MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 1499–1509 doi: 10.1111/mec.13137

Population genomic analysis uncovers African and European admixture in *Drosophila melanogaster* populations from the south-eastern United States and Caribbean Islands

JOYCE Y. KAO, ASIF ZUBAIR, MATTHEW P. SALOMON, SERGEY V. NUZHDIN and DANIEL CAMPO

Section of Molecular and Computational Biology, Department of Biology, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

Abstract

Drosophila melanogaster is postulated to have colonized North America in the past several 100 years in two waves. Flies from Europe colonized the east coast United States while flies from Africa inhabited the Caribbean, which if true, make the south-east US and Caribbean Islands a secondary contact zone for African and European _D. melanogaster_. This scenario has been proposed based on phenotypes and limited genetic data. In our study, we have sequenced individual whole genomes of flies from populations in the south-east US and Caribbean Islands and examined these populations in conjunction with population sequences from the west coast US, Africa, and Europe. We find that west coast US populations are closely related to the European population, likely reflecting a rapid westward expansion upon first settlements into North America. We also find genomic evidence of African and European admixture in south-east US and Caribbean populations, with a clinal pattern of decreasing proportions of African ancestry with higher latitude. Our genomic analysis of _D. melanogaster_ populations from the south-east US and Caribbean Islands provides more evidence for the Caribbean Islands as the source of previously reported novel African alleles found in other east coast US populations. We also find the border between the south-east US and the Caribbean island to be the admixture hot zone where distinctly African-like Caribbean flies become genomically more similar to European-like south-east US flies. Our findings have important implications for previous studies examining the generation of east coast US clines via selection.

Keywords: admixture, *Drosophila melanogaster*, ecological genomics, population genomics, population structure

Received 14 September 2014; revision received 23 February 2015; accepted 25 February 2015

Introduction

According to the currently accepted demographic model, *D. melanogaster* originated in sub-Saharan Africa with a migration event into the European continent 10 000 years ago (David & Capy 1988). Colonization of the Americas is hypothesized to have happened in two waves. The first wave occurred ~400–500 year ago with African flies being transported into the Caribbean Islands along with the transatlantic slave trade. The second wave, which happened in the mid-19th century, was the cosmopolitan flies arriving with European settlers into North America (David & Capy 1988). These two waves purportedly created a secondary contact zone in the south-east United States and Caribbean Islands of cosmopolitan-adapted flies from Europe and African-like flies from West Africa (Caracristi & Schlötterer 2003; Duchen *et al.* 2013; Bergland *et al.* 2014) (Fig. 1). The flies originating from the Caribbean

Correspondence: Joyce Y. Kao, Fax: 213 821 4257; E-mail: joycekao@usc.edu

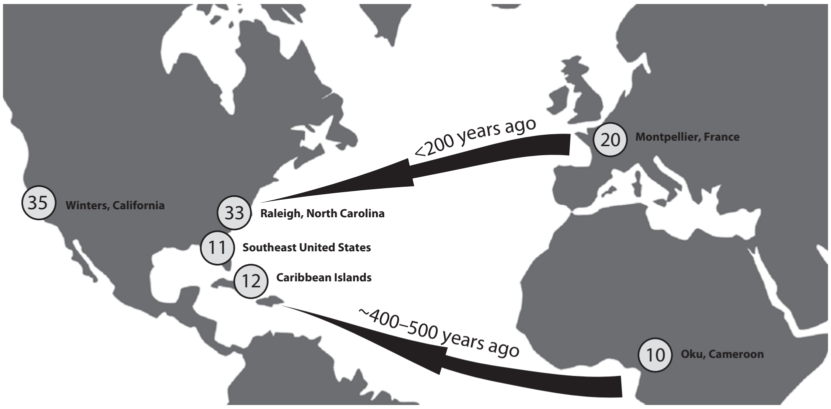


Fig. 1 Map of sequenced populations with number of whole-genome sequences in circles. Arrows indicate currently accepted migration history of *D. melanogaster* into the Americas.

islands have retained African-like behaviour and physical phenotypes despite its close proximity to the US cosmopolitan populations. Yukilevich & True (2008b) showed that pigmentation patterns of Caribbean flies resemble those from West Africa and become increasingly more cosmopolitan with increasing latitude into the United States. Additionally in the same study, morphological measurements such as thorax length, wing length and thorax luminosity also showed the same pattern in Caribbean flies. With regard to African behaviours, it has been suggested that Caribbean male courtship is more similar to African males than to American male courtship (Yukilevich & True 2008b). Moreover, based on mating preferences, it was shown that Caribbean flies freely mated with West African flies, which showed partial sexual isolation with US flies (Yukilevich & True 2008a).

