DNA sequencing technology have allowed genetic variation to be studied at the level of fully sequenced genomes. We have sequenced more than 100 *D. melanogaster* genomes originating from sub-Saharan Africa, which is thought to contain the ancestral range of this model organism. We found evidence for recent and substantial non-African gene flow into African populations, which may be driven by natural selection. The data also helped to refine our understanding of the species' history, which may have involved a geographic expansion from southern central Africa (*e.g.* Zambia). Lastly, we identified a large number of genes and functions that may have experienced recent adaptive evolution in one or more populations. An understanding of genomic variation in ancestral range populations of *D. melanogaster* will improve our ability to make population genetic inferences for worldwide populations. The results presented here should motivate statistical, mathematical, and computational studies to identify evolutionary models that are most compatible with observed data. Finally, the potential signals of natural selection identified here should facilitate detailed follow-up studies on the genetic basis of adaptive evolutionary change.

far more strongly reduced on the X chromosome than on the autosomes [12,13]. Sequence data also allowed larger-scale comparisons of polymorphism and divergence, leading to suggestions that significant fractions of substitutions at nonsynonymous sites [14] and non-coding sites [15] were driven by positive selection.

Although previous studies have found considerable evidence for a genome-wide influence of natural selection, a thorough and confident identification of recent selective sweeps in the genome requires an appropriate neutral null model that incorporates population history. Both biogeography [16] and genetic variation [17,18] indicate that *D. melanogaster* originated within sub-Saharan Africa. Even within Africa, *D. melanogaster* has only been collected from human-associated habitats, and so its original habitat and ecology, along with the details of its transition to a human commensal species, remain unknown [19]. A few studies have found populations from eastern and southern Africa to be the most genetically diverse [18,20,21], suggesting that the species' ancestral range may lie within these regions. Small but significant levels of genetic structure are present within sub-Saharan Africa [18], which could reflect either long-term restricted migration or short-term effects of bottlenecks associated with geographic expansions within Africa.

On the order of 10,000 years ago [22–24], *D. melanogaster* is thought to have first expanded beyond sub-Saharan Africa, perhaps by traversing formerly wetter parts of the Sahara [16] or the Nile Valley [18]. This expansion involved a significant loss of genetic variation [12,13], brought *D. melanogaster* into the palearctic region (northern Africa, Asia, and Europe), and largely gave rise to the "cosmopolitan" populations that live outside sub-Saharan Africa today. American populations were founded only within the past few hundred years [25], and their complex demography appears to involve admixture between European and African source populations [26].

Recent advances in DNA sequencing technology have allowed genetic variation to be studied on the whole-genome scale. The sequencing of six *D. simulans* genomes [27] provided the first comprehensive look at fluctuations of polymorphism and divergence across the genome and their potential causes, including potential targets of adaptive evolution. More recently, larger samples of *D. melanogaster* genomes have been sequenced, yielding further insight into the potential impact of natural selection on diversity across the *Drosophila* genome [28,29] and connections between genetic and phenotypic variation [29]. However, a large majority of the sequenced genomes are of North American origin, and before we can clearly understand the demographic history of that population, we must investigate genomic variation in its African and European antecedents.

Here, we use whole genome sequencing and population genomic analysis to examine genetic variation in wild-derived population samples of *D. melanogaster*. We use a new method to detect pervasive admixture from cosmopolitan into sub-Saharan populations. We use geographic patterns of genetic diversity and structure to investigate the history of *D. melanogaster* within Africa. Finally, we identify loci with unusual patterns of allele frequencies within or between populations, which may represent targets of recent directional selection.

Results

With the ultimate aim of identifying population samples of importance for future population genomic studies, we sequenced genomes from 139 wild-derived *D. melanogaster* fly stocks. These genomes represented 22 population samples from sub-Saharan Africa and one from Europe (Figure 1; Table S1; Table S2). Most of these genomes were obtained from haploid embryos [30]. These genomes were found to be essentially homozygous (with the exception of chromosome 2 from GA187 [28]). A smaller number of genomes were sequenced from homozygous chromosome extraction lines; those included in the published data were found to be homozygous for target chromosome arms (X, 2L, 2R, 3L, 3R). Three genomes (the ZK sample; Table S1) were sequenced from adult flies from inbred lines; these were found to have extensive residual heterozygosity (results not shown). The data we analyze below consists entirely of non-heterozygous sequences from haploid embryo genomes. Apparently heterozygous sites in target chromosome arms were observed at low rates in all genomes, potentially resulting from cryptic copy number variation or recurrent base-calling errors, and were excluded from analysis. Based on the rarity of such sites (approximately one per 20 kb on average), their exclusion seems unlikely to strongly influence genome-wide summary statistics.

Sequencing was performed using the Illumina Genome Analyzer IIx platform. Paired-end reads of at least 76 bp were sequenced for each genome (Table S2). Alignment was performed using BWA [31], with consensus sequences generated via SAMtools [32]. Reads with low mapping scores (<20) were discarded, and positions within 5 bp of a consensus indel were masked (treated as missing data) for the genome in question. Resequencing of the reference strain y^1, cn^1, bw^1, sp^1 and the addition of simulated genetic variation allowed the quality of assemblies to be assessed. Based on the inferred tradeoff between error rates and genome-wide coverage, a nominal BWA quality score of Q31 (corresponding to an estimated Phred score of Q48) was chosen as a quality threshold for subsequent analyses (Figure S1). Genomic regions with long blocks of identity-by-descent (IBD) consistent with relatedness were masked (Table S3; Table S4). A full description of data generation and initial analysis can be found in the Materials and Methods section. Processed and raw sequence data for all genomes can be found at http://www.dpgp.org/dpgp2/DPGP2.html and http://ncbi.nlm.nih.gov/sra.

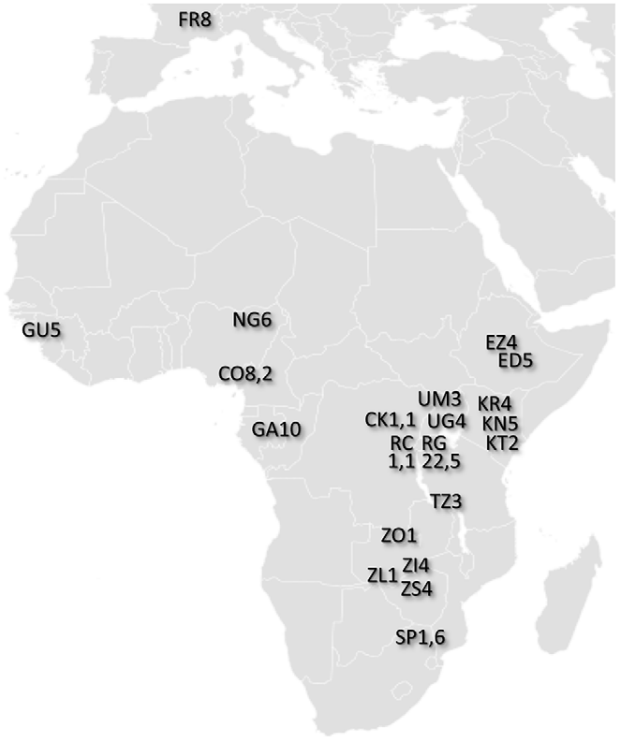

Figure 1. Locations of population samples from which the analyzed genomes were derived. Each population sample is indicated by a two letter abbreviation followed by the number of primary core genomes sequenced. For populations with secondary core genomes, that number follows a comma. Additional data and sample characteristics are described in Table S1.
doi:10.1371/journal.pgen.1003080.g001

Correlation of sequencing depth with genomic coverage and genetic distance

The sequenced genomes vary significantly in mean sequencing depth (average number of reads at a given bp) present in the assemblies. Among genomes with relevant data from all five target chromosome arms, mean depth ranges from 18X to 47X (Table S2). Depth was found to have a substantial influence on pairwise genetic distances. Mean depth showed positive, non-linear relationships with distance from the *D. melanogaster* reference genome, and with average distances to other African samples, such as Zambia-Siavonga (Figure 2A). The relationship between depth and genetic distance from Zambia is especially strong (Spearman $\rho = 0.63$; $P<0.00001$), suggesting that population ancestry has little influence on this quantity (a property of the ZI sample further discussed below). This correlation is especially pronounced for genomes with depth below 25X, while only a modest slope is present above this threshold.

Mean depth was also correlated with genomic coverage – the portion of the genome with a called base at the quality threshold (Figure 2B; Spearman $\rho = 0.62$; $P<0.00001$). The lowest depth genomes were found to have ~2% lower coverage than a typical genome with average depth. Some correlation of depth with genetic distance and genomic coverage might be expected if genomes with higher depth were more successful in mapping reads across genomic regions with high levels of substitutional (and perhaps structural) variation. Additionally, a consensus-calling bias in favor of the reference allele, such that higher depth genomes were more likely to have adequate statistical evidence favoring a non-reference allele, might contribute to the reduced genetic distances and genomic coverage exhibited by the lowest depth genomes.

The influence of depth on genetic distance has the potential to bias most population genetic analyses. We found that strict sample coverage thresholds (only analyzing sites covered in most or all assemblies) could ameliorate the depth-distance correlation, but at the cost of excluding most variation and introducing a substantial reference sequence bias (Figure S2). Instead, we addressed the depth-distance issue by focusing most analyses on genomes with >25X depth and made additional corrections when needed, as described below. Assemblies derived from haploid embryos with >25X depth were defined as the "primary core" data set (Table S1). Haploid embryo genomes with <25X depth were denoted as "secondary core". Genomes not derived from haploid embryos were labeled "non-core", and were not analyzed further in this study.

Identification of cosmopolitan admixture in sub-Saharan genomes

Previous work has suggested that introgression from cosmopolitan sources (*i.e.* populations outside sub-Saharan Africa) may be an important component of genetic variation for at least some African populations of *D. melanogaster* [18,33,34]. Preliminary examination of this data set revealed a number of sub-Saharan genomes with unusually low genetic distances to cosmopolitan

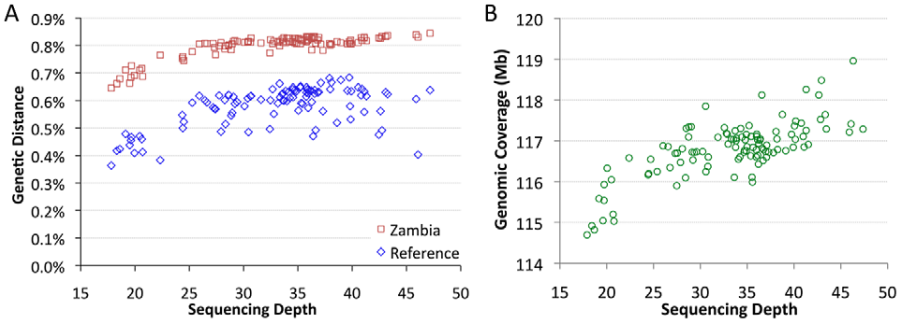

Figure 2. Mean sequencing depth. Mean sequencing depth is correlated with genetic distance (A) and genomic coverage (B). African core genomes with data from all major chromosome arms are depicted. The effect of depth on genetic distance applies whether genomes are compared to the published reference genome (blue) or the Zambia ZI population sample (red). Subsequent analyses focused largely on "primary core" genomes with >25X depth.
doi:10.1371/journal.pgen.1003080.g002

genomes (the latter represented by the European FR sample and the North American reference genome). Undetected admixture could undermine the demographic assumptions of many population genetic methods, altering genetic diversity and population differentiation, and creating long-range linkage disequilibrium. Hence, we attempted to identify specific chromosome intervals that have non-African ancestry, so that they could be filtered from downstream analyses when appropriate.

We developed a Hidden Markov Model (HMM) method to identify chromosome segments from sub-Saharan genomes that have cosmopolitan ancestry, as described under Materials and Methods. The method utilized a "European panel" (the FR sample) and an "African panel" (the RG sample) which may contain some admixture. Because of the diversity-reducing out-of-Africa bottleneck, non-African genomes should be more closely related to each other than they are to African genomes. Therefore, if we examine genomic windows of sufficient length, genetic distances between two FR genomes should be consistently lower than between an RG and an FR genome (Figure S3). To take advantage of this contrast, we constructed chromosome arm-wide emissions distributions by evaluating two locally rescaled quantities in ~50 kb windows. One distribution, representing African ancestry, was formed from genetic distances of each RG genome to the FR panel. The other distribution, representing non-African ancestry, was formed from genetic distances between each FR genome and the remainder of the FR panel. Individual African genomes were then compared to the FR panel to determine the likelihood of African or non-African ancestry in each window (essentially, using the emissions distributions to determine whether we are truly making an Africa-Europe genetic comparison, or if we are actually comparing two non-African alleles in the case of an admixed African genome). The HMM was then applied to convert likelihoods to admixture probabilities for each genome in each window. This approach was validated using simulations (see Materials and Methods; Figure S4). For the empirical data, the above approach was applied iteratively to the RG sample to eliminate non-African intervals from the "African panel" used to create emissions distributions. Emissions distributions generated using the FR and RG samples were also used to calculate admixture probabilities for the other sub-Saharan primary core genomes. Simple correction factors were applied to account for the effects of sequencing depth and other quality factors for each genome (Materials and Methods).

When applied to the RG primary core genomes, the admixture detection method produced generally sharp peaks along chromosome arms, with only 3.3% of window admixture probabilities between 0.05 and 0.95 (Figure S5; Table S5). When primary core genomes from other population samples were analyzed, results still appeared to be of reasonable quality, with 8.3% "intermediate" admixture probabilities as defined above (Figure S5; Table S5). However, inferences for the secondary core and non-core genomes appeared less reliable, with 22.5% intermediate admixture probabilities and more admixture predicted in general (Figure S5; Table S6). Hence, the influence of lower sequencing depth may have added significant "noise" into the admixture analysis. Below, we focus on admixture inferences from the primary core genomes only.

Inter-population variability in cosmopolitan admixture proportion

The estimated proportion of cosmopolitan admixture varied dramatically among the twenty sub-Saharan population samples represented in the primary core data set (Figure 3A, Table S7). In general, populations with substantial admixture were observed across sub-Saharan Africa, but admixture proportion varied substantially within geographic regions. At the extremes, one Zambia sample (ZI) had 1.4% inferred admixture among four genomes, while another Zambia sample (ZL) had 84% inferred admixture from the single genome sequenced. A Kruskal-Wallis test for the 14 populations with $n \geq 3$ primary core genomes supported a significant effect of population on admixture proportion ($P < 0.0001$).

Testing whether admixture might be related to anthropogenic activity, we found that human population size of the collection locality had a strong positive correlation with admixture proportion (Spearman $\rho = 0.60$; one-tailed $P = 0.003$; Figure S6). For the seven collection sites with population sizes below 20,000, all but one population sample had an admixture proportion below 7% (the exception, KR, may reflect a higher regional effect of admixture in Kenya). In contrast, for the eight cities with a population above 39,000, admixture proportion was always above 15%. These results mirror previous findings that urban African flies are genetically intermediate between rural African flies and European flies, when population samples from the Republic of Congo [33,35] and Zimbabwe [34] were examined. Our results suggest that African invasion by cosmopolitan *D. melanogaster* is not

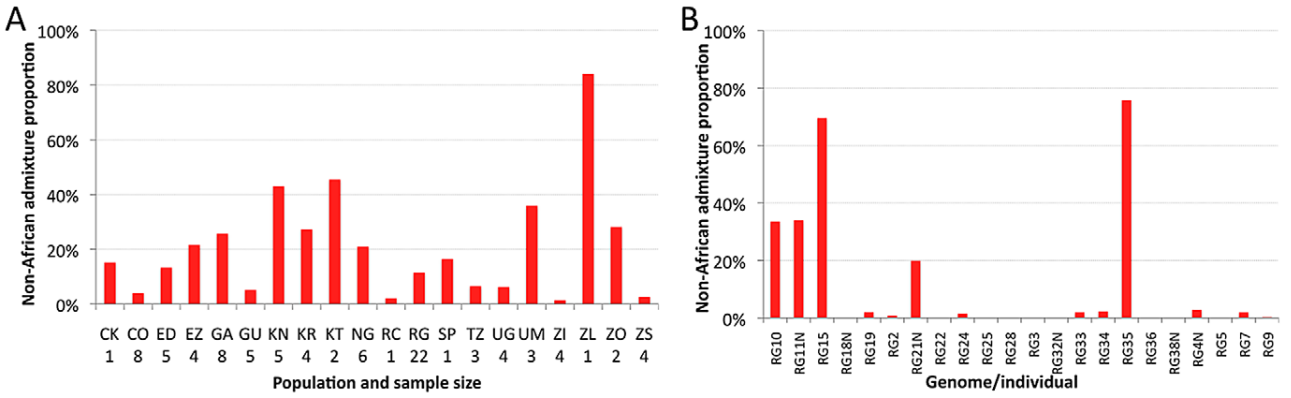

Figure 3. Heterogeneity in estimated cosmopolitan admixture proportions. Heterogeneity in estimated cosmopolitan admixture proportions, both among African populations (A) and within the Rwanda RG population sample (B).
doi:10.1371/journal.pgen.1003080.g003

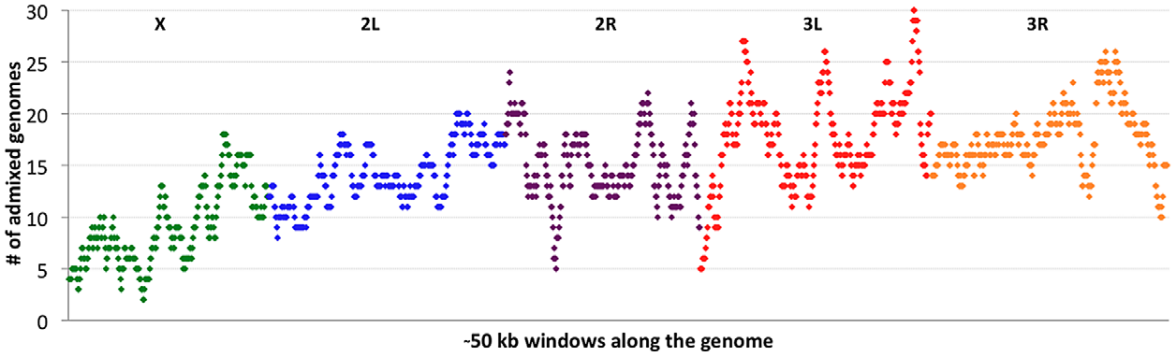

Figure 4. Cosmopolitan admixture levels are depicted across the genome. For each genomic window, the number of African primary core genomes (across all populations) with >50% admixture probability is plotted. Chromosome arms are labeled and indicated by color. Each window contains 1000 RG non-singleton SNPs (approximately 50 kb on average).
doi:10.1371/journal.pgen.1003080.g004

limited to the largest African cities, and has occurred in moderately sized towns and cities across sub-Saharan Africa.

In theory, higher admixture levels in urban African locations could result from either neutral or adaptive processes. If larger cities are more connected to international trade, then selectively neutral immigration would affect urban populations first. However, the large size of admixture tracts (*e.g.* a mean admixture tract length of 4.8 centiMorgans or 3.8 Mb for the RG sample; Table S7) suggests an unusually rapid spread of cosmopolitan alleles into Africa, which may not be compatible with plausible levels of passive gene flow. We used the method of Pool and Nielsen [36] to estimate, for the RG sample, the parameters of a two epoch migration rate change model. This method found the highest likelihood for a change in migration 59 generations before present, with near-zero migration before this time (point estimate 1.2×10^{-8}), and an unscaled migration rate of 0.0010 since the change. It is not clear whether a neutral model invoking thousands of immigrants per generation should be viewed as realistic. Note that the rate of admixture would have to be higher yet if this small Rwandan town was not the point of African introduction for cosmopolitan immigrants.

Alternatively, cosmopolitan admixture into sub-Saharan *D. melanogaster* could be a primarily adaptive phenomenon. Certain cosmopolitan alleles might provide a selective advantage in modern urban environments and may now be favored in modernizing African cities, but may be neutral or deleterious in rural African environments. Or, some cosmopolitan genotypes (such as those conferring insecticide resistance [37]) may now be advantageous in both urban and rural African environments, but have thus far spread primarily into urban areas. In either scenario, there is still a role for demography (*i.e.* migration rates within Africa) in governing the geographic spread of cosmopolitan alleles into African environments in which they are adaptive.

Intra-population variability in cosmopolitan admixture proportion

Perhaps more striking than the between-population pattern of admixture are the stark differences in ancestry observed within populations. This individual variability is well-illustrated by the RG sample (Figure 3B), but similar patterns are also observed in other populations (Table S8). Among the 22 RG primary core genomes, nine have no inferred admixture at all, eight others have less than 3% admixture, while the other five genomes contain 20–76% admixture. Based on forward simulations with recombination and migration [36], the observed variance among genomes in cosmopolitan admixture proportion for the RG sample was found to be unlikely under the point estimates of demographic parameters reported above (one-tailed $P=0.02$).

The unexpectedly high variance in admixture proportion may require a combination of biological explanations. Inversion frequency differences between African and introgressing chromosomes would reduce the rate of recombination, potentially keeping admixture in longer blocks. However, the genome-wide prevalence of long admixture tracts (including in regions that do not overlap common inversions) makes this explanation incomplete at best.

Alternatively, African populations may be subject to local heterogeneity for any number of environmental factors, and cosmopolitan alleles may confer a greater preference for and/or fitness in specific microhabitats. Such differences might provide a degree of spatial isolation between flies with higher and lower levels of admixture. However, the RG sample was collected from a handful of markets, restaurants, and bars in the center of the relatively small town of Gikongoro, Rwanda (an area less than 200 m across), and it's not clear whether any meaningful isolation could exist on this scale.

Finally, sexual selection may play a role in generating this pattern. African strains of *D. melanogaster* are known to display varying degrees of "Z-like" mating behavior, in which females discriminate against males from "M-like" strains, which include cosmopolitan populations [38,39]. Hence, one would expect many African females to avoid mating with males carrying the cosmopolitan alleles responsible for the M phenotype. And indeed, mating choice experiments [33] found that matings between rural Brazzaville females and urban (apparently admixed) Brazzaville males were much less frequent than homogamic pairings. This phenomenon might help to explain the prevalence of admixture in only a subset of genomes in RG and other samples. Further empirical and predictive studies will be needed to assess the ability of these and other hypotheses to explain the inferred patterns of cosmopolitan admixture among sub-Saharan genomes.

Intra-genomic variability in cosmopolitan admixture proportion

If cosmopolitan admixture is partly due to adaptive processes, it may be worthwhile to examine variability in admixture proportion across the genome. Figure 4 shows the number of primary core genomes with admixture probability above 50% for each window analyzed by the admixture HMM. By including admixture tracts from 95 sub-Saharan genomes across all populations, we may lose

some population-specific signals, but we gain resolution that would not exist within small samples.

Clear differences were observed between chromosomes in admixture levels. Averaging across all windows, arms 3L and 3R had the highest number of admixed genomes (18.1 and 18.0, respectively), while 2L and 2R were somewhat lower (both averaged 14.7). Both autosomes, however, were considerably more admixed than the X chromosome, which averaged just 9.3 admixed genomes per window. A qualitatively consistent pattern has been reported [34] in which cosmopolitan admixture was detected on the third chromosome but not the X chromosome in a sample from Harare, Zimbabwe.

A lesser contribution of the X chromosome to cosmopolitan admixture might be expected if males contributed disproportionately to introgression. However, the mating preferences described above might be expected to yield the opposite result, suppressing genetic contributions from cosmopolitan males into African populations. Additionally, the loci responsible for the M/Z behavioral polymorphism are thought to reside primarily on the autosomes [38], which should impede autosomal introgression rather than X-linked introgression. Another explanation for the deficiency of X-linked admixture is more efficient selection due to the X chromosome's hemizygosity [40]. The X chromosome might have experienced a higher rate of "out-of-Africa" selective sweeps [12], and even though some cosmopolitan adaptations may now be favored in Africa, it is conceivable that the X chromosome contains a greater density of cosmopolitan alleles that are still deleterious in sub-Saharan Africa and limit X-linked introgression. Even if cosmopolitan alleles that remain deleterious in Africa occur at similar rates on the X chromosome and autosomes, selection might be more effective against introgressing X chromosomes. Alternatively, the X chromosome's higher recombination rate may lead advantageous X-linked cosmopolitan alleles to introgress within smaller chromosomal blocks. The recombination rate difference predicted by mapping crosses [28] will be magnified by the lack of recombination in males, and perhaps also by the autosomes' generally higher levels of inversion polymorphism [41], which should decrease autosomal recombination rates in nature and increase X chromosome recombination (due to the interchromosomal effect [42]).

Considerable variation in the proportion of admixed individuals was also apparent within chromosomes. For example, the X chromosome's dearth of admixture was most dramatic for the telomere-proximal half of its windows (average 6.9 admixed genomes) and less severe for the centromere-proximal half (average 11.7). On a finer scale, the proportion of admixed genomes showed relatively narrow genomic peaks and valleys (Figure 4), with the most extreme admixture levels often limited to intervals on the order of 100 kb. If the adaptive hypothesis of cosmopolitan admixture is correct, genomic peaks and valleys of admixture could include cosmopolitan loci that are advantageous and deleterious, respectively, in sub-Saharan Africa. We return to the specific content of these intervals later, in the context of out-of-Africa sweeps.

Principal Components Analysis

In order to evaluate the effectiveness of admixture identification and to examine geographic gradients of genetic variation, principal Components Analysis (PCA) [43] was applied to admixture-filtered and unfiltered data. In both cases, the first principal component clearly reflected cosmopolitan versus African ancestry. Comparison of these results suggested that our admixture detection method had successfully filtered most, but not all, cosmopolitan admixture from sub-Saharan genomes (Figure 5A).

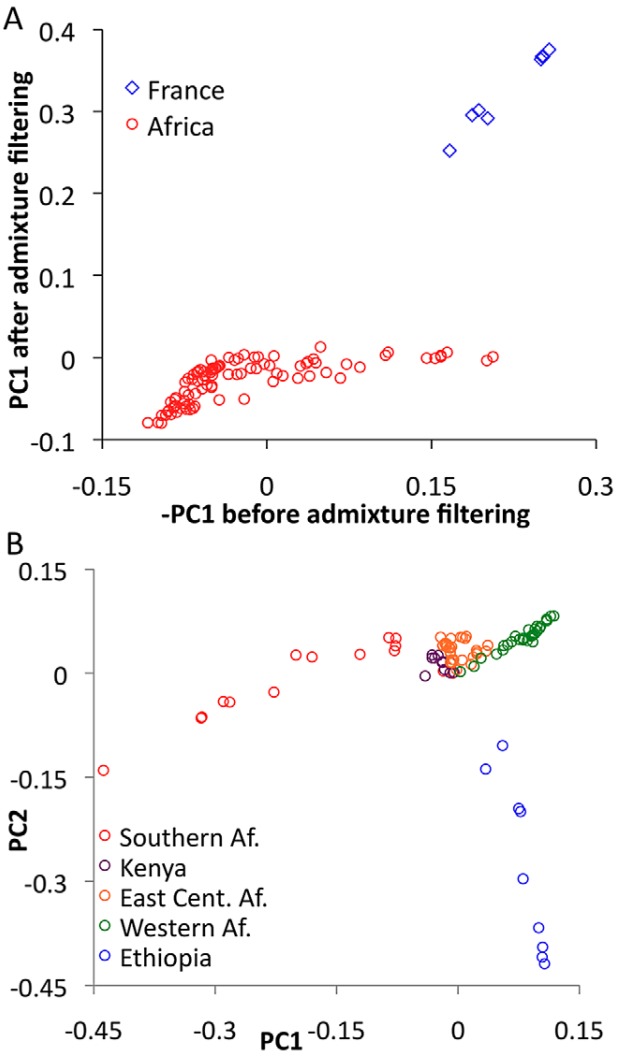

Figure 5. Principal Components Analysis (PCA). (A) PCA was done for the full primary core data set before and after masking putative cosmopolitan admixture from sub-Saharan genomes. Reductions in the magnitude of PC1 after filtering are consistent with the admixture identification method being largely successful. (B) PCA was applied to the sub-Saharan genomes only, after admixture filtering. Genomes were found to cluster by geographical region, including southern (SP, TZ, ZI, ZL, ZO, ZS), eastern (CK, RC, RG, UG, UM), and western (CO, GA, GU, NG) African groups.
doi:10.1371/journal.pgen.1003080.g005

For example, RG35 was by far the most admixed genome in its Rwanda population sample, with a pre-filtering -PC1 of 0.153. After filtering, its PC1 dropped to −0.001 – a considerable improvement, although slightly higher than the population average of −0.049. Hence, a minority of admixture may remain undetected, and for analyses that may be especially sensitive to low levels of admixture, users of the data could opt to exclude genomes with higher levels of detected admixture.

Focusing on PCA from admixture-filtered sub-Saharan data, PC1 separated southern African populations from western African and Ethiopian populations, with eastern African samples having intermediate values (Figure 5B). PC2 mainly distinguished Ethiopian samples from all others, while subsequent principal components lacked obvious geographic patterns (Table S9).

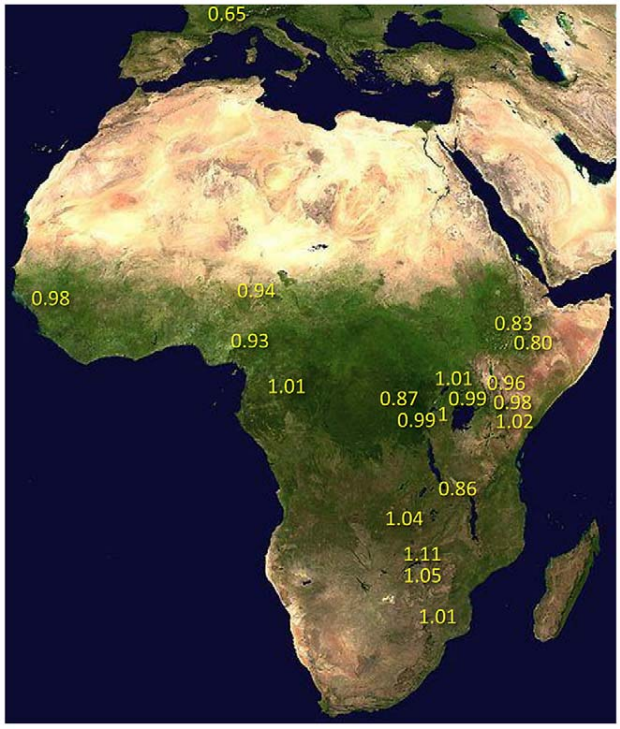

Figure 6. Relative nucleotide diversity, scaled by π_{RG}, was calculated for each population sample. This method allowed the comparison of diversity between populations with missing data in different genomic regions, and allowed the inclusion of secondary core genomes. Values were corrected for the modest predicted effects of sequencing depth (see Materials and Methods), and were based on non-centromeric, non-telomeric chromosomal regions, and equal weighting of chromosome arms.
doi:10.1371/journal.pgen.1003080.g006

Table 1. Relative nucleotide diversity (versus the RG sample) for each population sample is given for chromosome arms and the average of arms.

Population	X	2L	2R	3L	3R	Average
CK*	0.77	**	**	0.92	0.93	0.87
CO	0.88	0.97	0.94	0.97	0.87	0.93
ED	0.73	0.83	0.82	0.86	0.77	0.80
EZ	0.78	0.85	0.83	0.80	0.90	0.83
FR	0.41	0.66	0.59	0.75	0.84	0.65
(FR std)		(0.58)		(0.62)	(0.63)	(0.57)
GA	0.94	0.98	1.01	1.05	1.07	1.01
GU	0.91	1.00	0.98	1.00	1.02	0.98
KN	1.00	0.78	1.00	1.02	1.11	0.98
KR	0.96	0.72	1.03	0.99	1.08	0.96
KT	1.00	1.05	0.99	1.03	1.03	1.02
NG	0.91	0.90	0.93	0.96	1.00	0.94
RC*	1.04	0.98	0.98	1.00	0.98	0.99
SP*	1.05	1.08	1.05	1.01	0.86	1.01
TZ	0.66	0.68	1.02	0.93	1.01	0.86
UG	0.98	1.02	1.00	1.02	0.94	0.99
UM	0.96	0.99	1.05	1.02	1.07	1.02
ZI	1.05	1.15	1.07	1.10	1.17	1.11
ZO	1.05	**	1.03	1.03	1.06	1.04
ZS	0.96	1.13	1.03	1.00	1.13	1.05

Data consisted of non-centromeric, non-telomeric regions, with putatively admixed regions masked from African genomes. For the FR sample, values in parentheses reflect the exclusion of inverted chromosomes.
*denotes a value based on comparisons between primary and secondary core genomes.
**indicates arms for which diversity could not be estimated due to a lack of non-masked data.
doi:10.1371/journal.pgen.1003080.t001

Geographic patterns of sub-Saharan genetic diversity

Nucleotide diversity was evaluated for windows and for full chromosome arms, in terms of both absolute and relative π (the latter based on comparison with the RG sample). The use of relative π allowed unbiased comparisons of diversity involving populations with incomplete genomic coverage of admixture-filtered data, and it enabled populations lacking two or more genomes with >25X depth to be considered by using RG genomes with similar depth for comparison (see Materials and Methods).

Under simple demographic scenarios of geographic expansion, populations with the highest genetic diversity are the most likely to reflect the geographic origin of all extant populations. Hypotheses for the ancestral range of *D. melanogaster* within sub-Saharan Africa have ranged from western and central Africa (based on biogeography [16]) to eastern and southern Africa (based on a smaller sequence data set [18]). Among 19 African populations, the greatest diversity was found in the ZI sample collected from Siavonga, Zambia (Figure 6; Table 1; Table S10), followed by the geographically proximate ZS and ZO samples. The inferred nucleotide diversity of ZI (0.70%; 0.83% for higher recombination regions) is lower than estimates for geographically similar samples based on multilocus Sanger sequencing [13], and slightly lower than a recent population genomic analysis [28], but higher than an earlier population genomic estimate of θ [44]. While differences in the genomic coverage of these data sets may help to explain some differences, mapping and consensus-calling biases against non-reference reads may also play a role. Such factors are not expected to have a dramatic impact on comparisons of diversity levels between African populations. Hence, based on the samples represented in our study, southern-central Africa appears to contain the center of genetic diversity for *D. melanogaster*. Although this hypothesis requires further confirmation, these results are consistent with a southern African origin for *D. melanogaster*.

Much of Zambia and Zimbabwe is characterized by a subtropical climate and seasonally dry Miombo and Mopane woodland. Whether this landscape might reflect the original environment of *D. melanogaster* is unclear, because the species has never been collected from a completely wild environment [16] and the details of its transition to an obligately human-commensal species are unknown [19]. Compared with related species, African strains of *D. melanogaster* have superior resistance to desiccation [45] and temperature extremes [46]. These characteristics would be predicted by an evolutionary origin in subtropical southern Africa, as opposed to humid equatorial forests.

Most populations from eastern Africa (including Kenya, Rwanda, and Uganda) had modestly lower diversity compared to Zambia and Zimbabwe, while western populations (including Cameroon, Guinea, and Nigeria) showed an additional slight reduction. The two Ethiopian samples showed the lowest variation among African populations, with roughly three quarters the diversity of ZI, potentially indicating a bottleneck during or since

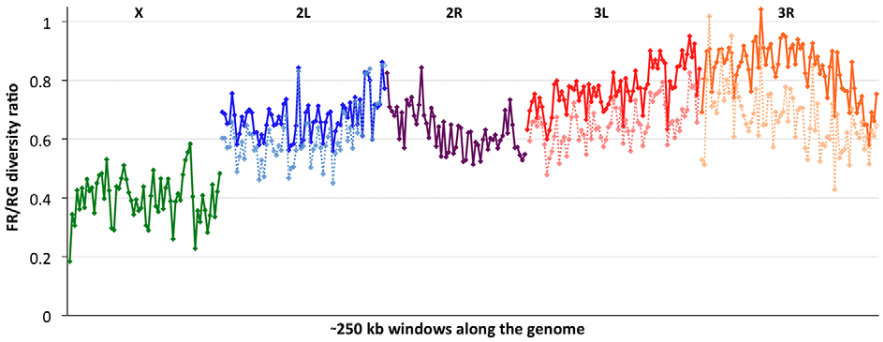

Figure 7. The ratio of nucleotide diversity between non-African (France, FR) and African (Rwanda, RG) genomes. Each window contains 5000 RG non-singleton SNPs. Chromosome arms are labeled and indicated by color. Dashed series for the three arms with segregating inversions in the FR sample reflect diversity ratios for standard chromosomes only, indicating that inversions add significant diversity at the scale of whole chromosome arms.
doi:10.1371/journal.pgen.1003080.g007

the species' occupation of Ethiopia. The distinctness of Ethiopian samples was also indicated by an analysis of mitochondrial and Wolbachia genomes from these same genomes [47]. Otherwise, only two population samples had reduced diversity relative to the overall geographic pattern described above, and one of these (CK) had limited pairwise comparisons in the admixture-filtered data.

The other sample with locally reduced variation, TZ, displays an unusual pattern of diversity loss on chromosome arms X and 2L specifically (Table 1), associated with all three sampled genomes carrying inversions *In(1)A* and *In(2L)t* [48]. Similarly, two of the three Kenya samples (KN and KR) show reduced diversity on arm 2L only, also apparently in association with *In(2L)t* [48]. These results suggest the possibility that selection on polymorphic inversions, which are common in sub-Saharan populations [41], can be an important determinant of genome-scale diversity levels. Although this hypothesis is contrary to some theoretical predictions [49] and empirical findings [50] that would lead one to expect the effects of inversions to be mainly restricted to breakpoint regions, it is supported by an analysis of inversion polymorphism and linked variation in the genomes studied here [48].

Chromosomal diversity in a non-African population

Consistent with previous work [12,13,51], variation for the cosmopolitan sample (FR) is much more strongly reduced on the X chromosome relative to the autosomes (Table 1). However, further genomic patterns in the ratio of π_{FR} to π_{RG} can be observed (Figure 7). This diversity ratio ranges from below 0.2 (at the X telomere) to above 1 (for a window on arm 3R), with similar patterns observed if π_{FR} is instead compared against $\pi_{\sim I}$ (Figure S7). Diversity ratios on autosomal arms showed distinct differences: FR retains 59% of the RG diversity level on arm 2R but 84% on arm 3R (Table 1).

Based on the inversions identified for each genome [48], we examined the influence of inversions on chromosome arm-wide diversity by recalculating π_{FR} and π_{RG} using standard chromosome arms only. For the RG sample, the exclusion of inversion-carrying arms had negligible influence on diversity, except that the inclusion of *In(3R)P* (present in four of 22 genomes) increased π_{RG} on arm 3R by 4% (Table S10). More dramatic contrasts were observed for the FR sample, in which inversions were found to result in arm-wide diversity increases of 10% on arm 2L (due to one of eight FR genomes carrying *In(2L)t*) and 18% on arm 3L (due to a pair of *In(3L)P* chromosomes). As further detailed in a separate analysis [50], arm 3R was even more strongly affected, with a 29% diversity increase due to the presence of *In(3R)P* (in three of eight genomes), *In(3R)K* and *In(3R)Mo* (one genome each). Although the French sample only contains inversions on these three arms, they contribute to a 12% genome-wide increase in nucleotide diversity.

In light of the above observations, it is possible that inversions have had important effects both in reducing chromosome arm-wide diversity (for the Tanzania and Kenya populations) and also in elevating it (for non-African autosomes). As further suggested in a separate analysis [48], the spatial scale of increased diversity associated with inversions in the France sample (Figure 7) may indicate a recent arrival of inverted chromosomes from one or more genetically differentiated populations. Given that similar levels of gene flow are not indicated by polymorphism on chromosome arms lacking inversions, the spread of genetically divergent inverted chromosomes into France may have been primarily driven by natural selection. In light of their powerful elevation of π_{FR}, inverted chromosomes in this sample may have originated from a more genetically diverse African or African-admixed population. Similarly, the more modest elevation of π_{RG} associated with *In(3R)P* might indicate the recent introgression of these inverted chromosomes from a genetically differentiated population. However, the nature of selective pressures acting on inversions in natural populations of *D. melanogaster* remains largely unknown.

Without inversions, relative π_{FR} for autosomal arms ranged from 0.58 to 0.63, with chromosome 3 showing higher values than chromosome 2. In light of the above hypothesis to account for the presence of divergent inverted chromosomes in the France sample, some of the remaining differences in relative π_{FR} among inversion-free chromosomes might stem from recombination between standard chromosomes and earlier waves of introgressing inverted chromosomes. Alternatively, given that *D. melanogaster* autosomes frequently carry recessive deleterious mutations [52], associative overdominance during the out-of-Africa bottleneck might have favored intermediate inversion frequencies [53–55]. This hypothesis is mainly plausible in small populations [56], which may have existed due to strong founder events during the out-of-Africa expansion [57]. Given the opportunity for recombination between standard and inverted chromosomes since that time, past associative overdominance related to inversions (or centromeric regions) might contribute to the modest difference in relative π_{FR} between inversion-free second and third chromosomes, as well as the larger gap between both autosomes and the X chromosome.

Table 2. Nucleotide diversity and genetic differentiation are shown, averaged across the non-centromeric, non-telomeric regions of each chromosome arm.

Population	CO	ED	FR	GA	GU	KR	NG	RG	TZ	UG	ZI	ZS
CO	**0.702**	0.780	0.765	0.759	0.745	0.759	0.738	0.770	0.781	0.766	0.841	0.811
ED	0.159	**0.614**	0.781	0.801	0.790	0.778	0.783	0.789	0.799	0.786	0.845	0.822
FR	0.224	0.297	**0.491**	0.774	0.772	0.751	0.764	0.783	0.783	0.779	0.828	0.805
GA	0.035	0.143	0.193	**0.763**	0.768	0.770	0.749	0.789	0.793	0.789	0.858	0.827
GU	0.031	0.144	0.205	0.020	**0.741**	0.769	0.743	0.781	0.790	0.778	0.851	0.821
KR	0.077	0.156	0.205	0.052	0.063	**0.707**	0.744	0.755	0.700	0.763	0.810	0.760
NG	0.048	0.161	0.221	0.020	0.028	0.055	**0.703**	0.772	0.772	0.772	0.843	0.809
RG	0.056	0.135	0.208	0.039	0.043	0.037	0.057	**0.754**	0.771	0.763	0.828	0.800
TZ	0.138	0.214	0.274	0.113	0.123	0.037	0.127	0.091	**0.650**	0.784	0.819	0.754
UG	0.052	0.135	0.206	0.041	0.042	0.047	0.058	0.015	0.110	**0.750**	0.838	0.811
ZI	0.090	0.146	0.205	0.072	0.077	0.053	0.090	0.043	0.094	0.057	**0.831**	0.817
ZS	0.082	0.149	0.208	0.063	0.070	0.015	0.078	0.036	0.046	0.052	0.008	**0.790**
D_{ZI}/π_{ZI}:	1.023*	1.027*	1.001	1.026*	1.029*	0.990	1.022*	1.003	0.999	1.015*	(1)	0.988

Values above the diagonal represent D_{xy} (in percent), while those below reflect F_{ST}. Bold values on the diagonal are π (%). The ratio of each population's genetic distance to the ZI sample versus diversity with the ZI sample is also given (bottom row). Ratios were corrected based on the (minor) predicted effects of sequencing depth for each population (see Materials and Methods). Ratios significantly greater than one (bootstrapping $P<0.001$) are noted (*). Admixture-filtered data from genomes with less than 15% estimated admixture were analyzed for each population that had two or more such genomes.

doi:10.1371/journal.pgen.1003080.t002

The ratio of relative π_{FR} for the X chromosome versus the inversion-free autosomes appears consistent with some previously explored founder event models [57] if chromosomes X and 2 are compared (ratio = 0.692, compared to a minimum of 0.669 in the cited study), but not if chromosome 3 is examined instead (ratio = 0.646). Some studies have concluded that the difference between X-linked and autosomal diversity reductions in cosmopolitan *D. melanogaster* exceeds the predictions of demographic models involving population bottlenecks and/or a shift in sex-specific variance in reproductive success [12,13]. Instead, the X chromosome's disproportionate diversity reduction might result from more efficient positive selection on this chromosome (due to male hemizygosity [40]) during the adaptation of cosmopolitan populations to temperate environments. However, it appears relevant that the above studies examined autosomal loci on chromosome 3, but not chromosome 2. Further theoretical, simulation, and inferential studies to elucidate the relative influence of selection, demography, and inversions on the X chromosome and autosomes is needed before their relative contribution to diversity in cosmopolitan *D. melanogaster* can be clearly understood.

Genetic structure and expansion history

Levels of genetic differentiation between populations were evaluated in terms of D_{xy} and F_{ST} [58] for each chromosome arm. In order to minimize the effects of any residual admixture in the filtered data, only genomes with admixture proportion below 15% were included. Populations with sufficient data for this analysis included CO, ED, FR, GA, GU, KR, NG, RG, TZ, UG, ZI, and ZS. Within Africa, F_{ST} values on the order of 0.05 were typical (Table 2). Geographically proximate population pairs often had lower F_{ST} (at minimum, a value of 0.009 between ZI and ZS). Comparisons involving the ED sample gave uniformly higher F_{ST} than other African comparisons (median 0.147), consistent with the loss of diversity observed for Ethiopian samples. As expected, comparisons of African samples with the European FR sample yielded the highest F_{ST} values (median 0.208). Genetic differentiation at putatively unconstrained short intron sites [59,60] showed similar patterns (Table S11), but as expected, magnitudes of D_{xy} and π were more than twice as high as for all non-centromeric, non-telomeric sites (for ZI, short intron $\pi = 0.0194$).

In order to assess the compatibility our data with a model of geographic expansion from southern Africa, we examined the ratio of each population's D_{ZI} (average pairwise genetic distance, or D_{xy}, between this population and the ZI sample) and π_{ZI}. This ratio will be near 1 if a population's genomes are no more divergent from ZI genomes than ZI genomes are from each other, consistent with the recent sampling of this population's diversity from a ZI-like ancestral population. In contrast, ratios exceeding 1 indicate that a population contains unique genetic diversity not present in ZI. Populations from eastern Africa (KR, RG, TZ, UG) and Europe (FR) had ratios compatible with a recent ZI-like origin (Table 2). However, populations from western Africa (CO, GA, GU, NG) and Ethiopia (ED) showed modest levels of unique variation. The highest ratio, for Guinea (GU), indicated a 2.9% excess of D_{ZI} over π_{ZI}. Elevated ratios could indicate a relatively ancient occupation of at least some of the above regions (perhaps on the order of tens of thousands of years). Alternatively, under the hypothesis of an expansion from southern Africa, these regions may have received a genetic contribution from a different part of a structured southern African ancestral range (*e.g.* migration into Gabon and western Africa from Angola, which also contains Miombo woodlands but has not been sampled).

Examination of genomewide genetic differentiation may also shed light on the sub-Saharan origins of cosmopolitan *D. melanogaster*. Geographic hypotheses for expansion of *D. melanogaster* from sub-Saharan Africa have ranged from a Nile route starting from the equatorial rift zone [18] to a more western crossing of the Sahara via formerly wetter areas of "Paleochad" [16]. A simple prediction is that the sub-Saharan samples most closely related to the cosmopolitan source population should show the lowest values of D_{xy} and F_{ST} relative to the cosmopolitan FR sample. However, even low levels of undetected cosmopolitan admixture in sub-Saharan genomes could

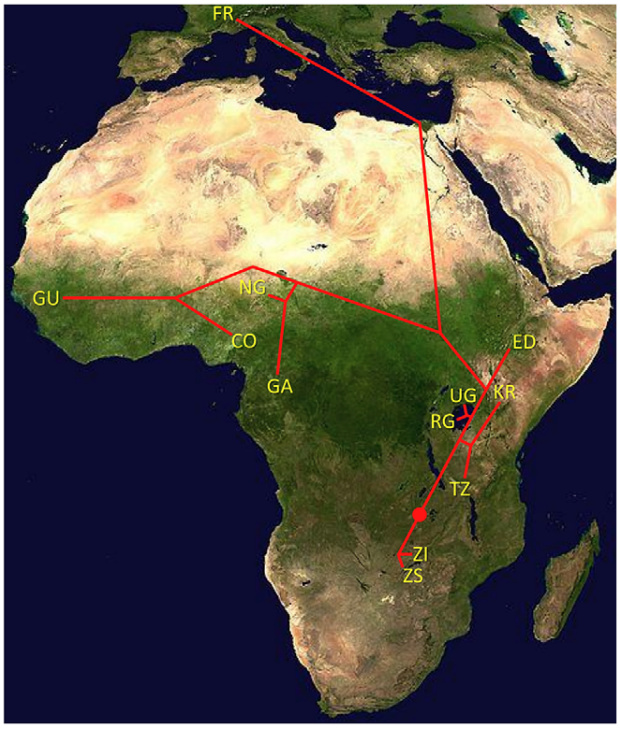

Figure 8. Topology of a neighbor-joining population distance tree based on the matrix of D_{xy} values (Table 2). Red dot indicates root based on midpoint rooting. Branch lengths are not to scale.
doi:10.1371/journal.pgen.1003080.g008

obscure this signal, and so only genomes with <15% detected admixture were considered below. Among the eleven African populations analyzed (see above), the Kenyan KR sample showed the lowest genome-wide D_{FR} (Table 2) and would have had the lowest FR F_{ST} if not for its anomalous pattern of variation for arm 2L (Table S11). However, KR is the sample with the highest proportion of detected cosmopolitan admixture, which clouds the interpretation of these results. After KR, the lowest D_{FR} values come from the western group of samples (NG, CO, GA, and GU), of which two (CO and GU) had relatively low levels of detected admixture. Despite its northeast sub-Saharan location, the Ethiopian ED sample does not appear to represent a genetic intermediate between cosmopolitan and other sub-Saharan populations, and may instead represent a separate branch of this species' geographic expansion. Further sampling and analysis may be needed to obtain compelling evidence regarding the geographic origin of cosmopolitan *D. melanogaster*.

One scenario for the sub-Saharan expansion of *D. melanogaster* is illustrated by the geographic fit of a simple neighbor-joining population tree based on D_{xy} values (Figure 8; Figure S8). This tree is consistent with the hypothesis of a southern Africa origin for *D. melanogaster*, with an initial expansion into eastern Africa, followed by offshoots reaching Ethiopia, the palearctic (northern Africa and beyond), and western Africa. Of course, even after the filtering of cosmopolitan admixture, a tree-like topology is not likely to fully describe the history of sub-Saharan *D. melanogaster* populations. However, the history described above seems consistent with levels and patterns of population diversity (Figure 6; Table 2), and may capture some important general features of the species' history.

Even if the general expansion history described above ultimately proves to be accurate, many historical details await clarification. Diversity differences among African populations could indicate population bottlenecks during a sub-Saharan range expansion, and population growth during such an expansion is also possible. Further analysis of population genomic data is also needed to establish whether ancestral range populations have also been affected by population growth [23] or a bottleneck [21]. Lastly, although migration within Africa has not erased the observed diversity differences and genetic structure, the historical and present magnitudes of such gene flow are not clear. The quantitative estimation of historical parameters may be addressed by detailed follow-up studies. However, for a species like *D. melanogaster*, in which very large population sizes may allow relatively high rates of advantageous mutation and efficient positive selection, one concern is that the effects of recurrent hitchhiking may be important on a genomewide scale [27,28,61,62]. Hence, the application of standard demographic inference methods to random portions of the *D. melanogaster* genome (or even putatively unconstrained sites) may yield estimates that are biased by violations of the assumption of selective neutrality. Under the assumption of demographic equilibrium, Jensen *et al.* [62] estimated a ~50% reduction in diversity due to positive selection for Zimbabwe *D. melanogaster*, and selective sweeps may have similarly important influences on the means and variances of other population genetic statistics as well. Hence, further methodological development may be needed before accurate demographic estimates can be obtained for species in which large population sizes facilitate efficient natural selection.

Influence of recombination and selection on genetic variation

Focusing on our largest population sample (22 primary core RG genomes), we investigated relationships between genetic diversity and mapping-based recombination rate estimates [28]. To minimize the effects of direct selective constraint on the sites examined, we focused on the middles of short introns (bp 8 to 30 of introns ≤65 bp in length), which are among the most polymorphic and divergent sites observed in the *Drosophila* genome [59,60]. Since each 23 bp intronic locus is too small to be considered individually, we show broad-scale patterns of diversity from all relevant sites within a given cytological band. Consistent with previous findings [9], strong relationships between recombination and variation were observed for all chromosome arms (Figure 9), with Pearson's r ranging from 0.68 (for 3L) to 0.95 (for 2R), with $P = 0.0005$ or lower for all arms (Table S12). Curiously, bp position along the chromosome arm was a stronger predictor of diversity than estimated recombination rate for arms 3L and 3R (Table S12), which could reflect imprecision in recombination rate estimates for chromosome 3, or the influence of polymorphic inversions on recombination in nature. Across all autosomal arms, the strongest correlation between recombination and diversity was for low rates of crossing-over (adjusted rate below 1 cM/Mb, equivalent to an unadjusted 2 cM/Mb rate, Pearson $r = 0.56$ and $P = 0.0002$). However, a strong correlation persisted above this threshold as well (Pearson $r = 0.44$, $P = 0.002$). Correlations within these categories were not significant for the X chromosome, potentially due to smaller numbers of chromosome bands, especially for the low recombination category ($n = 4$). Overall, the above results are consistent with the well-supported role for natural selection in reducing variation in regions of low recombination. However, the relative contributions of specific selection models such as hitchhiking [10] and background selection [11] to this pattern have not been quantitatively estimated.

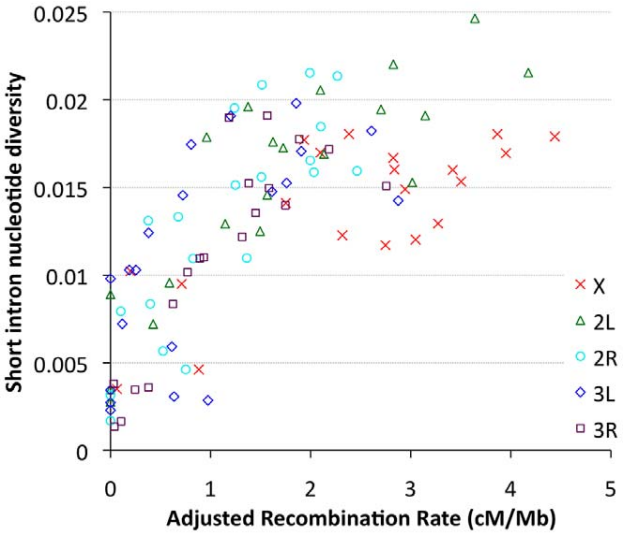

Figure 9. Nucleotide diversity versus recombination rate for short intron sites (bp 8–30 in <65 bp introns) is plotted by cytological band. Recombination rate estimates are from Langley et al. (2011), multiplied by one half for autosomes and two thirds for the X chromosome, and weighted by cytological sub-band recombination rate estimates and site counts.
doi:10.1371/journal.pgen.1003080.g009

Examining the RG sample's allele frequencies at short intron sites, we observed an excess of singleton polymorphisms (sites with a minor allele count of 1) for all chromosome arms relative to the predictions of selective neutrality and demographic equilibrium (Figure 10A). The degree of this excess varied among chromosome arms: compared to a null expectation of 31% singleton variants, the autosomal arms ranged from 33% to 37%, while the X chromosome had 44% singletons. The general excess of rare alleles could reflect population growth, as suggested for a Zimbabwe population sample [23], and growth has some potential to influence X-linked and autosomal variation differently [63]. Recurrent hitchhiking may contribute to the genomewide excess of rare alleles [64]. Under this hypothesis, the difference in singleton excess between the X chromosome could reflect more efficient X-linked selection due to hemizygosity [40]. Without a difference in the rate of X-linked and autosomal adaptation, this contrast could instead result from a greater fraction of X-linked selective sweeps acting on new beneficial mutations, with relatively more autosomal sweeps via selection on standing variation. The autosomes may have more potential to harbor recessive and previously deleterious functional variants, and sweeps from standing variation do not strongly influence the allele frequency spectrum [65].

We also used short intron allele frequencies to conduct a preliminary analysis of the relationship between recombination and rare alleles. Specifically, we tested whether the proportion of singletons among variable sites differed between low recombination regions (defined here as <1 cM/Mb) and moderate to high recombination regions (>1 cM/Mb). No clear relationship between recombination and allele frequency was observed above this cutoff (results not shown). For the 1 cM/Mb threshold, the X chromosome showed an elevated proportion of singletons in the low recombination category (53% *vs.* 43%; Pearson χ^2 $P = 0.032$; Figure 10B). Data from the autosomes are inconclusive: while three arms show non-significant trends toward more rare alleles in low recombination regions (Figure 10B), arm 2L showed a significant pattern in the opposite direction (30% *vs.* 36%; $P = 0.025$), possibly reflecting specific evolutionary dynamics of the 2L centromere-proximal region. The X chromosome result is qualitatively consistent with the predictions of the recurrent hitchhiking model [64] and some (but not all) previous findings from *D. melanogaster* [51,66,67]. Under this hypothesis, the lack of a comparable autosomal pattern might indicate a lesser influence of classic selective sweeps on the autosomes relative to the X chromosome, or a greater effect of inversion-related selection on the autosomes obscuring predictions of the recurrent hitchhiking model. Background selection may also increase the proportion of singletons [68,69], although a greater X-linked effect of background selection has not been suggested. Further study is needed to quantify the influence of positive and negative selection at linked sites on nucleotide diversity and allele frequencies in the *D. melanogaster* genome.

Linkage disequilibrium and its direction

Linkage disequilibrium (LD) was examined using a standard correlation coefficient (r^2) between single nucleotide polymorphism (SNP) pairs, and also via the directional LD metric r_ω [70,71]. The r_ω statistic is positive when minor frequency alleles at two sites tend to occur on the same haplotype, negative if they tend to be on

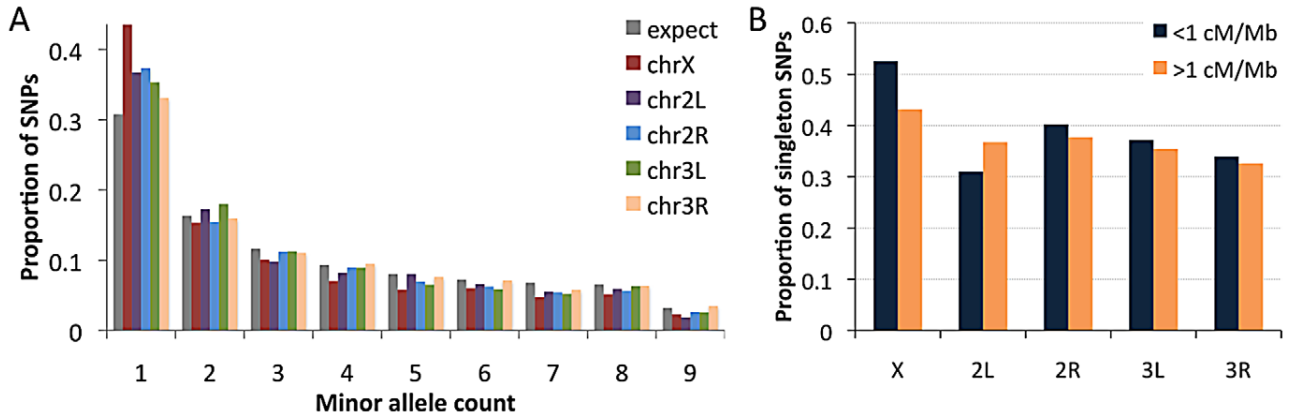

Figure 10. Allele frequencies for the RG sample (using a sample size of 18) at short intron sites. (A) The folded frequency spectrum for each chromosome arm. (B) Comparison of the proportion of SNPs with a minor allele count of 1 in regions of lower versus higher recombination.
doi:10.1371/journal.pgen.1003080.g010

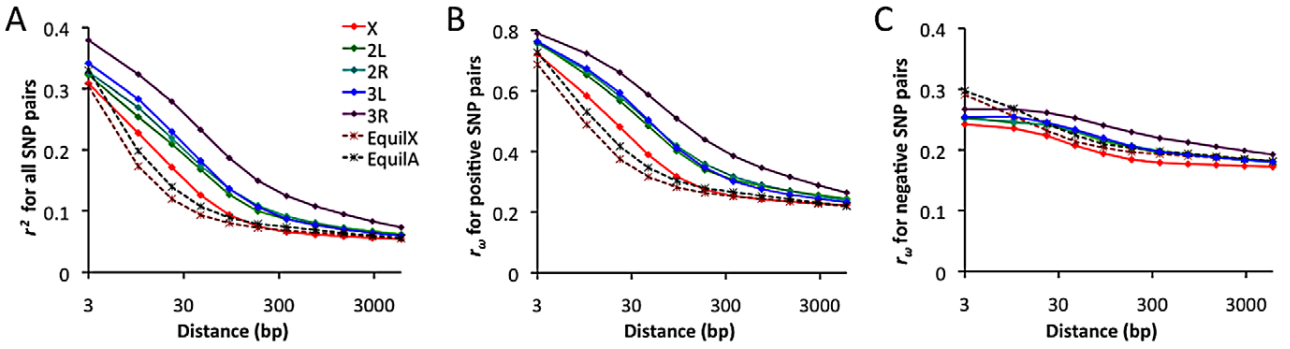

Figure 11. Linkage disequilibrium (LD), excluding singleton polymorphisms. Series refer to the observed LD for each major chromosome arm, and the expected LD from neutral equilibrium simulations for X-linked and autosomal loci, as given in panel A. (A) Average r^2 for a series of SNP pair distance bins. (B) Average r_ω for SNP pairs with positive LD. (C) Average r_ω for SNP pairs with negative LD.
doi:10.1371/journal.pgen.1003080.g011

different haplotypes. Although we lack a comprehensive understanding of the evolutionary forces capable of influencing r_ω, it is known that hitchhiking strengthens positive r_ω (since recombination near a selective sweep leaves groups of positively linked SNPs [72]), while negative r_ω may result from epistatic interactions among beneficial or deleterious alleles [70]. Empirical data from the RG sample was compared against neutral simulations with equilibrium demographic history. Importantly, equilibrium may not accurately reflect the history of the RG sample: recent population growth may have occurred, and the RG sample's modest diversity reduction compared to the ZI sample may imply a mild population bottleneck. Although the full effects of demography can not be eliminated by any simple procedure, we can reduce the influence of growth or other forces responsible for this population's excess of singleton polymorphisms by excluding singletons from the empirical and simulated data.

In general, an excess of LD was observed over neutral, equilibrium predictions for all chromosome arms (Figure 11A). The X chromosome's lower LD is consistent with its higher average recombination rate (54% higher for the regions examined [28]). The RG pattern contrasts with data from a North American population, which showed elevated X-linked LD [28,29] that likely reflects a stronger influence of demography and possibly selection on the X chromosome during the species' out-of-Africa expansion. For the RG sample, the X chromosome's LD excess was largely confined to the 10–100 bp scale. In contrast, autosomal arms showed an excess of LD at all scales 10 bp and above (Figure 11A). Since the simulations account for differences in average (inversion-free) recombination rate between the X and autosomes, the autosomes' more pronounced LD excess could result from a stronger influence of inversions on these arms. As noted above, the autosomes' higher inversion polymorphism should reduce autosomal recombination rates in nature and increase X chromosome recombination rates. Arm 3R contains the largest number of common inversions in Africa [41], and LD for this arm is by far the highest. Arm 3R's somewhat lower average recombination rate (7–27% lower than other autosomal arms for the analyzed regions) may contribute to this pattern as well. The above observations regarding LD are concordant with estimates of the population recombination rate for the RG sample, which are elevated for the X chromosome (in spite of its potentially lower population size) and reduced for 3R [73].

Notably, the observed LD excess is driven entirely by SNP pairs with positive r_ω (Figure 11B), while negative SNP pairs show no departure from equilibrium expectations (Figure 11C). Although cosmopolitan admixture has been largely removed from the analyzed data set, it remains possible that demographic events of this nature might inflate positive LD specifically. Inversions may well play a role in boosting positive LD, since inversion-associated polymorphisms may often be present at similar frequencies on the same haplotypes. However, given the excess of LD on all chromosome arms and on relatively short spatial scales, it is not yet clear whether inversions are a sufficient explanation. Recurrent hitchhiking may also contribute to the genome-wide excess of positive LD [72]. Further studies will be needed to evaluate the compatibility of specific hypotheses with genome-wide LD patterns.

Potential targets of selective sweeps in a Rwanda sample

Identifying the genes and mutations underlying Darwinian selection is an important aspect of evolutionary biology, and of population genomics in particular. The lack of a precise demographic model limits our ability to formally reject the null hypothesis of neutral evolution for specific loci, since certain demographic models can mimic the effects of selective sweeps [74]. However, we have still sought to learn about general patterns of directional selection in the genome by conducting a series of local outlier analyses to detect unusual patterns of allele frequencies within and between populations that are consistent with recent adaptive evolution. These outlier analyses necessarily involve a strong assumption about the proportion of the genome affected by selection. However, the enrichment analyses we perform on these outliers should be robust to some level of random false positives within the outliers, and should still be informative if not all adaptive loci are detected.

We searched for putative signals of selective sweeps in the RG sample using a modified version of the SweepFinder program [75,76] that looks for both allele frequency spectra and diversity reductions consistent with recent selective sweeps. As further described in the Materials and Methods section, we analyzed the RG data in windows and used the Λ_{max} statistic in an outlier framework, rather than making an explicit assumption regarding the appropriate demographic null model – as would have been required for typical simulations defining statistical significance. Here, we focus on the most extreme 5% of windows from each chromosome arm. After merging neighboring outlier windows, a total of 343 outlier regions were obtained (Table S13). For each outlier region, the gene with the closest exon to the Λ_{max} peak was recorded. Genes within extreme outlier regions included *Ankyrin 2* (cytoskeleton, axon extension), *Girdin* (actin filament organization,

Table 3. Gene ontology enrichment analysis based on outlier windows for high Λ_{max} in the Rwanda RG sample, indicating potential targets of recent selective sweeps.

Gene Ontology Category Description	Outlier Genes	Total Genes	*P* value
mRNA binding	22	120	0
microtubule associated complex	20	184	0
lipid particle	15	148	0
polytene chromosome	9	37	0
mRNA 3′-UTR binding	7	13	0
ribonucleoprotein complex	6	12	0
positive regulation of translation	4	7	0
heterochromatin	4	8	0
nuclear pore	6	21	0.0001
male meiosis	5	18	0.0001
SMAD protein import into nucleus	4	10	0.0001
ubiquitin-protein ligase activity	8	48	0.0002
precatalytic spliceosome	8	73	0.0002
nuclear mRNA splicing, via spliceosome	9	102	0.0003
nucleus	65	699	0.0004
polytene chromosome puff	4	15	0.0005
regulation of alternative nuclear mRNA splicing, via spliceosome	7	34	0.0006
neurogenesis	22	316	0.0008
female meiosis chromosome segregation	5	21	0.0013
positive regulation of transcription, DNA-dependent	9	34	0.0014
DNA binding	25	248	0.0017
salivary gland cell autophagic cell death	8	39	0.0023
protein ubiquitination	4	15	0.0024
nucleoplasm	4	14	0.0028
spermatogenesis	8	56	0.0032
mitotic cell cycle	4	17	0.0036
regulation of apoptotic process	4	11	0.0039
regulation of mitosis	4	18	0.0049
chromatin organization	4	13	0.0051
negative regulation of transcription, DNA-dependent	9	41	0.0057
cytokinesis	7	42	0.0059
phagocytosis, engulfment	13	112	0.0068
autophagic cell death	7	36	0.0073
protein complex	7	36	0.0079
fusome	5	16	0.0083
nuclear envelope	4	18	0.0086

Listed are GO categories with $P<0.01$ and outlier genes >3. Full results are given in Table S14.
doi:10.1371/journal.pgen.1003080.t003

regulation of cell size), *Laminin A* (behavior, development, meiosis), *narrow abdomen* (ion channel, circadian rhythm), *Odorant receptor 22a* [77], and ribosomal proteins *S2* and *S14b* (separate regions). Several strong outliers corresponded to genes also implicated in a recent genome scan based on outliers for low polymorphism relative to divergence [28], including *bendless* (axonogenesis, flight behavior) *CENP-meta* (mitotic spindle organization, neurogenesis), *female sterile (1) homeotic* (regulation of transcription), *Heterogeneous nuclear ribonucleoprotein at 27C* (regulation of splicing), *loquacious* (RNA interference, nervous system development, germ-line stem cell division), and *no distributive junction* (meiotic chromosome segregation).

Despite a similar number of outlier regions as the F_{ST} analyses described below, the Λ_{max} scan yielded a much larger number of significantly enriched gene ontology categories: 115 categories had $P<0.05$ based on random permutation of target windows within chromosomal arms (Table 3; Table S14). Consistent with previous results from a population genomic outlier analysis of diversity and divergence [28], numerous biological processes related to gene regulation were observed, including positive and negative regulation of transcription, positive regulation of translation, regulation of alternative splicing, mRNA cleavage, chromatin organization, regulation of chromatin silencing, and gene silencing. Many enriched cellular components (*e.g.* nucleus, precatalytic

spliceosome, mRNA cleavage and polyadenylation complex, ribonucleoprotein complex, heterochromatin, and euchromatin) and molecular activities (*e.g.* DNA binding, mRNA binding and especially mRNA 3′-UTR binding) were also consistent with a broad importance for regulators of gene expression in recent adaptive evolution. A number of the GO terms listed in Table 3 were also reported from the above-mentioned genome scan [28], including negative regulation of transcription, positive regulation of translation, ribonucleoprotein complex, precatalytic spliceosome, protein ubiquitination, nuclear pore, lipid particle, and spermatogenesis. Other enriched biological processes included oogenesis, neurogenesis, male meiosis and female meiosis chromosome segregation, regulation of mitosis and apoptosis, and phagocytosis. Additional cellular components included microtubule-associated complex, kinetochore, and fusome while enriched molecular activities also included ATP binding and voltage-gated calcium channel activity.

Locally elevated genetic differentiation between African populations

Nine African population samples with larger sample sizes after admixture filtering were included in an analysis of local genetic differentiation. F_{ST} was evaluated for each pair of populations, and the mean F_{ST} for each window was noted. Examination of the 2.5% highest mean F_{ST} values for each chromosome arm and the merging of neighboring outlier windows resulted in 294 outlier regions (Table S15). For each outlier region, the gene with the closest exon to the center of the most extreme window was noted. Genes associated with unusually strong F_{ST} outlier regions included *Odorant receptor 22b* (tandem paralog of the above-mentioned *Or22a*), *Cuticular protein 65Au*, *Dystrophin*, *P-element somatic inhibitor*, and *CG15696* (predicted homeobox transcription factor). Of course, many of the strongest putative signals of adaptive differentiation are wide, and further investigation will be needed to confirm specific targets of selection. Permutation of putative target windows indicated that genes from 34 GO categories were significantly over-represented among our outliers at the $P = 0.05$ level (Table 4; Table S16). These GO categories included biological processes (*e.g.* oocyte cytoskeleton organization, regulation of alternative splicing, regulation of adult cuticle pigmentation), cellular components (*e.g.* mitochondrial matrix, dendrite), and molecular functions (*e.g.* olfactory receptor activity, mRNA binding).

Locally elevated genetic differentiation between Africa and Europe

A windowed F_{ST} outlier approach was also applied to detect loci that may contain adaptive differences between sub-Saharan (RG) and European (FR) populations.

Some of these loci might have had adaptive importance during the expansion of *D. melanogaster* into temperate environments, but others could reflect recent selection within Africa. A total of 346 outlier regions resulted from analyzing the upper 2.5% tail of Rwanda-France F_{ST} (Table S17). Genes associated with strong F_{ST} outliers included *Or22a* (which may be under selection in Africa, see above), *CHKov1* (insecticide and viral resistance [78,79]), *ACXC* (spermatogenesis), and *Jonah 98Ciii* (digestion), plus a number of genes involved in morphological and/or nervous system development (*e.g. Bar-H1*, *Death-associated protein kinase related*, *Enhancer of split*, *hemipterous*, *highwire*, *mastermind*, *rictor*, *sevenless*, *Serendipity δ*, and *wing blister*). Other genes at the center of strong outlier regions were also detected by a genome-wide analysis of diversity ratio between U.S. and Malawi populations [28], including *dpr13* (predicted chemosensory function), *Neuropeptide Y receptor-like*, *rugose* (eye development), and *Sno oncogene* (growth factor signaling, neuron development).

The genes identified in this analysis still yielded 31 significantly enriched GO categories (Table 5; Table S18). Biological processes among these GO categories included chromosome segregation, locomotion, female germ-line cyst formation, histone phosphorylation, and alcohol metabolism. Cellular components included basal lamina and polytene chromosome interband, while molecular activities included transcription coactivators and neuropeptide receptors. The detected GO categories were essentially distinct from those obtained from the diversity ratio analysis of Langley *et al.* [28]. The lack of overlap may stem at least partially from differences in the statistics and populations used in each analysis. The well-known challenges of identifying positive selection in the presence of bottlenecks [74], along with uncertainty regarding the portion of the genome affected by adaptive population differences, may also contribute to these findings. Both analyses, however, should motivate new adaptive hypotheses to be tested via detailed population genetic analyses and experimental approaches.

If the rapid introgression of non-African genotypes into African populations documented above is driven by natural selection, then sharp peaks and valleys of admixture along the genome (Figure 4) should contain functional differences between sub-Saharan and cosmopolitan populations. Such differences may have been driven by natural selection after these populations diverged, and hence may be detectable by the Africa-Europe F_{ST} outlier scan presented above. Given that the scale of these F_{ST} outliers (on the order of 10 kb) is narrower than our admixture peaks and valleys (on the order of 100 kb), population genetic signals of elevated differentiation may be helpful in localizing genes responsible for driving or opposing non-African gene flow into African populations.

We selected eight clear genomic peaks of admixture within the higher recombination regions analyzed for F_{ST}. These peaks were delimited by windows containing the local maximum number of admixed genomes, and identified F_{ST} outlier regions that either overlapped them or were within 100 kb. Valleys of admixture were more difficult to clearly distinguish from gaps between peaks and minor fluctuations (Figure 4) – three were identified, one of which overlapped several F_{ST} outlier regions (Table S19). For peaks of admixture, seven of these eight regions were associated with F_{ST} outlier regions (Table S19), exceeding random expectations (permutation $P = 0.017$). Stronger outlier regions associated with admixture peaks included the genes *Bar-H1*, *Enhancer of split*, *Neuropeptide Y receptor-like*, and *sevenless*. Further studies will be needed to evaluate the possibility that cosmopolitan alleles at one or more of these loci may now confer a fitness advantage in urban African environments.

Conclusions and Prospects

Here, we have described variation across more than one hundred *D. melanogaster* genomes, focusing on the species' sub-Saharan ancestral range. We observed clear evidence of cosmopolitan admixture at varying levels in all sub-Saharan populations. While admixture initially appeared to be merely a barrier to studying African variation, inferred patterns of admixture suggested that this process is associated with intriguing biological dynamics. Based on the apparent speed of introgression, the association of admixture with urban environments, and dramatically differing admixture levels across the genome (with peak admixture levels correlated with outliers for Africa-Europe F_{ST}), it appears that admixture may be a primarily non-neutral process. Unexpected variance in admixture proportion within populations provides another departure from simple models, and could indicate isolation mechanisms within African populations.

Table 4. Gene ontology enrichment analysis based on outlier windows for high mean F_{ST} for African population comparisons.

Gene Ontology Category Description	Outlier Genes	Total Genes	*P* value
DNA-directed RNA polymerase activity	3	17	0.00103
oocyte microtubule cytoskeleton organization	3	7	0.0033
regulation of alternative nuclear mRNA splicing, via spliceosome	6	33	0.00416
olfactory receptor activity	6	32	0.00419
mitochondrial matrix	4	31	0.00485
positive regulation of protein phosphorylation	2	5	0.00519
regulation of adult chitin-containing cuticle pigmentation	3	8	0.00638
regulation of R8 cell spacing in compound eye	3	4	0.00786
notum cell fate specification	3	3	0.00834
receptor signaling protein serine/threonine kinase activity	4	17	0.0096
regulation of nuclear mRNA splicing, via spliceosome	2	6	0.01232
sensory perception of smell	8	49	0.01485
RNA polymerase II transcription cofactor activity	2	12	0.01549
mRNA binding	13	114	0.01581
mediator complex	2	14	0.01648
nucleobase-containing compound metabolic process	2	6	0.01681
SMAD protein import into nucleus	2	10	0.01889
transcription from RNA polymerase II promoter	3	20	0.01932
lipid particle	8	138	0.02089
muscle cell homeostasis	3	7	0.02217
spermatocyte division	2	6	0.02634
embryonic axis specification	2	5	0.02869
cytosolic small ribosomal subunit	2	24	0.02915
haltere development	2	3	0.03257
MAPK cascade	2	8	0.03302
mucosal immune response	2	5	0.03376
odorant binding	6	61	0.03402
dendrite	3	19	0.03486
small nuclear ribonucleoprotein complex	2	22	0.03584
notum development	2	3	0.03783
neurexin family protein binding	2	2	0.03902
induction of apoptosis	2	9	0.0406
myofibril assembly	2	4	0.04232
oocyte axis specification	3	11	0.04339

Listed are GO categories with $P<0.05$ and outlier genes >1. Full results are given in Table S16.
doi:10.1371/journal.pgen.1003080.t004

We observed the greatest genetic diversity in a Zambian sample and nearby populations, suggesting a possible geographic origin for the species. Even at a broad genomic scale, however, it appears that genetic diversity does not always reflect demographic expectations. We observed chromosome arm-specific deviations in population diversity ratios, most notably for comparisons involving the European population: genetically differentiated inverted chromosomes strongly influence autosomal diversity in our France sample, potentially due to recent natural selection elevating the frequency of introgressing inversions with African origin. Considering this hypothesis alongside our admixture inferences, it is conceivable that selection has driven gene flow in both directions across the sub-Saharan/cosmopolitan genetic divide, with consequences for genome-wide levels and patterns of diversity. Additional studies are needed to evaluate models of population history, natural selection, and inversion polymorphism that may account for the above patterns.

We have identified numerous genes and processes that may represent targets of positive selection within and between populations. However, further investigations will be needed to confirm targets of selection and their functional significance. Such studies may help reveal the biological basis of this species' adaptation to temperate environments, as well as contrasting environments within Africa, while potentially also providing more general insights into the genetic basis of adaptive evolution.

Although the aims of this publication are primarily descriptive, data such as that presented here may play an important role in resolving some long-standing controversies in population genetics. It's clear that natural selection plays an important role in shaping sequence divergence between *Drosophila* species and in reducing

Table 5. Gene ontology enrichment analysis based on outlier windows for high F_{ST} between Rwanda and France population samples.

Gene Ontology Category Description	Outlier Genes	Total Genes	*P* value
chromosome segregation	5	20	0.00106
dephosphorylation	3	11	0.00315
digestion	2	4	0.00389
locomotion	4	8	0.00601
basal lamina	3	5	0.00675
polytene chromosome interband	3	17	0.0087
pyruvate metabolic process	2	6	0.00879
female germ-line cyst formation	2	3	0.01018
GTPase activity	8	76	0.01732
regulation of protein localization	2	4	0.01821
tissue development	2	4	0.01851
iron ion binding	3	17	0.01888
organ morphogenesis	2	4	0.01895
FMN binding	2	7	0.02285
actin filament bundle assembly	3	8	0.02419
histone phosphorylation	2	5	0.02495
nucleus localization	2	3	0.0271
germ cell development	4	19	0.03221
eye development	3	6	0.03279
ATPase activity, coupled	6	40	0.03319
alcohol metabolic process	3	12	0.03355
organic anion transport	2	7	0.03944
metal ion binding	5	44	0.0395
organic anion transmembrane transporter activity	2	7	0.03967
mitotic cell cycle	3	17	0.04069
transcription coactivator activity	3	8	0.04122
larval chitin-based cuticle development	2	5	0.04311
lipid particle	8	138	0.04768
anion transport	2	4	0.04834
neuropeptide receptor activity	8	30	0.04931
choline dehydrogenase activity	3	13	0.04994

Listed are GO categories with $P<0.05$ and outlier genes >1. Full results are given in Table S18.
doi:10.1371/journal.pgen.1003080.t005

polymorphism in genomic regions of low recombination. However, the relative importance of natural selection and neutral forces in governing levels and patterns of variation in regions of higher recombination is unresolved. We still do not know if, for example, linked hitchhiking events have an important influence on diversity at most sites in the genome. The relative impact of population history and natural selection on genetic diversity during the out-of-Africa expansion of *D. melanogaster* is also uncertain. And in regions of low recombination, the relative contributions of hitchhiking and background selection in reducing genetic variation have not been quantified. It is our hope that population genomic data sets like this one will motivate theoretical and simulation studies that advance our fundamental understanding of how evolutionary forces shape genetic variation.

Note added in proof

Consensus sequences with reduced reference bias are now available from http://www.dpgp.org/dpgp2/dpgp2.html

Materials and Methods

Drosophila stocks and DNA preparation

Genomes reported here are derived from the population samples listed in Table S1 and depicted in Figure 1. The collection methods for samples collected in 2004 or later correspond to a published protocol [80]. Information about individual fly stocks is presented in Table S2. Most of the relevant stocks are isofemale lines, each founded from a single wild-caught female. In some cases, intentional inbreeding was conducted by sib-mating for five generations; such lines have an 'N' appended to the isofemale line label. Although not a focus of our analysis, we have also released genomic data from a small number of chromosome extraction lines, created using balancer stocks.

Except for the three ZK genomes, DNA for all inbred and isofemale lines was obtained from haploid embryos [30]. Briefly, a female fly from the stock of interest was mated to a male homozygous for the *ms(3)K81*^{1} allele [81]. This mating produces

some eggs which are fertilized but fail to develop because the clastogenic paternal genome. Rarely, such eggs bypass apparent checkpoints and develop as haploid embryos. Eggs with partially developed first instars were visually identified under a microscope. DNA was isolated from haploid embryos and genome-amplified as previously described [30]. For the ZK genomes and chromosome extraction lines, DNA was isolated from 30 adult flies (generally females; mixed sexes in the case of autosomal extraction lines). For all samples, library preparation for sequencing (ligation of paired end adapters, selection of ~400 bp fragments, and PCR enrichment) was conducted as previously described [30]. In some cases, bar code tags (6 bp) were added to allow multiplexing of two or more genomes in one flow cell lane.

Sequencing, assembly, and data filtering

Sequencing was performed using standard protocols for the Illumina Genome Analyzer IIx. Initial data processing and quality analysis was performed using the standard Illumina pipeline. Sequence reads were deposited in the NIH Short Read Archive as project SRP005599. Alignments to the *D. melanogaster* reference genome (BDGP release 5) using BWA version 0.59 [31] with default settings and the "-I" flag. Program defaults included a 32 bp seed length; reads could therefore map to the reference only if two or fewer reference differences were present within a seed. Although read lengths varied from 76 bp to 146 bp within this data set, only the first 76 bp of longer reads was used for the assemblies reported here. In order to exclude ambiguously mapping reads, those with a BWA mapping quality score less than 20 were eliminated from the assemblies.

Consensus sequences for each assembly were obtained using the SAMtools (version 0.1.16) pileup module [32]. These diploid consensus sequences generally included a few thousand heterozygous calls, scattered across the genome. Such sites are not expected to represent genuine heterozyosity in these haploid/homozygous samples (with the exception of ZK, in which large-scale heterozygosity was observed, presumbaly due to incomplete inbreeding). All putatively heterozygous sites were masked to 'N'. Sites within 5 bp of a consensus indel were also masked to 'N' – this criterion was found to reduce errors associated with indel alignment; no appreciable benefit was observed if 10 bp was masked instead (data not shown).

Data were only considered for "target" chromosome arms, as defined in Table S1. These are chromosome arms expected to derive from the population sample of interest (as opposed to originating from laboratory balancer stocks), and observed to be free of heterozygous intervals. Chromosome arms were further defined as "focal" (the genomic regions analyzed here, namely the euchromatic portions of X, 2L, 2R, 3L, and 3R) or "non-focal" (the mitochondria and heterochromatin, including chromosomes 4 and Y). The assemblies analyzed here were defined as "release 2" data and are available for download at http://www.dpgp.org/dpgp2/DPGP2.html. Assemblies of mitochondrial and bacterial symbiont genomes are reported and analyzed separately [47].

Estimation of consensus error rate

Although the above assemblies provide nominal quality scores, we performed a separate evaluation of statistical confidence in the accuracy of assemblies. This analysis utilized five haploid embryo, reference strain (y^1 cn^1 bw^1 sp^1) genomes resequenced with comparable depth and read characteristics as the non-reference genomes reported here (Table S2). In order to simulate the effects of genetic variation, the maq fakemut program [82] was used to introduce artificial substitutions and indels into the resequenced reference genomes. Substitutions were introduced at rate 0.012/bp, while 1 bp indels were introduced at rate 0.0024/bp. Alignment and consensus sequence generation was then performed as described above.

The artificially mutagenized reference genomes allowed us to examine the tradeoff between minimizing error rates and maximizing genomic coverage. Based on the joint pattern of these quantities for various nominal quality scores (Figure S1), we selected a nominal quality threshold of Q31 as the basis for downstream analyses. The observed consensus sequence error rate for the nominal Q31 cutoff suggested was equivalent to an average Phred score of Q48 (roughly one error per 100 kb).

Detection of identical-by-descent genomic regions

Long tracts of identity-by-descent (IBD) between genomes may result from the sampling of related individuals. Because such relatedness violates the assumptions of many population genetic models, we sought to identify and mask instances of IBD caused by relatedness. Target chromosomes from all possible pairs of genomes were compared to search for long intervals of identity-by-descent (IBD) that may result from close relatedness. Following Langley *et al.* [28], windows 500 kb in length were moved in 100 kb increments across the genome, and sequence identity was defined as less than 0.0005 pairwise differences per site. A large number of pairwise intervals fit this criterion (Table S3). Some chromosomal intervals, including centromeres and telomeres, had recurrent IBD in between-population comparisons (Table S4). Cross-population IBD occurred at scales up to 4 Mb within these manually delimited "recurrent IBD regions", and its occurrence between different populations suggests that processes other than close relatedness are responsible. Such intervals were not masked from the data. We identified clear instances of "relatedness IBD" between two genomes when within-population IBD exceeded the scale observed between populations: when more than 5 Mb of summed genome-wide IBD tracts occurred outside recurrent IBD regions, or when tracts greater than 5 Mb overlapped recurrent IBD regions. Only nine pairs of genomes met one or both of these criteria (Table S4), and two of these pairs were expected based on the common origin of isofemale and chromosome extraction lines (Table S2). For these pairs, one of the two genomes was chosen for filtering, and all identified IBD intervals from this pairwise comparison were masked to 'N' for most subsequent analyses.

Admixture detection method—overview

Relevant for the inference of non-African admixture is a panel of eight primary core genomes from France (the "FR" sample). *D. melanogaster* populations from outside sub-Saharan Africa show reduced genetic diversity and are more closely related to each other than to sub-Saharan populations [22,26]. Hence, whether admixture came from Europe or elsewhere in the diaspora, FR should represent an adequate "reference population" for the source of non-African admixture. However, we lack an African population that is known to be free of admixture. And while a variety of statistical methods exist for the detection of admixture, options for detecting unidirectional admixture using a single reference population are more limited. We therefore developed a new method to detect admixture in this data set.

We constructed a windowed Hidden Markov Model (HMM) machine learning approach based on a given haplotype's average pairwise divergence from the non-African reference population (D_{FR}). The admixed state is based on comparisons of individual FR haplotypes to the remainder of the FR sample. The non-admixed state is based on comparisons of haplotypes from a provisional "African panel" to the FR sample. Here, 22 genomes from the Rwanda "RG" sample are used as the African panel. We allow for

the possibility of admixture within the African panel as described below.

Formally, the emissions distribution for the non-admixed state was constructed as follows. For each window, each RG haplotype was evaluated for average pairwise divergence from the FR sample ($D_{RG,FR}$). Each of these values was rescaled in terms of standard deviations of $D_{RG,FR}$ from the window mean $D_{RG,FR}$. Standardized values were added to the emissions distribution in bins of 0.1 standard deviations, and these bins were ultimately rescaled to sum to 1. Hence, the emissions distribution reflects the genome-wide pattern of $D_{RG,FR}$, accounting for local patterns of diversity.

The emissions distribution for the admixed state was constructed similarly. For each window, each FR haplotype was evaluated for average pairwise divergence from the remainder of the FR sample ($D_{FR,FR}$). However, these $D_{FR,FR}$ values were still rescaled by the window mean and standard deviation of $D_{RG,FR}$. An alternative version of the method in which the admixed state's emissions distribution was instead rescaled by the local mean and standard deviation of $D_{FR,FR}$ was slightly less accurate when applied to simulated data.

Given these genome-wide emissions distributions, we can examine $D_{RG,FR}$ for each African allele for each window, and obtain its likelihood if we are truly making an "Africa-Europe comparison" with this $D_{RG,FR}$ (non-admixed state) or if we are actually making a "Europe-Europe comparison" (admixed state). These likelihoods form the input for the HMM process, which was performed using an implementation [83] of the forward-backward algorithm. A minimum admixture likelihood of 0.005 was applied to HMM input, in order to reduce the influence of a single unusual window. Admixed intervals were defined as windows with >50% posterior probability for the admixed state. For the purpose of masking admixed genomic intervals for downstream analyses, one window on each side of admixed intervals was added (to account for uncertainty in the precise boundaries of admixture tracts).

Admixture detection method—validation

The admixture detection method was tested using simulated data containing known admixture tracts. Population samples of sequences 10 Mb in length were simulated using MaCS [84], which can approximate coalescent genealogies across long stretches of recombining sequence. Demographic parameters were based on a published model for autosomal loci [13,23]. The command line used was "./macs04 63 10000000 -s 12345 -i 1 -h 1000 -t 0.0376 -r 0.171 -c 5 86.5 -I 2 27 36 0 -en 0 2 0.183 -en 0.0037281 2 0.000377 -en 0.00381 2 1 -ej 0.00381000001 2 1 -eN 0.0145 0.2", specifying simulations with present population mutation rate 0.0376 and population recombination rate 0.171, gene conversion parameters based on a weighted average of loci from Yin *et al.* [85], and historic tree retention parameter $h = 1000$ [84].

The above simulations generate population samples that may resemble data from sub-Saharan and cosmopolitan populations of *D. melanogaster*, but they do not involve any admixture. If admixture was specified with the command line, then without modifications to the simulation program, there would not be an output record of admixture tract locations. Instead, extra "non-African" haplotypes were simulated (one for each African haplotype), and these "donor alleles" became the source for admixture tracts which were spliced into the African population's data after MaCS simulation was completed.

The locations and lengths of admixture tracts were determined by a separate simulation process. The forward simulation program developed by Pool and Nielsen [36] accounts for drift, recombination, and migration, recording intervals with migrant history. By using this program to simulate a region symmetric to the African MaCS data, we identified intervals that should contain admixture tracts after g generations of admixture. These intervals were then spliced from the non-African donor alleles into African haplotypes from the MaCS simulated polymorphism data.

The simulated data with admixture was then analyzed using the admixture HMM method described above. In this case, windows of 10 kb were analyzed. Times since the onset of admixture (g) of 100, 1000, and 10000 generations were examined. Migration rates were specified to approximate a total admixture proportion of 10% (hence testing the robustness of the method to this level of admixture in the "African panel").

As indicated by representative simulation results shown in Figure S8, the admixture detection method was highly accurate for $g = 100$ and $g = 1000$, and moderately accurate for $g = 10000$. Based on preliminary observations from the data, we suspected that much of the admixture in our data set was on the order of $g = 100$ or less.

Admixture detection method—implementation

The admixture HMM was initially applied to the RG sample alone. Compared with the simulated data, the empirical data showed more overlap between the admixed and non-admixed emissions distributions. This contrast could result from demographic differences between the African population used here (from Rwanda) and the one from which demographic parameter estimates were obtained (from Zimbabwe), and/or an effect of positive selection making Africa-Europe diversity comparisons more locally heterogeneous than expected under neutrality. We responded by expanding the window size used in the empirical data analysis. Windows were based on numbers of non-singleton polymorphic sites among the 22 RG primary core genomes. We chose a window size of 1000 such SNPs, which corresponds to a median window size close to 50 kb. Smaller windows led to noisier likelihoods (results not shown), while larger windows might exclude short admixture tracts without an appreciable gain in accuracy.

Another concern regarding the empirical data was the effect of sequencing depth on pairwise divergence values. After restricting the admixture analysis to genomes with >25X mean depth, we still observed a minor degree of "wavering" in admixture probabilities for genomes with the lowest depth. We therefore applied a simple correction factor to approximate each genome's quality effects on divergence metrics. In theory, we wish to know the effect of depth and other aspects of quality on D_{FR}. In practice, however, genomes differ in D_{FR} in part based on their level of admixture. Instead, D_{RG} (average pairwise divergence from the rest of the Rwanda sample) was used as a proxy. For each chromosome arm, a genome's D_{RG} was compared to the RG population average. Each genome's D_{FR} was then multiplied by the correction factor $\frac{\overline{D_{RG}}}{D_{RG}}$. Following this correction, no effect of depth on admixture inferences was observed within the primary core data set.

Although simulations suggested that our admixture method is robust to ~10% admixture in the African panel, we sought to maximize the method's accuracy by applying it iteratively to the RG sample. Identical-by-descent regions (as defined above) were masked during the creation of emissions distributions, but likelihoods were then evaluated for full RG chromosome arms. After one full "round" of the method (emissions, likelihoods, and HMM), admixture tracts were masked from the RG sample. This masked RG sample became the revised African panel for a second round of analysis, this one with a more accurate emissions

distribution for the non-admixed state (since it contains more true "Africa-Europe" comparisons, and is presumably less influenced by admixture). Admixture masking for RG was redone based on round 2 admixture intervals, and the re-masked RG data was used to create a third and final set of emissions distributions. The round 3 emissions distributions were used to generate final admixture calls not only for the RG sample, but also for the other African genomes in the primary core data set.

The use of RG as an "African panel" when examining admixture in other African populations is not without concern. Fortunately, in addition to being the largest African sample, RG also occupies a genetically intermediate position within Africa (see results section), which reduces the potential impact of genetic structure on the accuracy of admixture inferences for non-RG genomes. It also appears that aside from the effects of admixture, no other African sample has a much closer relationship to FR than RG does (see results section), thus mitigating a potential source of bias.

Analysis of admixture detection results

Standard linear regression was used to investigate the possible relationship between cosmopolitan admixture proportion (for a population sample) and the human population size of the collection locality (city, town, or village population size). Census-based population estimates were obtained from online sources for 15 of 20 population samples. For the remainder, satellite-based estimates were obtained from fallingrain.com (Table S1). While a set of uniform and perfectly accurate population figures is not available for these locations, the estimates used here may still allow a significant effect of human population size on cosmopolitan admixture proportion to be detected.

The centiMorgan length of each admixture interval was calculated based on recombination rates inferred from smoothed genetic map data [28]. The extra buffer windows added to each side of conservative admixture tract delimitations described above were not included in these length estimates. CentiMorgan tract lengths were then used with a method [36] that estimates three parameters of a migration rate change model: the current migration rate, the previous migration rate, and the time of migration rate change. A minimum detectable tract length of 0.5 cM was chosen, corresponding to roughly 200 kb or 4 windows on average. Forward simulations [36] including recombination, migration, and drift were performed under the estimated demographic model. Simulated data were compared to empirical data, to test how often simulated variance in cosmopolitan admixture proportion exceeded that observed in the RG sample.

Genetic diversity and structure of populations

Regions of lower recombination proximal to centromeres and telomeres were excluded from most analyses, except where indicated below. Recombination rates were taken from mapping-based estimates [28], and the threshold between "low" and "high" recombination rates was defined as 2×10^{-8} cross-overs per bp per generation. In most cases, a single transition point was apparent where a chromosome arm transitioned from low to high recombination, moving away from a centromere or telomere. A few narrow "valleys" of recombination rate estimates slightly below this threshold within broader high recombination regions, along with one peak of recombination rate slightly above this threshold close to the 3L centromere, were ignored in the definition of centromere-proximal and telomere-proximal boundaries. "Mid-chromosomal intervals" reflecting the higher recombination intervals used in this analysis for each chromosome arm were: X:2,222,391–20,054,556, 2L:464,654–15,063839, 2R:9,551,429–20,635,011, 3L:1,979,673–12,286,842, 3R:12,949,344–25,978,664.

Principal components analysis (PCA) was conducted using the method of Patterson *et al.* [43]. Mid-chromosomal data from all primary core genomes were included. The analysis was run twice, on data sets with and without admixture filtering. Applying additional filters (excluding sites with >5% missing data or <2.5% minor allele frequency) had little effect on results.

Nucleotide diversity (π) was initially calculated in 100 kb windows, and weighted values for each population sample (based on the number of sites in each window with data from at least two genomes) were then averaged to obtain a population's mean absolute π for each chromosome arm. Relative π was calculated by obtaining the ratio of window π from a given population versus that for the RG population (the largest African sample), and window ratios were weighted by the number of sites with data from two or more RG genomes. Relative π values should therefore be robust to cases where a population has large blocks of masked data in a genomic region with especially high or low diversity (since π in each window is standardized by that observed for the RG sample), which could bias estimates of absolute π. Genome-wide relative π was calculated as the unweighted average value of the five major chromosome arms. Three samples (CK, RC, SP) had only one primary core genome, but one or more secondary core genomes. Relative π for these samples was calculated based on comparisons between primary and secondary core genomes, both for the target sample and for RG (which also contains primary and secondary core genomes). A similar re-estimation of relative π for the CO sample yielded genome-wide relative π of 0.914 from primary-secondary comparisons, versus 0.927 from primary core genomes only.

D_{xy}, the average rate of nucleotide differences between populations, was calculated for a subset of populations with high levels of genomic coverage in the admixture-filtered data (CO, ED, FR, GA, GU, KR, NG, RG, TZ, UG, ZI, ZS). F_{ST} was calculated using the method of Hudson *et al.* [58], with equal population weightings regardless of their sample sizes. Arm-wide and genome-wide estimates of both statistics were calculated as described above for relative π.

Using the above summary statistics, we calculated the ratio of a population's D_{ZI} (genetic distance from the four Zambia ZI genomes) to π_{ZI}. Here, the intention was to test which populations contained unique genetic diversity not observed in the maximally diverse ZI population, leading to ratios greater than one. The significance of ratios greater than one was assessed via a bootstrapping approach. Windows 100 kb in length were sampled with replacement until 667 were drawn, to match the number present in non-centromeric, non-telomeric regions of the empirical data. One million such replicates were conducted for each population, and the proportion of replicates with a ratio less than one became the bootstrapping P value. The use of windows much larger than the scale of linkage disequilibrium implies a conservative test.

For each population's genome-wide relative π (Figure 6), and for the D_{ZI} to π_{ZI} ratio (Table 2; described), we applied a correction factor to reduce the predicted influence of sequencing depth on these quantities. From a linear regression of primary core genomes' sequencing depth *versus* D_{ZI} (Figure 2), the slope and y intercept of this relationship were obtained. Based on population mean sequencing depth, a population's predicted D_{ZI} was compared to the predicted D_{ZI} of the reference population (RG for π, ZI for the ratio analysis). Observed summary statistic were multiplied by the ratio of these predicted values to obtain a corrected estimate. For both statistics, this adjustment led to changes of ~1% or less.

Linkage disequilibrium from empirical and simulated data

In addition to the standard correlation coefficient (r^2) of linkage disequilibrium (LD), we also examined directional LD via the r_ω statistic [70,71]. Here, LD is defined as positive if minor alleles preferentially occur on the same haplotype, and otherwise LD is negative. Empirical LD patterns were compared to data simulated under neutral evolution and equilibrium demography using *ms* [86]. In these simulations, the population mutation rate was taken from observed π. The population recombination rate was then inferred from the ratio of empirical estimates of recombination rates (the average rate from Langley *et al.* [28] for the analyzed X-linked and autosomal regions, simulated separately) and mutation rate [87]. Estimates for the rate of gene conversion relative to crossover events (5x) and the average gene conversion tract length (86.5 bp) were taken from a weighted average of the locus-specific estimates obtained by Yin *et al.* [85].

Genomic scans for loci with unusual allele frequencies

The Λ_{max} statistic of Sweepfinder [75] uses allele frequencies to evaluate the relative likelihood of a selective sweep *versus* neutral evolution. To add information regard diversity reductions, we implemented the approach of Pavlidis *et al.* [76] to include a fraction of the invariant sites. One invariant site was added to the input for every 10 invariant sites that had <50% missing data. Likelihoods were evaluated for 1000 positions from each window. The folded allele frequency spectrum from short intron sites (see below) was used for background allele frequencies, assumed by the method to represent neutral evolution.

Local outliers for Λ_{max} and F_{ST} were examined in overlapping windows of 100 RG non-singleton SNPs (roughly 5 kb on average). For F_{ST}, overlapping windows were offset by increments of 20 RG non-singleton SNPs, in order to identify outlier loci that could result from adaptive population differentiation. Outlier windows were defined by the upper 2.5% (F_{ST}) or 5% (Λ_{max}) quantile for each chromosome arm. The lower threshold for F_{ST} avoids an excessive number of outliers due to the greater number of (overlapping) windows, compared to the non-overlapping windows for Λ_{max}. Outliers with up to two non-overlapping non-outlier windows between them were considered as part of the same "outlier region", since they might reflect a single evolutionary signal. For F_{ST}, the center of an outlier region was defined as the midpoint of its most extreme window. The nearest gene to an outlier region was calculated based on the closest exon (protein-coding or untranslated) to the above location, based on *D. melanogaster* genome release 5.43 coordinates obtained from Flybase.

Two F_{ST} outlier analyses were conducted. One, with the aim of identifying loci that may have contributed to the adaptive difference between African and cosmopolitan populations, focused on F_{ST} between the FR and RG population samples. The other scan was intended to search for potential adaptive differences among African populations. The nine population samples with a mean post-filtering sample size above 3.75 were included (CO, ED, GA, GU, NG, RG, UG, ZI, ZS). The mean F_{ST} from all pairwise population comparisons was evaluated for each window, and outlier regions for this overall F_{ST} were obtained. Each population was also analyzed separately, in terms of the mean F_{ST} from eight pairwise population comparisons. Here, outliers were analyzed separately for each African population, but the lists of population-specific outliers were also combined for more statistically powerful enrichment tests.

The enrichment of gene ontology (GO) categories among sets of outliers was evaluated. For each GO category, the number of unique genes that were the closest to an outlier region center (see above) was noted. A P value was then calculated, representing the probability of observing as many (or more) outlier genes from that category under the null hypothesis of a random distribution of outlier region centers across all windows. Calculating null probabilities based on windows, rather than treating each gene identically, accounts for the fact that genes vary greatly in length, and hence in the number of windows that they are associated with. P values were obtained from a permutation approach in which all outlier region center windows were randomly reassigned 10,000 times (results not shown).

Supporting Information

Figure S1 Evaluation of the tradeoff between genomic coverage and error rate (estimated Phred score) for a series of nominal quality score thresholds. Resequenced genomes from the reference strain (y^1 cn^1 bw^1 sp^1) were modified to simulate realistic levels of variation. Assembly and filtering were conducted as described for the other genomes. Based on the above relationship, we chose a nominal quality score of Q31 (marked in red) to jointly maximize genomic coverage and estimated true quality score.
(PDF)

Figure S2 A: Within-population genetic distances for 27 RG genomes, with each series representing a different sample coverage threshold. Cov2 is the absence of any threshold. Cov26 and Cov27 require that a site have a called allele (at nominal Q31) in at least 26 or all 27 of the RG genomes, respectively. Cov117 and Cov118 require that a site have a called allele in at least 117 or all 118 core genomes from all populations. Sample coverage thresholds were associated with large decreases in variation, as they preferentially excluded variable sites. The most stringent thresholds (e.g. Cov118) lessened the dependence of genetic distances on sequencing depth. B: For the 27 RG genomes, a comparison of within-population genetic distances and distance to the published reference genome. For the unfiltered data (Cov2), within-population and reference divergences are of similar magnitude for genomes with >25X depth (here, outliers for low reference divergence may represent non-African admixture). A consistent "reference bias" (closer relationship to the reference genome than to genomes from the same population) was observed for genomes with <25X depth. For the stringent sample coverage threshold (Cov118), all genomes show strong reference sequence bias. In fact, the reference sequence becomes the closest relative of each African genome. No sample coverage thresholds were used in downstream analyses.
(PDF)

Figure S3 Expectations and observations for genetic distances with regard to population ancestry. (A) An illustration of basic diversity relationships between sub-Saharan and cosmopolitan populations. Cosmopolitan genetic variation is essentially a subset of that observed in sub-Saharan Africa. Due to the diversity loss associated with the out-of-Africa expansion, genetic distances amongst cosmopolitan haplotypes are lower than if cosmopolitan and sub-Saharan haplotypes are compared. The admixture inference method compares sub-Saharan and cosmopolitan genomes, assessing whether each genomic window truly looks like a "S:C" comparison above (in the case of African ancestry for the sub-Saharan genome) or if it instead resembles a "C:C" comparison between cosmopolitan genomes, based on genetic distance to France being lower than expected for a truly African haplotype. (B) Plots of the local ratio of D_{FR} (genetic distance to the France sample) for single Rwanda RG genomes relative to the FR

average (genetic distance among France genomes). Shown are RG2 in light green (for which no admixture was called) and RG21 in red (for which two admixture intervals were called, see yellow boxes), for windows along the complete arm 2R. As shown here, the two RG lines have generally similar genetic distances to the France sample, but within the putative admixture intervals, RG21 becomes more similar to the cosmopolitan genomes.
(PDF)

Figure S4 Performance of the admixture detection HMM method on simulated data. Each chart depicts the estimated admixture probability (Y axis) for each window along the chromosome (X axis), with true admixture tracts shaded. Shown here are representative simulation results for admixture beginning 100 generations ago (A–C), 1,000 generations ago (D–F), or 10,000 generations ago (G–I).
(PDF)

Figure S5 A log-scale plot of admixture probabilities from all genomic windows of four subsets of the sequenced African genomes. For groups of genomes within the Rwanda RG population sample and outside it, and for genomes in the primary core and secondary core categories (the former with greater than 25X sequencing depth), the proportion of window admixture probabilities within each 5% bin is plotted. The greater occurrence of intermediate admixture probabilities for secondary core genomes may indicate less accurate performance, relative to that observed for primary core genomes.
(PDF)

Figure S6 The relationship between inferred population admixture proportion and the human population size of the collection locality. Admixture proportion is the average level of non-African ancestry estimated for a population's genomes by the HMM method described in the text. A maximum population size of 100,000 was based on the assumption that flies in larger cities continue to occupy similarly uniform urban environments. The relationship was statistically significant (Spearman $\rho = 0.60$; one-tailed $P = 0.003$).
(PDF)

Figure S7 Population diversity ratios across the genome. (A) France (FR) vs. Rwanda (RG) illustrates different levels of non-African diversity loss for each major chromosome. (B) FR vs. Zambia (ZI) demonstrates that results from (A) are not driven by the RG-specific patterns. (C) RG vs. ZI shows less heterogeneity, and suggests that the peak observed in (B) is due to a ZI-specific loss of diversity around the chromosome 3 centromere. Chromosome arms are labeled and indicated by color. Each window contains 5000 RG non-singleton SNPs.
(PDF)

Figure S8 A neighbor-joining population distance tree based on the matrix of D_{xy} values. Branch lengths are to scale, and basal node was obtained by midpoint rooting.
(PDF)

Table S1 Population samples from which the sequenced genomes originate. Negative latitudes and longitudes indicate southern and western hemispheres, respectively. For each sample, numbers of primary core and secondary core genomes are given (>25X and <25X mean sequencing depth, respectively). Addendum genomes are listed by major chromosome, and consist of chromosome extraction lines with highly variable depth, except where noted.
(XLS)

Table S2 Characteristics of the sequenced genomes and their corresponding fly stocks. Labels of isofemale, inbred, and chromosome extraction stocks are given, along with NIH SRA access numbers. Focal and non-focal chromosome arms originating from the population of interest are listed. Read length, genomic coverage, and mean sequencing depth are provided.
(XLS)

Table S3 Regions of identity by descent (defined as sequence divergence <0.0005) were identified using 500 kb windows, advanced at 100 kb. All pairs of genomes in the data set were examined (within and between population samples) for target arms on chromosomes X, 2, and 3. All detected tracts of identity are listed here, but only a subset of these were masked from the analyzed data (Table S4).
(XLS)

Table S4 Regions of identity-by-descent masked from the analyzed data. Regions of identity by descent (defined as sequence divergence <0.0005) were identified using 500 kb windows, advanced at 100 kb. All pairs of genomes in the data set were examined (within and between population samples) for target arms on chromosomes X, 2, and 3. Our interest was to identify data that departs from population genetic assumptions due to close relatedness within population samples, and to mask this data in the FASTA files only. IBD regions were only masked if they occurred in within-population comparisons and if they exceeded the scale of IBD observed in between-population comparisons. Some genomic intervals, including centromeres and telomeres, had recurrent IBD in between- population comparisons (list below). Within these manually defined regions, IBD blocks up to 4 Mb occurred in between-population comparisons, and we elected to only filter within-population IBD blocks greater than 5 Mb. Outside of these recurrent IBD zones, we identified within-population pairs of individuals with more than 5 Mb of total genome-wide IBD (this was beyond the scale of total between-population IBD observed outside recurrent IBD zones). All IBD segments for these pairs (including those in recurrent IBD zones) were masked from one of the identical alleles. A buffer region of 100 kb was added to each IBD interval, to account for IBD extending between window increments. Note that position numbers for each arm are given starting with 1 (not 0), and are in closed format (the start and stop positions are the first and last bp included in a tract).
(XLS)

Table S5 Admixture probabilities across genomic windows for primary core genomes. For windows of 1000 RG non-singleton SNPs (coordinates listed), each genome's admixture probability from the HMM forward-backward algorithm is listed. Each chromosome arm is presented in a separate tab. GA187 is not present for 2L and 2R.
(XLS)

Table S6 Admixture probabilities across genomic windows for secondary core and addendum genomes. For windows of 1000 RG non-singleton SNPs (coordinates listed), each genome's admixture probability from the HMM forward-backward algorithm is listed. Each chromosome arm is presented in a separate tab. These probabilities are provided only to illustrate the HMM's performance under challenging conditions of low sequencing depth. Aside from possibly the RG secondary core genomes, these probabilities may be less accurate than those for the primary core genomes (Table S5), and are not intended for ancestry assignment in downstream analyses.
(XLS)

Table S7 Admixture characteristics of African populations. Sample sizes before and after admixture filtering are given. The proportion of non-African admixture estimated for each population is shown, along with the average length of admixture tracts in centiMorgans. Finally, the estimated town population size is given.
(XLS)

Table S8 Estimated cosmopolitan admixture proportion for each of the primary core genomes, based on HMM analysis of whole chromosome arms (center column), or restricted to non-centromeric, non-telomeric intervals (right column).
(XLS)

Table S9 Individual results from Principle Components Analysis. PCA was applied to the full primary core data, and to sub-Saharan genomes only, both before admixture filtering and after it. Columns following an individual ID refer to the vector of PC1, PC2, etc.
(XLS)

Table S10 Nucleotide diversity for populations samples with >95% genomic coverage of n>1 in the filtered data. Values are listed for each chromosome arm, and the average of arm estimates. Estimates are given for non-centromeric, non-telomeric chromosomal regions (left), and for the full data (right).
(XLS)

Table S11 D_{xy} and F_{ST} between pairs of populations, for each major chromosome arm. Admixture-filtered data from genomes with <15% estimated admixture were analyzed for non-centromeric, non-telomeric regions. Results are presented in separate tabs for all sites, and for middles of short introns (see Materials and Methods).
(XLS)

Table S12 Regression results for correlations of short intron diversity. Recombination rate estimates are compared against nucleotide diversity in the RG sample. Chromosomal position is also regressed against RG nucleotide diversity.
(XLS)

Table S13 Outlier regions for Sweepfinder likelihood ratio for a Rwanda population sample. Outliers were identified and delimited as described in the Materials and Methods. "Region" coordinates include all outlier windows within an outlier region. "Window" coordinates refer to a region's window with the highest Λ_{max}. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region.
(XLS)

Table S14 Gene ontology enrichment analysis based on outlier windows for high Λ_{max} in the Rwanda RG sample, indicating potential targets of recent selective sweeps. GO ID number and description are given for each biological process (b), cellular component (c), or molecular activity (m) represented. *P* values were generated by randomly permuting outlier locations. Categories with >1 outlier genes are listed first, sorted by *P* value. Categories with <2 outlier genes follow.
(XLS)

Table S15 Outlier regions for mean pairwise F_{ST} among 9 sub-Saharan population samples. Outliers were identified and delimited as described in the Materials and Methods. "Region" coordinates include all outlier windows within an outlier region. "Window" coordinates refer to a region's window with the highest F_{ST}. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region.
(XLS)

Table S16 Gene ontology enrichment analysis based on outlier windows for high mean pairwise F_{ST} among 9 sub-Saharan population samples. GO ID number and description are given for each biological process (b), cellular component (c), or molecular activity (m) represented. *P* values were generated by randomly permuting outlier locations. Categories with >1 outlier genes are listed first, sorted by *P* value. Categories with <2 outlier genes follow.
(XLS)

Table S17 Outlier regions for F_{ST} between France and Rwanda population samples. Outliers were identified and delimited as described in the Materials and Methods. "Region" coordinates include all outlier windows within an outlier region. "Window" coordinates refer to a region's window with the highest F_{ST}. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region.
(XLS)

Table S18 Gene ontology enrichment analysis based on outlier windows for high F_{ST} between FR and RG population samples. GO ID number and description are given for each biological process (b), cellular component (c), or molecular activity (m) represented. *P* values were generated by randomly permuting outlier locations. Categories with >1 outlier genes are listed first, sorted by *P* value. Categories with <2 outlier genes follow.
(XLS)

Table S19 Genomic locations of selected admixture peaks and valleys are listed in separate tables. For each of these regions, information is given regarding any outlier regions for France-Rwanda F_{ST}. A significant excess of overlap between admixture peaks and F_{ST} outlier regions was observed.
(XLS)

Author Contributions

Conceived and designed the experiments: JEP RBC-D RPS KAS DJB CHL. Performed the experiments: CMC MWC PS. Analyzed the data: JEP RBC-D RPS KAS PD JJE CHL. Wrote the paper: JEP CHL.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.
2. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330: 1787–1797.
3. Dobzhansky T, Sturtevant AH (1938) Inversions in the Chromosomes of *Drosophila pseudoobscura*. Genetics 23: 28–64.
4. Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54: 595–609.
5. Singh RS, Hickey DA, David J (1982) Genetic differentiation between geographically distant populations of *Drosophila melanogaster*. Genetics 101: 235–256.
6. Mettler LE, Voelker RA, Mukai T (1977) Inversion clines in populations of *Drosophila melanogaster*. Genetics 87: 169–176
7. Hudson RR, Kreitman M, Aguadé (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153–159.
8. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351: 652–654.
9. Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356: 519–520.

10. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
11. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289–1303.
12. Kauer M O., Dieringer D, Schlötterer C (2003) A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of *Drosophila melanogaster*.
13. Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W (2007) Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177: 469–480.
14. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415: 1022–1024.
15. Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149–1152.
16. Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L, et al. (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. In: Hecht MK, Wallace B, Prance GT, eds. Evolutionary biology. New York: Plenum. pp. 159–225.
17. Begun DJ, Aquadro CF (1993) African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365: 548–550.
18. Pool JE, Aquadro CF (2006) History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics 174: 915–929.
19. Lachaise D, Silvain J-F (2004) How two Afrotropical endemics made two cosmopolitan commensals: the *Drosophila melanogaster – D. simulans* paleogeographic riddle. Genetica 120: 17–39.
20. Veuille M, Baudry E, Cobb M, Derome N, Gravot E (2004) Historicity and the population genetics of *Drosophila melanogaster* and *D. simulans*. Genetica 120: 61–70.
21. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila* melanogaster populations. Genome Res 15: 790–799.
22. Baudry E, Viginier B, Veuille M (2004) Non-African populations of *Drosophila melanogaster* have a unique origin. Mol Biol Evol 21: 1482–1491.
23. Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genet 2: e166. doi:10.1371/journal.pgen.0020166
24. Thornton KR, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172: 1607–1619.
25. Lintner JA (1882) First annual report on the injurious and other insects of the State of New York. Albany, New York: Weed, Parsons, and Co.
26. Caracristi G, Schlötterer C (2003) Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol Biol Evol 20: 792–799.
27. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, et al. (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5: e310. doi:10.1371/journal.pbio.0050310
28. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, et al. (2012) Genomic variation in natural populations of *Drosophila melanogaster*. Genetics, In Press.
29. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The *Drosophila melanogaster* genetic reference panel. Nature 482: 173–178.
30. Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. Genetics 188: 239–246.
31. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 2078–2079.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
33. Capy P, Veuille M, Paillette M, Jallon J-M, Vouidibio J, et al. (2000) Sexual isolation of genetically differentiated sympatric populations of *Drosophila melanogaster* in Brazzaville, Congo: the first step towards speciation? Heredity 84: 468–475.
34. Kauer M, Dieringer D, Schlötterer C (2003) Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. Mol Biol Evol 20:1329–1337.
35. Vouidibio J, Capy P, Defaye D, Pla E, Sandrin E, et al. (1989) Short-range genetic structure of *Drosophila melanogaster* populations in an Afrotropical urban area and its significance. Proc Natl Acad Sci USA 86: 8442–8446.
36. Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719.
37. Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, et al. (2004) A world-wide survey of an *Accord* insertion and its association with DDT resistance in *Drosophila melanogaster*. Mol Ecol 13: 2491–2504.
38. Wu C-I, Hollocher H, Begun DJ, Aquadro CF, Xu Y, et al. (1995) Sexual isolation in *Drosophila melanogaster*: a possible case of incipient speciation. Proc Natl Acad Sci USA 92: 2519–2523.
39. Hollocher H, Ting C-T, Pollack F, Wu C-I (1997) Incipient speciation by sexual isolation in *Drosophila melanogaster*: variation in mating preference and correlation between the sexes. Evolution 51: 1175–1181.
40. Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. Am Nat 130: 113–146.
41. Aulard S, David JR, Lemeunier F (2002) Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. Genet Res 79: 49–63.
42. Lucchesi JC, Suzuki DT (1968) The interchromosomal control of recombination. Ann Rev Genet 2: 53–86.
43. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190
44. Sackton TB, Kulathinal RJ, Bergman CM, Quinlan AR, Dopman EB, et al. (2009) Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. Genome Biol Evol 1: 449–465.
45. van Herrewege J, David JR (1997) Starvation and desiccation tolerances in *Drosophila*: comparison of species from different climatic origins. Ecoscience 4: 151–157.
46. Stanley SM, Parsons PA, Spence GE, Weber L (1980) Resistance of species of the *Drosophila melanogaster* subgroup to environmental extremes. Aust J Zool 28: 413–421.
47. Richardson MF, Weinert LM, Welch JJ, Linheiro RS, Magwire MM, et al. (2012) Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. (Joint Submission)
48. Corbett-Detig RB, Hartl DL (2012) Population genomics of inversion polymorphisms in *Drosophila melanogaster*. (Joint Submission)
49. Guerrero RF, Rousset F, Kirkpatrick M (2012) Coalescent patterns for chromosomal inversions in divergent populations. Phil Trans R Soc B 367:430–438.
50. Hasson E, Eanes WF (1996) Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics 144:1565–1575.
51. Andolfatto P, Przeworski M (2001) Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. Genetics 158: 657–665.
52. Greenberg R, Crow JF (1960) A comparison of the effect of lethal and detrimental chromosomes from *Drosophila* populations. Genetics 8:1153–1168.
53. Dobzhansky T (1954) Evolution as a creative process. Proceedings of the 9th International Congress on Genetics, Bellagio, Italy, 1:435–449.
54. Ohta T (1971) Associative overdominance caused by linked detrimental mutations. Genet Res 18:277–286.
55. Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. Genetics 173:419–434.
56. Bierne N, Tsitrone A, David P (2000) An inbreeding model of associative overdominance during a population bottleneck. Genetics 155:1981–1990.
57. Pool JE, Nielsen R (2008) The impact of founder events on chromosomal variability in multiply mating species. Mol Biol Evol 25: 1728–1736.
58. Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583–589.
59. Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. Genome Res 16: 875–884.
60. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the utility of short intron sequences as a reference for the detection of positive and negative selectioni in *Drosophila*. (2010) Mol Biol Evol 27:1226–1234,
61. Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the *Drosophila* genome? PLoS Genet 5: e1000495. doi:10.1371/journal.pgen.1000495
62. Jensen JD, Thornton KR, Andolfatto P (2008) Approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. PLoS Genet 4: e1000198. doi:10.1371/journal.pgen.1000198
63. Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. Evolution 61: 3001–3006.
64. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783–796.
65. Pennings PS, Hermisson J (2006) Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet 2: e186. doi:10.1371/journal.pgen.0020186
66. Charlesworth D, Charlesworth B, Morgan MT (1995) The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632.
67. Zeng K, Charlesworth B (2011) The joint effects of background selection and genetic recombination on local gene genealogies. Genetics 189:251–266.
68. Ometto L, Glinka S, De Lorenzo D, Stephan W (2005) Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol 22: 2119–2130.
69. Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, et al. (2007) Adaptive genic evolution in the *Drosophila* genomes. Proc Natl Acad Sci USA 104: 2271–2276.
70. Langley CH, Crow JF (1974) The direction of linkage disequilibrium. Genetics 78: 937–941.
71. Langley CH, Tobari YN, Kojima K-I (1974) Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78: 921–936.
72. Stephan W, Song YS, Langley CH (2006) The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics 172: 2647–2663.
73. Chan AH, Jenkins P, Song YS (2012) Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. (Joint Submission)
74. Jensen JD, Kim Y, Bauer DuMont V, Aquadro CF, Bustamante CD (2005) Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170: 1401–1410.

75. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566–1575.
76. Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. Genetics 185: 907–922.
77. Aguadé M (2009) Nucleotide and copy-number polymorphism at the odorant receptor genes *Or22a* and *Or22b* in *Drosophila melanogaster*. Mol Biol Evol 26: 61–70.
78. Aminetzach YT, Macpherson JM, Petrov DA (2005) Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. Science 309: 764–767.
79. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. (2011) Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a duplication. PLoS Genet 7: e1002337. doi:10.1371/journal.pgen.1002337
80. Pool JE (2009) Notes regarding the collection of African *Drosophila melanogaster*. Dros Inf Serv 92:130–134.
81. Fuyama Y (1984) Gynogenesis in *Drosophila melanogaster*. Jpn J Genet 59: 91–96.
82. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858.
83. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science 320: 1629–1631.
84. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Res 19: 136–142.
85. Yin J, Jordan MI, Song YS (2009) Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics 25: i231–i239.
86. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model. Bioinformatics 18: 337–338.
87. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, et al. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res 19: 1195–1201.

SUPPORTING TABLES:

(Available at: http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1003080#s5)

Table S1. Population samples from which the sequenced genomes originate. Negative latitudes and longitudes indicate southern and western hemispheres, respectively. For each sample, numbers of primary core and secondary core genomes are given (>25X and <25X mean sequencing depth, respectively). Addendum genomes are listed by major chromosome, and consist of chromosome extraction lines with highly variable depth, except where noted. doi:10.1371/journal.pgen.1003080.s009. (XLS).

Table S2. Characteristics of the sequenced genomes and their corresponding fly stocks. Labels of isofemale, inbred, and chromosome extraction stocks are given, along with NIH SRA access numbers. Focal and non-focal chromosome arms originating from the population of interest are listed. Read length, genomic coverage, and mean sequencing depth are provided. doi:10.1371/journal.pgen.1003080.s010. (XLS).

Table S3. Regions of identity by descent (defined as sequence divergence ¡0.0005) were identified using 500 kb windows, advanced at 100 kb. All pairs of genomes in the data set were examined (within and between population samples) for target arms on chromosomes X, 2, and 3. All detected tracts of identity are listed here, but only a subset of these were masked from the analyzed data (Table S4). doi:10.1371/journal.pgen.1003080.s011. (XLS).

Table S4. Regions of identity-by-descent masked from the analyzed data. Regions of identity by descent (defined as sequence divergence ¡0.0005) were identified using 500 kb windows, advanced at 100 kb. All pairs of genomes in the data set were examined (within and between population samples) for target arms on chromosomes X, 2, and 3. Our interest was to identify data that departs from population genetic

assumptions due to close relatedness within population samples, and to mask this data in the FASTA files only. IBD regions were only masked if they occurred in within-population comparisons and if they exceeded the scale of IBD observed in between-population comparisons. Some genomic intervals, including centromeres and telomeres, had recurrent IBD in between- population comparisons (list below). Within these manually defined regions, IBD blocks up to 4 Mb occurred in between-population comparisons, and we elected to only filter within-population IBD blocks greater than 5 Mb. Outside of these recurrent IBD zones, we identified within-population pairs of individuals with more than 5 Mb of total genome-wide IBD (this was beyond the scale of total between-population IBD observed outside recurrent IBD zones). All IBD segments for these pairs (including those in recurrent IBD zones) were masked from one of the identical alleles. A buffer region of 100 kb was added to each IBD interval, to account for IBD extending between window increments. Note that position numbers for each arm are given starting with 1 (not 0), and are in closed format (the start and stop positions are the first and last bp included in a tract). doi:10.1371/journal.pgen.1003080.s012. (XLS).

Table S5. Admixture probabilities across genomic windows for primary core genomes. For windows of 1000 RG non-singleton SNPs (coordinates listed), each genome's admixture probability from the HMM forward-backward algorithm is listed. Each chromosome arm is presented in a separate tab. GA187 is not present for 2L and 2R. doi:10.1371/journal.pgen.1003080.s013. (XLS).

Table S6. Admixture probabilities across genomic windows for secondary core and addendum genomes. For windows of 1000 RG non-singleton SNPs (coordinates listed), each genome's admixture probability from the HMM forward-backward algorithm is listed. Each chromosome arm is presented in a separate tab. These probabilities are provided only to illustrate the HMM's performance under challenging conditions of low sequencing depth. Aside from possibly the RG secondary core

genomes, these probabilities may be less accurate than those for the primary core genomes (Table S5), and are not intended for ancestry assignment in downstream analyses. doi:10.1371/journal.pgen.1003080.s014. (XLS).

Table S7. Admixture characteristics of African populations. Sample sizes before and after admixture filtering are given. The proportion of non-African admixture estimated for each population is shown, along with the average length of admixture tracts in centiMorgans. Finally, the estimated town population size is given. doi:10.1371/journal.pgen.1003080.s015. (XLS).

Table S8. Estimated cosmopolitan admixture proportion for each of the primary core genomes, based on HMM analysis of whole chromosome arms (center column), or restricted to non-centromeric, non-telomeric intervals (right column). doi:10.1371/journal.pgen.1003080.s016. (XLS).

Table S9. Individual results from Principle Components Analysis. PCA was applied to the full primary core data, and to sub-Saharan genomes only, both before admixture filtering and after it. Columns following an individual ID refer to the vector of PC1, PC2, etc. doi:10.1371/journal.pgen.1003080.s017. (XLS).

Table S10. Nucleotide diversity for populations samples with ¿95% genomic coverage of n¿1 in the filtered data. Values are listed for each chromosome arm, and the average of arm estimates. Estimates are given for non-centromeric, non-telomeric chromosomal regions (left), and for the full data (right).
doi:10.1371/journal.pgen.1003080.s018. (XLS).

Table S11. Dxy and FST between pairs of populations, for each major chromosome arm. Admixture-filtered data from genomes with ¡15% estimated admixture were analyzed for non-centromeric, non-telomeric regions. Results are presented in separate tabs for all sites, and for middles of short introns (see Materials and Methods). doi:10.1371/journal.pgen.1003080.s019.(XLS).

Table S12. Regression results for correlations of short intron diversity. Re-

combination rate estimates are compared against nucleotide diversity in the RG sample. Chromosomal position is also regressed against RG nucleotide diversity. doi:10.1371/journal.pgen.1003080.s020. (XLS).

Table S13. Outlier regions for Sweepfinder likelihood ratio for a Rwanda population sample. Outliers were identified and delimited as described in the Materials and Methods. Region coordinates include all outlier windows within an outlier region. Window coordinates refer to a region's window with the highest ?max. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region. doi:10.1371/journal.pgen.1003080.s021. (XLS).

Table S14. Gene ontology enrichment analysis based on outlier windows for high ?max in the Rwanda RG sample, indicating potential targets of recent selective sweeps. GO ID number and description are given for each biological process (b), cellular component (c), or molecular activity (m) represented. P values were generated by randomly permuting outlier locations. Categories with ¿1 outlier genes are listed first, sorted by P value. Categories with ¡2 outlier genes follow. doi:10.1371/journal.pgen.1003080.s022. (XLS).

Table S15. Outlier regions for mean pairwise FST among 9 sub-Saharan population samples. Outliers were identified and delimited as described in the Materials and Methods. Region coordinates include all outlier windows within an outlier region. Window coordinates refer to a region's window with the highest FST. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region. doi:10.1371/journal.pgen.1003080.s023. (XLS).

Table S16. Gene ontology enrichment analysis based on outlier windows for high mean pairwise FST among 9 sub-Saharan population samples. GO ID number and description are given for each biological process (b), cellular component (c), or

molecular activity (m) represented. P values were generated by randomly permuting outlier locations. Categories with ¿1 outlier genes are listed first, sorted by P value. Categories with ¡2 outlier genes follow. doi:10.1371/journal.pgen.1003080.s024. (XLS).

Table S17. Outlier regions for FST between France and Rwanda population samples. Outliers were identified and delimited as described in the Materials and Methods. Region coordinates include all outlier windows within an outlier region. Window coordinates refer to a region's window with the highest FST. The gene with the closest exon to the predicted sweep target is listed, along with information about the putative target position within the gene region. doi:10.1371/journal.pgen.1003080.s025. (XLS).

Table S18. Gene ontology enrichment analysis based on outlier windows for high FST between FR and RG population samples. GO ID number and description are given for each biological process (b), cellular component (c), or molecular activity (m) represented. P values were generated by randomly permuting outlier locations. Categories with >1 outlier genes are listed first, sorted by P value. Categories with <2 outlier genes follow. doi:10.1371/journal.pgen.1003080.s026. (XLS).

Table S19. Genomic locations of selected admixture peaks and valleys are listed in separate tables. For each of these regions, information is given regarding any outlier regions for France-Rwanda FST. A significant excess of overlap between admixture peaks and FST outlier regions was observed. doi:10.1371/journal.pgen.1003080.s027. (XLS).

SUPPORTING FIGURES:

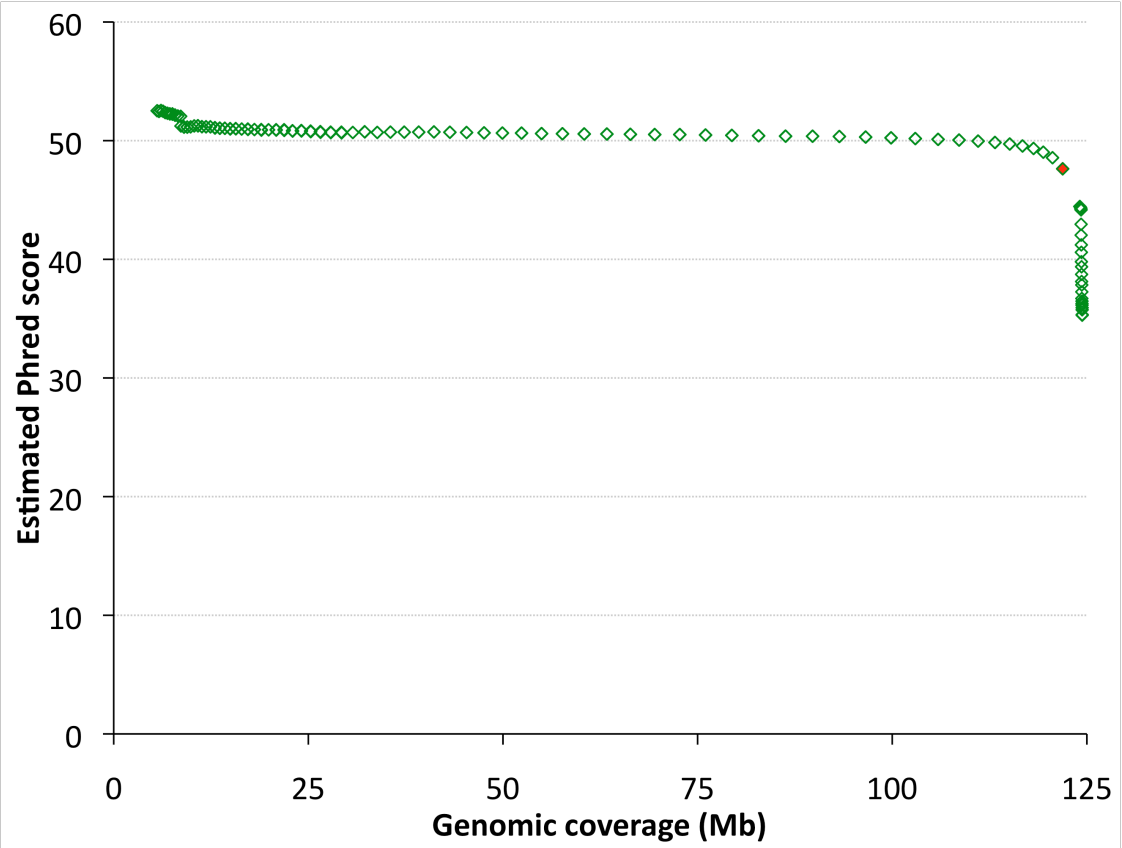

Figure S1. Evaluation of the tradeoff between genomic coverage and error rate (estimated Phred score) for a series of nominal quality score thresholds. Resequenced genomes from the reference strain (y^1 cn^1 bw^1 sp^1) were modified to simulate realistic levels of variation. Assembly and filtering were conducted as described for the other genomes. Based on the above relationship, we chose a nominal quality score of Q31 (marked in red) to jointly maximize genomic coverage and estimated true quality score.

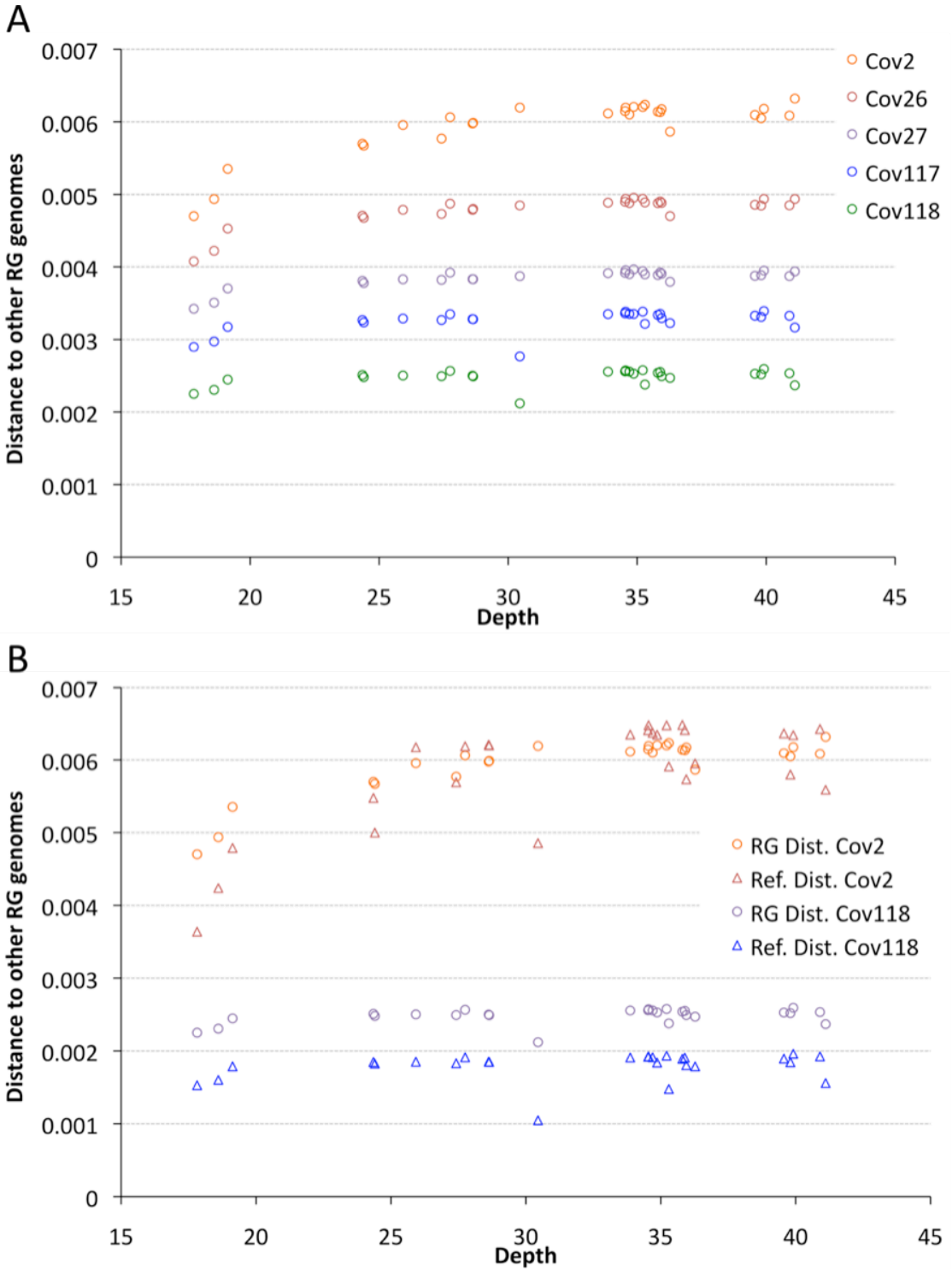
A
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
Distance to other RG genomes
15
20
25
30
35
40
45
Depth
Cov2
Cov26
Cov27
Cov117
Cov118
B
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
Distance to other RG genomes
15
20
25
30
35
40
45
Depth
RG Dist. Cov2
Ref. Dist. Cov2
RG Dist. Cov118
Ref. Dist. Cov118

Figure S2. A: Within-population genetic distances for 27 RG genomes, with each series representing a different sample coverage threshold. Cov2 is the absence of any threshold. Cov26 and Cov27 require that a site have a called allele (at nominal Q31) in at least 26 or all 27 of the RG genomes, respectively. Cov117 and Cov118 require that a site have a called allele in at least 117 or all 118 core genomes from all populations. Sample coverage thresholds were associated with large decreases in variation, as they preferentially excluded variable sites. The most stringent thresholds (e.g. Cov118) lessened the dependence of genetic distances on sequencing depth. B: For the 27 RG genomes, a comparison of within-population genetic distances and distance to the published reference genome. For the unfiltered data (Cov2), within-population and reference divergences are of similar magnitude for genomes with >25X depth (here, outliers for low reference divergence may represent non-African admixture). A consistent "reference bias" (closer relationship to the reference genome than to genomes from the same population) was observed for genomes with <25X depth. For the stringent sample coverage threshold (Cov118), all genomes show strong reference sequence bias. In fact, the reference sequence becomes the closest relative of each African genome. No sample coverage thresholds were used in downstream analyses.

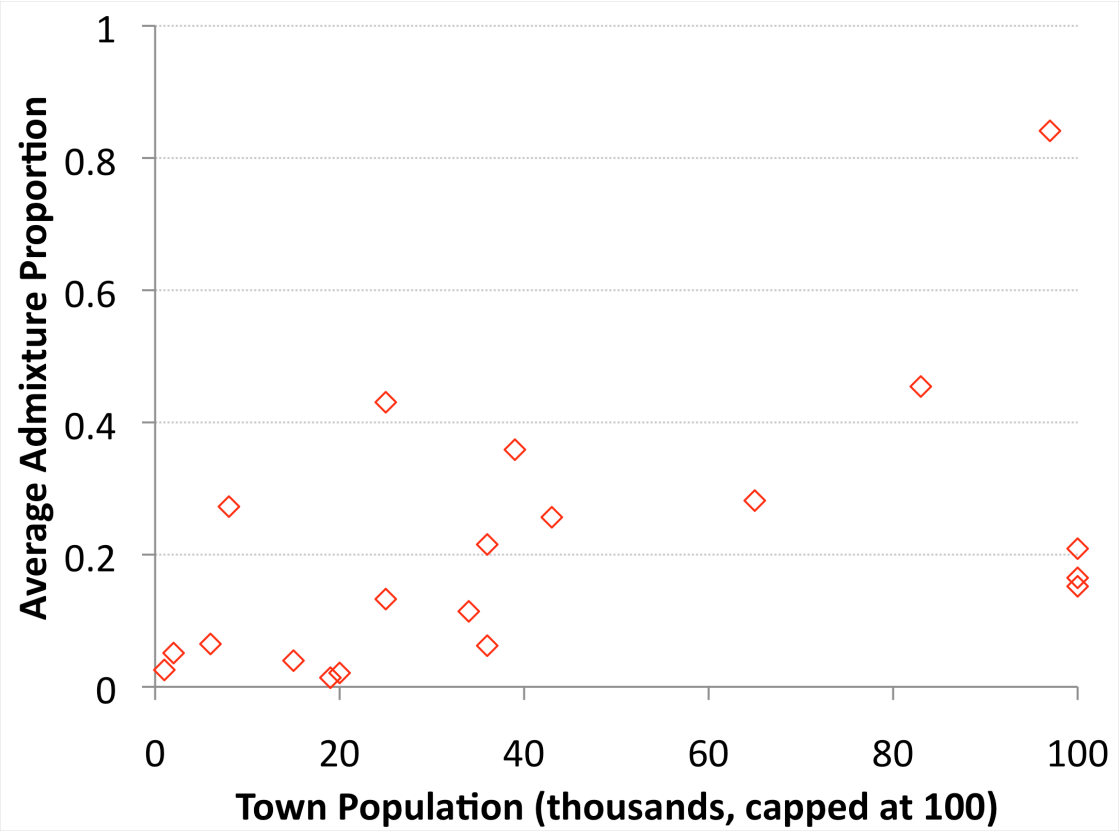

Figure S3. The relationship between inferred population admixture proportion and the human population size of the collection locality. Admixture proportion is the average level of non-African ancestry estimated for a population's genomes by the HMM method described in the text. A maximum population size of 100,000 was based on the assumption that flies in larger cities continue to occupy similarly uniform urban environments. The relationship was statistically significant (Pearson $r = 0.52$; one-tailed $P < 0.01$).

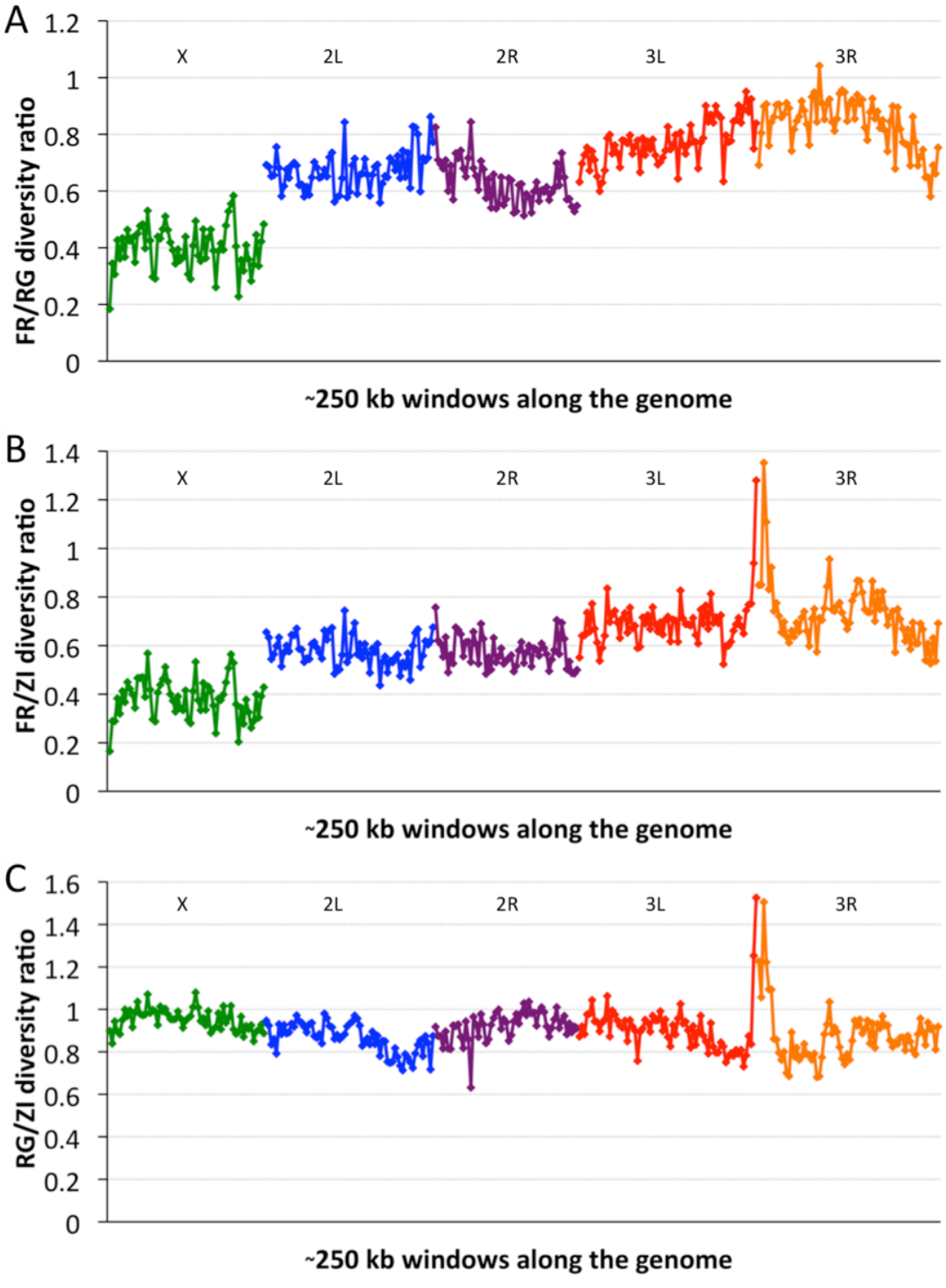
A
1.2
1
0.8
0.6
0.4
0.2
0
X
2L
2R
3L
3R
FR/RG diversity ratio
~250 kb windows along the genome
B
1.4
1.2
1
0.8
0.6
0.4
0.2
0
X
2L
2R
3L
3R
FR/ZI diversity ratio
~250 kb windows along the genome
C
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
X
2L
2R
3L
3R
RG/ZI diversity ratio
~250 kb windows along the genome

Figure S4. Population diversity ratios across the genome. (A) France (FR) vs. Rwanda (RG) illustrates different levels of non-African diversity loss for each major chromosome. (B) FR vs. Zambia (ZI) demonstrates that results from (A) are not driven by the RG-specific patterns. (C) RG vs. ZI shows less heterogeneity, and suggests that the peak observed in (B) is due to a ZI-specific loss of diversity around the chromosome 3 centromere. Chromosome arms are labeled and indicated by color. Each window contains 5000 RG non-singleton SNPs.

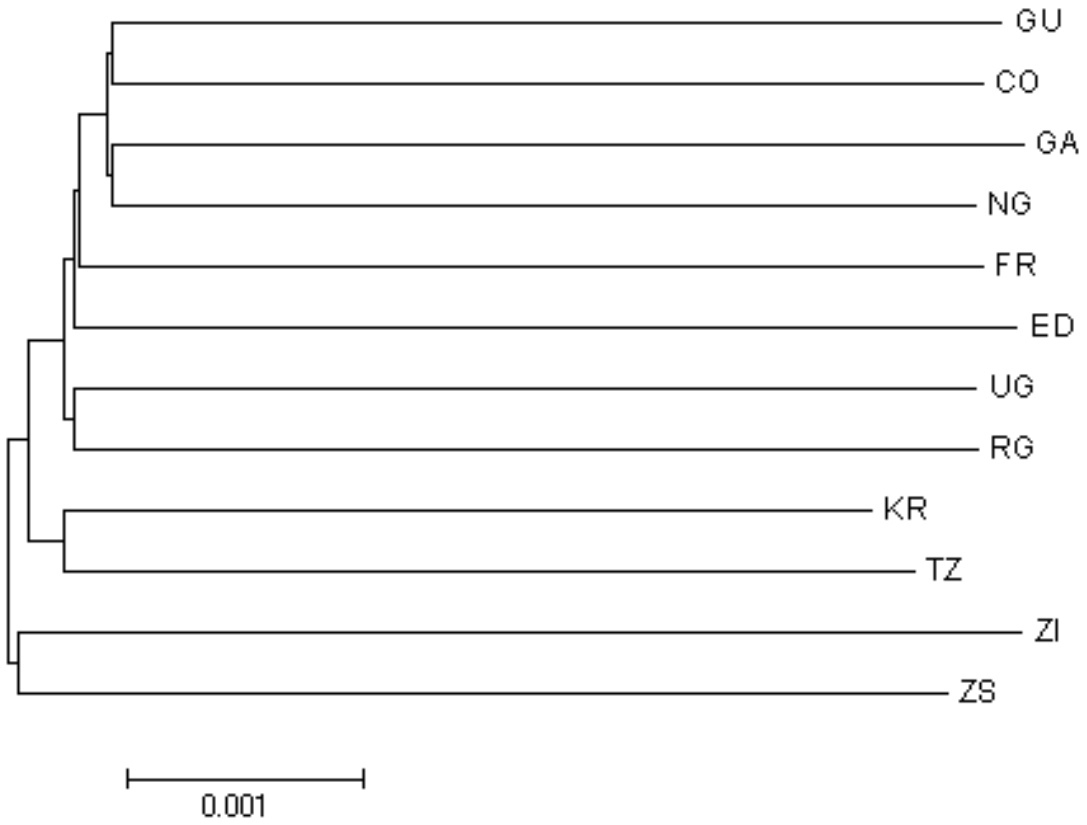

Figure S5. A neighbor-joining population distance tree based on the matrix of Dxy values. Branch lengths are to scale, and basal node was obtained by midpoint rooting.

General Discussion

The population history of *D. melanogaster* has been a focus of research since several decades. Lachaise et al. (1988) have summarized the history of this species in their seminal work "Historical biogeography of the *Drosophila melanogaster* species subgroup". There they compile all ecological data available and propose a historical reconstruction of the distribution pattern of this group. They suggest that *D. melanogaster* originated in western Africa (Tsacas and Lachaise, 1974) when it separated from the ancestor of *D. simulans* around 2.5 million years ago, during the aridification of the Rift (Figure 4, section e.). This divergence time was later confirmed by Li et al. (1999): they estimated it to be around 2.3 million years ago using maximum likelihood methods. Later on, the putative western African origin of *D. melanogaster* was revisited by Pool and Aquadro (2006) and Pool et al. (2012). The former study suggested a potential eastern African origin (Uganda), whereas in Pool et al. (2012) we suggest a southern African origin (Zambia).

Although the origin of this species is clearly sub-Saharan (supported by molecular studies by Begun and Aquadro, 1993; Andolfatto, 2001; Kauer et al., 2002; Ometto et al., 2005) there is still some disagreement on the exact region where this happened, whether it is the west, the east, or the south. Even though our study suggests that the highest diversity can be found in Zambia (Pool et al., 2012) this is not a sufficient argument supporting the placement of the center of origin. We are aware that other demographic factors that can reduce the diversity of founding populations and we accept that the exact origin is still unknown. Ideally, we would use statistical

methods to determine the ancestral range of a taxon given its evolutionary tree (e.g. Ronquist, 1994, 1997; Maddison and Maddison, 2009), but these methods apply to between-species phylogenies. Until similar statistical methods are developed for genealogies this question will remain open.

After the establishment of *D. melanogaster* in sub-Saharan Africa it started spreading throughout the continent and the rest of the world. A few decades ago, when the use of genetic data was still in its beginnings David and Capy (1988) reconstructed the colonization paths taken by the fruit fly and described its current distribution. Their reconstruction was based on allozyme, physiological and ecological data gathered from their own studies and from previous ones (Anxolabéhère et al., 1985; David et al., 1985; Fleuriet, 1986; Hale and Singh, 1987; Boussy and Kidwell, 1987). Here, we aimed at revisiting this history by using state of the art statistical methods and sequencing technologies. Given that the availability of worldwide full genomes is still limited we focused only on North American, European, and African populations.

The main difference between our approach and previous studies is the quantification of population parameters and model testing. Lachaise et al. (1988) and David and Capy (1988) correlated some events with geological data but it wasn't until the last ten years that genetic data was used to estimate divergence times and population sizes in populations of *D. melanogaster* (Baudry et al., 2004; Haddrill et al., 2005; Li and Stephan, 2006; Thornton and Andolfatto, 2006; Stephan and Li, 2007; Laurent et al., 2011; Duchen et al., 2013). Among these studies the present one is the only one that uses data from full genomes. We have also tested several possible demographic scenarios and performed model testing on a Bayesian framework in order to find statistical support for our hypothesis. Such tests in *D. melanogaster* have been done only in Laurent et al. (2011) and the population models for North America, Africa and Europe presented here are completely new. The overall goal of

this research was to study the demography of several *D. melanogaster* populations by making use of full-genome sequences and ABC methods for parameter estimation.

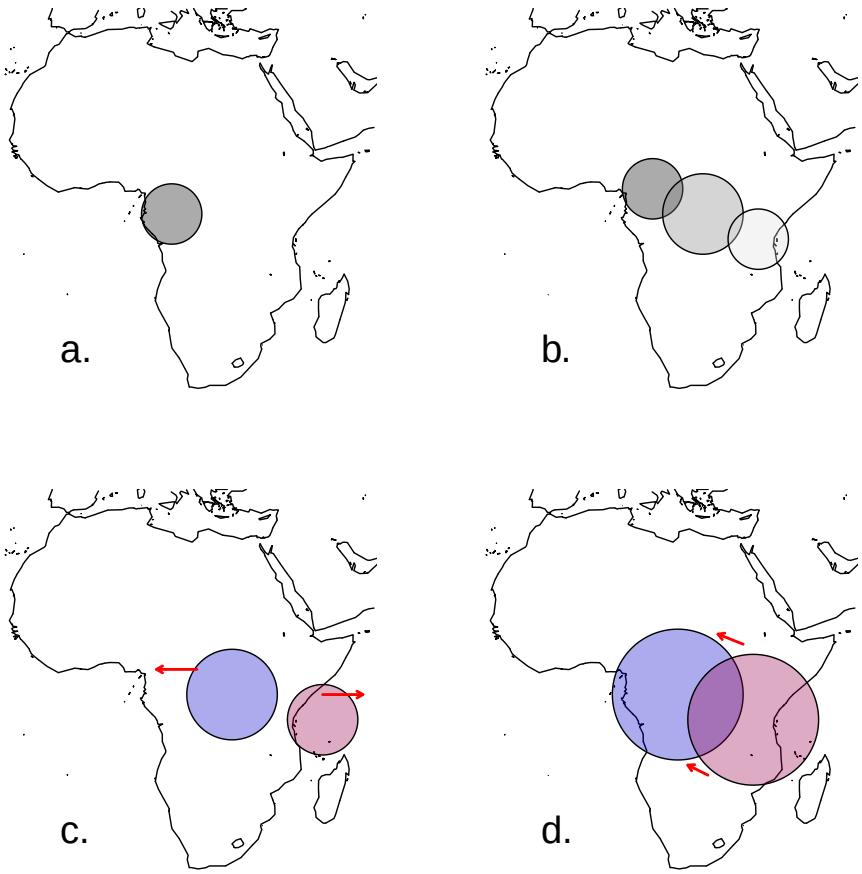

Figure 4: The origin of *D. melanogaster* in Africa according to Lachaise et al. (1988). a) The ancestor of the *D. melanogaster* subgroup arrived from Asia in the middle Miocene. b) Several speciation events took place leading to *D. orena*, *D. erecta*, *D. tessieri*, and *D. yakuba*. c) Split between *D. melanogaster* (west) and the ancestor of *D. simulans* (east) triggered by the continuous aridification of the Rift around 2.5 million years ago. d) Expansion and restored contact between the two species.

Admixture models

In the first chapter we focused on the population history of North American *D. melanogaster*, whose demography has been poorly known. One of the most inter-

esting aspects that draw our interest to this population was its rich diversity, which was unexpected given the young age of this population. North America was first colonized around 200 years ago but has more diversity than European populations, which have diverged from Africa some 19,000 years ago (Li and Stephan, 2006; Thornton and Andolfatto, 2006; Laurent et al., 2011; Duchen et al., 2013). This puzzling aspect led us to analyze different demographic models for North America, in order to find the one that best explains this high diversity.

One of our hypotheses was that the North American population could have derived directly from the African population, with a bigger founding size and a richer starting diversity. However, such a model was not able to explain the actual differentiation between Europe and North America, as well as the differentiation between Africa and North America. Our observed dataset shows that the European and North American populations are the closest; therefore a model in which the North American population splits independently from Africa was not the best. Population differentiation estimates were actually in concordance with the colonization history depicted by entomologists of the 19th century who described *D. melanogaster* as a non-native dipteran insect coming from Europe (Howard, 1900). For this reason we analyzed another model where North America derives directly from Europe, and Europe from Africa (as in Laurent et al., 2011). This model was able to explain F_{st} values between all pairs of populations but did not explain the diversity observed in North America. We then used these same models but allowed for migration to happen between all populations. Migration models were able to explain diversity patterns in all three populations, but again, they were not able to explain F_{st} values. We think that migration is indeed playing an important role, but it also has a homogenizing effect that wipes out signals of differentiation. If there were significant differences in the migration rates between continents, ranging from almost non-existent up to extremely high rates it might be possible to come up with a sce-

nario that explains the current patterns of diversity and differentiation. However, based on the results of chapter 2 we are certain that such differences do not exist, since Nm is always higher than 1, ranging from 5 to 30.

An admixture model was the model that best fitted the observed data. This model was able to explain the diversity observed in North America, as well as the patterns of differentiation between populations. In this model we decided to simulate an admixture event only recently, resembling the recent colonization that happened a few hundred years ago. Given the evidence of colonization in North America from both Europe and Africa (Caracristi and Schlötterer, 2003) we modeled this exact same event in a coalescent framework. We acknowledge migration kept taking place after admixture but most of the current diversity can be already explained by the admixture model, and we did not want to overparametrize our models. Admixture could have happened anywhere, but evidence suggests that it happened in North America (David and Capy, 1988), when populations coming from the North (with European ancestry) met populations coming from the south (with African ancestry). We estimated that around 85% of the ancestry is European and the remaining 15% is African, and this estimation was confirmed by visually inspecting the alignments. Aside from diversity and differentiation values the admixture model was able to explain all other summary statistics as well, including the JSFS.

Migration models

As stated above, although admixture between African and European populations played a major role in generating the diversity of North American populations we are aware that there is constant and ongoing gene flow between these populations. In Chapter 2 we estimated migration rates between African and European populations of *D. melanogaster*. Previous results (Singh and Rhomberg, 1987) showed that the product Nm of population size and migration rate between African and European

populations was in the order of 2. Our results show that Nm is around 10 representing a significant increase of gene flow in the last 25 years. Since *D. melanogaster* is a human commensal we think that this increase in gene flow is correlated with an increase in agricultural trade in the last few decades. Additionally, we found that migration rates between Africa and Europe are not symmetrical, with Africa receiving more migrants per generation from Europe than the other way around, although the difference does not appear to be significant.

Regarding other population parameters we find differences with previous studies (Baudry et al., 2004; Haddrill et al., 2005; Li and Stephan, 2006; Thornton and Andolfatto, 2006; Stephan and Li, 2007; Laurent et al., 2011; Duchen et al., 2013). For instance, estimates of population size in Rwanda are different from that of Zimbabwe, and the same applies to the population of France compared to the Netherlands. Although the confidence intervals of these estimates overlap we do not expect different populations have similar population sizes even if they are close to each other, since they could still have different histories. Divergence time between Rwanda and France does not seem to be significantly different to the one reported between Zimbabwe and the Netherlands. We think this might be the case if the founding population of Europe had representatives of both Rwanda and Zimbabwe in similar proportions. Finally, by looking at Tajima's D and the SFS of Rwanda using neutral loci we find footprints of a bottlenecked and a expanding population. This tells us either that Rwanda (or Zimbabwe) is not at the center of origin of *D. melanogaster*, or that selection is affecting the loci that we are studying. We think both of these cases are taking place simultaneously.

Population genomics

Availability of French and Rwandan sequences was possible thanks to the DPGP2 sequencing effort, which is described in Chapter 3. Together with France and Rwanda

20 additional African populations were sequenced, making up a total of 139 full genomes. Sequencing was performed using Illumina technology and all reads were mapped to *D. melanogaster*'s reference genome. For mapping and assembly we used the programs *bwa* (Li and Durbin, 2009) and *samtools* (Li et al., 2009), both of which are designed for mapping and assembly. One of the key aspects of our assembly was quality control. Before assembling all 139 genomes we first assembled test genomes with known artificially generated mutations (generated with the program *maq* (`http://maq.sourceforge.net/maq-man.shtml`)) to see how *bwa* and *samtools* dealt with sequencing errors. From this we established a threshold for minimum phred score, which we found to be 31. This threshold minimizes erroneous base calls and maximizes depth and coverage. After setting up the best protocol we then assembled all genomes. A total of 130 African lines and 9 French lines were sequenced and assembled. All reported genomes were controlled for quality, and SNPs were called making sure they have good quality thresholds and sequencing depth. Average depth for all reported genomes was 25x. All this data was then used to analyze diversity patterns among populations, as well as detection of identity by descent, and detection of admixture tracts for each chromosome arm in each line.

After analyzing the basic properties of these populations we found that the most diverse population is Siavonga (Zambia). This population is now thought to be much closer to the center of origin of *D. melanogaster* and is now subject of further investigation and acquisition of around 300 full-genome sequences (to be available shortly). Additionally, we found high non-cosmopolitan admixture in African lines. By non-cosmopolitan admixture we refer to non-sub-Saharan lines admixing with sub-Saharan lines. In this research a new method is presented to uncover the regions where admixture took place. Another important aspect of this work is the use of haploid embryos as sequencing targets. We are aware that population genetics studies often require single chromosomes, but this becomes problematic when

the organism under study is not haploid. For single genes the cloning technique is frequently used to separate the two alleles, but for full genomes cloning is no longer a way out. Langley et al. (2011) developed a method to generate haploid *D. melanogaster* embryos. These embryos are then used as targets of sequencing and the resulting sequence is almost completely haploid, with very little residual heterozygosity.

Bayesian estimation

A final comment concerning ABC methods. Being able to dodge the calculation of likelihoods and approximate posterior distributions of parameters is a great step in statistical methods for population genetics. We were able to use ABC not only for parameter estimation but also for model choice, since the ratio of posterior probabilities is proportional to the ratio of likelihoods and this allows us to calculate Bayes factors. One way to calculate posterior probabilities of models simulated by ABC is explained in Fagundes et al. (2007). There exist some concerns and criticisms for the use of ABC for model choice and parameter estimation, but most of these criticisms arise from a frequentist point of view and apply to Bayesian methods in general, not only to ABC. However, in order to validate the use of ABC methods it is important to carefully choose the prior distributions and to test different sets of prior information. Also, a good way to validate the performance of a model with ABC estimates is to use these estimates to run predictive simulations, the way it is done in chapter 1. If the results of the predictive simulations match the characteristics of the observed dataset this is considered a sign of a good ABC calculation. All in all, we are aware that models are just a simplistic representation of a more complicated natural scenario, but these models will help us learn the way populations evolve under simplified assumptions and simplified histories, and that is how models become very useful in evolutionary biology.

The choice of summary statistics also plays an important role in the outcome of an ABC analysis. Traditional ABC approaches make *ad hoc* choices of summary statistics. However, there are other ways to improve the estimation. First, it is possible to carefully choose only the summary statistics that improve the estimation of parameters and leave aside those that do not. Joyce and Marjoram (2008) and Fearnhead and Prangle (2012) developed algorithms for this purpose. Alternatively, one can use a machine-learning approach to estimate a posterior density, as proposed by Blum and François (2010). They fit a non-linear regression of parameter-summary statistics pairs and then enhance the estimation by importance sampling. Another way to tackle this problem is by transforming all statistics using partial least squares (Wegmann et al., 2009). By doing this it is possible to extract only the first few components of this transformation and use them for parameter estimation. This way noise is significantly reduced and it is possible to extract most of the information present in the original set of summary statistics. In the present study we opted for this last method, provided it was fast and the results were trustable.

Bibliography

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Molecular Biology and Evolution 18:279–290.

Anxolabéhère, D., D. Nouaud, G. Périquet, and P. Tchen. 1985. P-element distribution in Eurasian populations of *Drosophila melanogaster*: a genetic and molecular analysis. Proceedings of the National Academy of Sciences 82:5418–5422.

Baudry, E., B. Viginier, and M. Veuille. 2004. Non-African populations of *Drosophila melanogaster* have a unique origin. Molecular Biology and Evolution 21:1482–1491.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162:2025–2035.

Begun, D. J. and C. F. Aquadro. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548–550.

Blum, M. G. and O. François. 2010. Non-linear regression models for approximate bayesian computation. Statistics and Computing 20:63–73.

Boussy, I. A. and M. G. Kidwell. 1987. The PM hybrid dysgenesis cline in eastern Australian *Drosophila melanogaster*: discrete P, Q and M regions are nearly contiguous. Genetics 115:737–745.

Caracristi, G. and C. Schlötterer. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Molecular Biology and Evolution 20:792–799.

Charlesworth, B., M. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289–1303.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genetics 2:e64.

Clarke, B. 1964. Frequency-dependent selection for the dominance of rare polymorphic genes. Evolution 18:364–369.

Clarke, B. and P. O'donald. 1964. Frequency-dependent selection. Heredity 19:201–206.

David, J., P. Capy, V. Payant, and S. Tsakas. 1985. Thoracic trident pigmentation in *Drosophila melanogaster*: differentiation of geographical populations. Genet. Sel. Evol 17:211–224.

David, J. R. and P. Capy. 1988. Genetic variation of *Drosophila melanogaster* natural populations. TRENDS in Genetics 4:106–111.

Depaulis, F., M. Veuille, et al. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution 15:1788–1790.

Duchen, P., D. Živković, S. Hutter, W. Stephan, and S. Laurent. 2013. Demographic Inference Reveals African and European Admixture in the North American *Drosophila melanogaster* Population. Genetics 193:291–301.

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier. 2007. Statistical evaluation of alternative mod-

els of human evolution. Proceedings of the Natural Academy of Sciences U.S.A. 104:17614–17619.

Fearnhead, P. and D. Prangle. 2012. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74:419–474.

Fleuriet, A. 1986. Perpetuation of the hereditary sigma virus in populations of its host, *Drosophila melanogaster*. Geographical analysis of correlated polymorphisms. Genetica 70:167–177.

Fu, Y.-X. and W.-H. Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Grant, P. R. and B. R. Grant. 2006. Evolution of character displacement in Darwin's finches. Science 313:224–226.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Research 15:790–799.

Haldane, J. B. S. 1930. A mathematical theory of natural and artificial selection.(part vi, isolation.). Pages 220–230 *in* Mathematical Proceedings of the Cambridge Philosophical Society vol. 26 Cambridge Univ Press.

Hale, L. R. and R. S. Singh. 1987. Mitochondrial DNA variation and genetic structure in populations of *Drosophila melanogaster*. Molecular Biology and Evolution 4:622–637.

Hartl, D. L. and A. G. Clark. 2007. Principles of Population Genetics. 4th ed. Sinauer Associates, Sunderland, Massachusetts.

Hernandez, R. D., S. Williamson, and C. D. Bustamante. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Molecular Biology and Evolution 24:1792–1800.

Howard, L. O. 1900. A Contribution to the Study of the Insect Fauna of Human Excrement: (With Especial Reference to the Spread of Typhoid Fever by Flies.). Proceedings of the Washington Academy of Sciences 2:541–604.

Johnson, C. W. 1913. The distribution of some species of *Drosophila*. Psyche 20:202–205.

Joyce, P. and P. Marjoram. 2008. Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology 7.

Kauer, M., B. Zangerl, D. Dieringer, and C. Schlötterer. 2002. Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics 160:247–256.

Keller, A. 2007. *Drosophila melanogaster*'s history as a human commensal. Current Biology 17:R77–R81.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics 146:1197–1206.

Kennington, W. J., J. Gockel, and L. Partridge. 2003. Testing for asymmetrical gene flow in a drosophila melanogaster body-size cline. Genetics 165:667–673.

Kettlewell, H. D. 1958. A survey of the frequencies of *Biston betularia* (L.)(Lep.) and its melanic forms in Great Britain. Heredity 12:51–72.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas, and M. Ashburner. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. Evolutionary Biology 22:159–225.

Lachaise, D. and J.-F. Silvain. 2004. How two afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. Genetica 120:17–39.

Langley, C. H., M. Crepeau, C. Cardeno, R. Corbett-Detig, and K. Stevens. 2011. Circumventing heterozygosity: sequencing the amplified genome of a single haploid drosophila melanogaster embryo. Genetics 188:239–246.

Laurent, S. J., A. Werzner, L. Excoffier, and W. Stephan. 2011. Approximate bayesian analysis of drosophila melanogaster polymorphism data reveals a recent colonization of southeast asia. Molecular biology and evolution 28:2041–2051.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics 25:1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. 2009. The sequence alignment/map format and samtools. Bioinformatics 25:2078–2079.

Li, H. and W. Stephan. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. PLoS Genetics 2:1580–1589.

Li, Y.-J., Y. Satta, and N. Takahata. 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. Genes and Genetic Systems 74:117–127.

Maddison, W. P. and D. R. Maddison. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.6.

Maynard-Smith, J. and J. Haigh. 1974. The hitch-hiking effect of a favourable gene. Genetics Research 23:23–35.

Ometto, L., S. Glinka, D. De-Lorenzo, and W. Stephan. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Molecular Biology and Evolution 22:2119–2130.

Pool, J. E. and C. F. Aquadro. 2006. History and structure of Sub-Saharan populations of *Drosophila melanogaster*. Genetics 174:915–929.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. Emerson, P. Saelao, D. J. Begun, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genetics 8:e1003080.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosome: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16:1791–1798.

Ronquist, F. 1994. Ancestral areas and parsimony. Systematic Biology 43:267–274.

Ronquist, F. 1997. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. Systematic Biology 46:195–203.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science 312:1614–1620.

Singh, R. S. and L. R. Rhomberg. 1987. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. Genetics 115:313–322.

Stephan, W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. Philosophical Transactions of the Royal Society B: Biological Sciences 365:1245–1253.

Stephan, W. and H. Li. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Stephan, W., T. H. E. Wiehe, and M. W. Lenz. 1992. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology 41:237–254.

Sturtevant, A. H. 1920. Genetic studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. Genetics 5:488–500.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly. 1997. Inferring coalescence times from DNA sequence data. Genetics 145:505–518.

Thornton, K. R. and P. Andolfatto. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619.

Tsacas, L. and D. Lachaise. 1974. Quatre nouvelles espèces de la Côte-d'Ivoire du genre *Drosophila*, groupe *melanogaster*, et discussion de l'origine du sous-groupe *melanogaster* (Diptera: Drosophilidae). Geographie. Universite d'Abidjan E 7:193–211.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256–276.

Wegmann, D., C. Leuenberger, and L. Excoffier. 2009. Efficient Approximate Bayesian Computation compled with Markov chain Monte Carlo without likelihood. Genetics 182:1207–1218.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97–159.

Zayed, A. and C. W. Whitfield. 2008. A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. Proceedings of the Natural Academy of Sciences 105:3421–3426.

Acknowledgements

First of all this work would not have been possible without the support of my supervisor Prof. Wolfgang Stephan. I am really thankful for this opportunity he has given to me, for all the valuable help and scientific advice he provided during these three years, and for making me a better scientist. I feel very lucky I managed to get into his group, which is right at the forefront in population genetics.

I am also in great debt with Stefan Laurent. He did not only provide with many of the main ideas that served as backbone of this thesis, but he also was of incredible help during the most critical phases of my stay in this group. His constant enthusiasm and devotion with the topics here covered was a key factor in the culmination of these projects.

I would like to continue my acknowledgements to the people that helped me and inspired me with their own work: Dirk Metzler, John Parsch, Stephan Hutter, Pavlos Pavlidis and Daniel Živković. They were always ready to collaborate even with their busy schedules. To the people in Davis, California, where I spent several months working with Charles Langley, John Pool, and Kristian Stevens. My greatest acknowledgements to all of them and their group.

I would like to thank Meike for her constant support at all levels, and for being such a great example to follow, a real inspiration to me and to all of us. Ricardo, for his constant help and for showing such great enthusiasm towards population genetics. Ana and Sebastian, for their support and their hard–working attitude at all times. To my fellow colleagues: Susanne, Francesco, Katharina, Vedran, Korbi,

Myriam, Ina, Amanda, Paul, Felix, Soumya, Andreas, Dimitrios, Mathilde, Noemi, Tetyana, Sonja, and Lisha. To Frau Kroiss, for her constant support with all the administrative stuff and for being always ready to help.

Last but not least I want to express my gratitude to the newer EES and MEME students. I could really write down all your names here, from the cohorts that I had the pleasure to meet. Instead I want to let you know, that you are very special to me, and to thank you deeply for all the good times that we had. You guys made me like what I do and I really appreciate it.

A manera de culminar este trabajo quiero agradecer a Deisy, Carlos y Rocío. A ustedes va dedicado este trabajo y son ustedes la fuente mas profunda de inspiración para seguir adelante. Si pudiera escoger tres personas como las que yo quisiera llegar a ser en un futuro, esas personas serian ustedes, sin lugar a duda. Gracias por todo.

Curriculum Vitae

Curriculum Vitae - Pablo Duchén Bocángel

1 Personal Information

Name	Pablo Duchén Bocángel
Nationality	Bolivian - Mexican
Affiliation	Department of Biology II - Section of Evolutionary Biology Ludwig Maximilians University of Munich Germany
Language skills	Spanish - mother tongue; English - fluent; German - fluent; French - basic

2 Education and Research experience

2010-2013	Ph.D. in Theoretical Population Genetics. University of Munich (LMU), Germany.
2011	Training in Next-generation Sequencing (NGS) bioinformatics. University of California-Davis, United States of America.
2007-2009	Master of Sciences (M.Sc.) in Evolution, Ecology and Systematics. Research in Molecular Phylogenetics and Population Genetics. University of Munich (LMU), Germany
2004-2006	Licenciatura in Biological Sciences. Research in Systematics. Universidad Mayor de San Andres, La Paz - Bolivia.
2001-2004	Bachelor of Sciences (B.Sc.) in Biology, Universidad Mayor de San Andres, La Paz - Bolivia.

3 Conference presentations and invited talks

March 2013	Duchen, P. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Forschergruppe Meeting (Munich, Germany).
June 2012	Duchen, P., Živković, D., Hutter, S., Stephan, W., Laurent, S. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Poster at the Society for Molecular Biology and Evolution (SMBE) meeting (Dublin, Ireland).

April 2012	Duchen, P. Demography of a North American *Drosophila melanogaster* population: an ABC approach. Forschergruppe Meeting (Munich, Germany).
March 2012	Duchen, P., Laurent, S., Stephan, W. Approximate Bayesian Computation and Demographic inference of North American *Drosophila melanogaster*. Invited speaker at Excoffier's Lab. University of Bern, Switzerland.
June 2011	Duchen, P., Laurent, S., Stephan, W. Demographic inference of North American *Drosophila melanogaster*. Poster at the Evolution meeting (University of Oklahoma, United States of America).
October 2009	Duchen, P. and Renner, S. The evolution of *Cayaponia* (Cucurbitaceae): Repeated shifts from bat to bee pollination and long-distance dispersal to Africa 2-6 million years ago. Talk at the III EES Conference (University of Munich, Germany).
October 2008	Duchen, P., Grath, S. Parsch, J. Population genetics of two sex-biased genes in Drosophila ananassae. Poster at the II EES Conference (University of Munich, Germany).
September 2006	Duchen, P. and Beck, S. Fabaceae systematics of the National Park of Cotapata (La Paz - Bolivia). Talk at the V Student Scientific Conference (Universidad Mayor de San Andrés, La Paz, Bolivia).

4 Grants and awards

2009	Award to the best Master thesis: The evolution of *Cayaponia* (Cucurbitaceae): Repeated shifts from bat to bee pollination and long-distance dispersal to Africa 2-6 million years ago. University of Munich, Germany.
2009	Research grant from the Evolution, Ecology and Systematics Master program. University of Munich, Germany.
2005	Research grant from the "Instituto de Ecología" for the Licenciatura thesis. Universidad Mayor de San Andrés, La Paz, Bolivia.

5 Teaching experience

2013	Teaching Assistant in Evolutionary Biology. University of Munich, Germany.
2012	Guest lecture on Bayesian Inference, Approximate Bayesian Computation and its applications to Coalescent Theory. University of Munich, Germany.
2005-2006	Mathematics Distance Tutor for High Schools in the United States.
2006	Guest lecture on Systematics. Universidad Tomás de Aquino, La Paz, Bolivia.
2003	Teaching Assistant in Systematic Botany. Universidad Mayor de San Andrés, La Paz, Bolivia.

6 Publications

2013 R. Wilches, S. Voigt, **P. Duchen**, S. Laurent, and W. Stephan. Selective sweeps versus moderate allele frequency shifts: the signatures of selection at a QTL for cold tolerance in *D. melanogaster*. Submitted to *PLoS Genetics*.

2013 **Duchen, P.**, D. Živković, S. Hutter, W. Stephan, and S. Laurent (2013). Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. *Genetics* 193(1).

2012 Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, **P. Duchen**, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley (2012). Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genetics* 8(12).

2012 **Duchen, P.** and S. Beck (2012). Fabaceae systematics of the National Park of Cotapata (La Paz - Bolivia). *Revista de la Sociedad Boliviana de Botanica* 6(2).

2010 **Duchen, P**. and S. Renner (2010). The evolution of *Cayaponia* (Cucurbitaceae): repeated shifts from bat to bee pollination and long-distance dispersal to Africa 2-5 million years ago. *American Journal of Botany* 97(7).

2009 Morrison III W., J. Lohr, **P. Duchen**, R. Wilches, D. Trujillo, M. Mair, and S. Renner (2009). The impact of taxonomic change on Conservation: Does it kill, can it save, or is it just irrelevant? *Biological Conservation* 142(12).

7 Professional service

Membership	Society for the Study of Evolution
Reviewing for	Molecular Ecology Genetics

8 Programming skills

C, C++, Perl, R, Unix shell-scripting, Latex.