Previous studies looking at genomewide effects of divergence in these populations used tiling microarrays to detect highly differentiated regions between the pooled genomes of cosmopolitan populations (including Caribbean fly lines) and Zimbabwean populations and then sequenced a subset of fragments to look at genetic divergence (Yukilevich *et al.* 2010). Most differentiation was found between populations living in African vs. out of Africa and evidence supporting that most of the variation in North America and African populations originated from the sorting of African standing genetic variation into the New World through Europe (Yukilevich *et al.* 2010). However, Caracristi & Schlötterer (2003) found high levels of polymorphism in North American *D. melanogasterI* populations where the proportion of shared alleles between African and North American flies were greater than the proportion of shared alleles between African and European populations. This evidence supports the hypothesis that there was a separate migration event to the Caribbean and that this might be the source of putative African alleles in North America (Li & Stephan 2006). More recently, Duchen *et al.* (2013) showed that one east coast North American population of *D. melanogaster* are most likely the result of an admixture event between European and African populations with the African ancestry accounting for 15% of the mixture. However, it is not clear from their study whether there was a second migration event to the Caribbean from Africa and how the influx of African alleles would look like coming from the Caribbean Islands. Yukilevich *et al.* (2010) began the investigation of African alleles in the Caribbean and through a pooled DNA chip sequencing approach detected that the Caribbean populations had more African ancestral alleles relative to US populations and a lower amount of genetic divergence from Africa compared to US African comparisons. Their approach provided a first coarse peek at African/European admixture in this area of the world. The Caribbean islands have been claimed to be the source of additional African alleles in the North American populations (Caracristi & Schlötterer 2003) although it has never been further investigated.

Examining African and European admixture in south-east US and Caribbean *D. melanogaster* populations not only extends the knowledge of North American colonization, but also has important implications for previous clinal studies on populations on the entirety of the east coast US (Sezgin *et al.* 2004; Schmidt & Paaby 2008; Paaby *et al.* 2010, 2014). Particularly, having the presence of admixture would call into question the generation of clinal variation by natural selection in this particular cline. Before the recent studies of admixture on the east coast US (Duchen *et al.* 2013; Bergland *et al.* 2014), spatially varying selection was the main force generating clinal patterns. However, with the introduction of admixture, one can now imagine a scenario where directional gene flow from admixed populations could also contribute to patterns varying with latitudes along the east coast US. Bergland *et al.* (2014) has illustrated that indeed clinal variation could very well be attributed to secondary contact and local adaptation. However, none of the recent studies have examined populations past the borders of the United States. Because there is evidence indicating that the source of novel east coast US African alleles is harboured in Caribbean populations (Caracristi & Schlötterer 2003; Yukilevich *et al.* 2010), it is important to examine the interactions between south-east US and Caribbean populations to elucidate how the genomic patterns of admixture potentially affected the rest of the east coast US.

For this work, we have sequenced 23 *D. melanogaster* genomes from various locations in the south-east United States and the Caribbean Islands. Combined with the current sequencing efforts of other fly populations from Raleigh (NC, USA), Winters (CA, USA), Montpellier (France) and Oku (Cameroon), we can explore the interface of African and European admixture in North American populations in an attempt to elucidate the history of *D. melanogaster*'s migration to the Americas and to understand how Caribbean *D. melanogaster* populations can retain African-like phenotypes while being in such close proximity to European-like neighbouring populations from the United States.

Materials and methods

Drosophila melanogaster lines for sequencing

A subset of 23 isofemale lines of *D. melanogaster* from 12 locations collected and used by Yukilevich & True (2008b) in a previous study was selected for sequencing. Origins are as follows with ID numbers corresponding to locations in Yukilevich & True (2008b) where map coordinates can be found: Selba, AL (ID#: 20, 28 and 20, 17); Thomasville, GA (ID#: 13, 34 and 13, 29); Tampa Bay, FL (ID#: 4, 12 and 4, 27); Birmingham, AL (ID#: 21, 39 and 21, 36); Meridian, MS (ID#: 24, 2 and 24, 9); Sebastian, FL (ID#: 28, 8); Freeport, Grand Bahamas-west (ID#: 33, 16 and 33, 11); George Town, Exumas (ID#: 36, 9 and 36, 12); Bullock's Harbor, Berry Islands (ID#: 40, 23 and 40, 10); Cockburn Town, San Salvador (ID#: 42, 23 and 42, 20); Mayaguana, Mayaguana (ID#: 43, 19 and 43, 18); Port Au Prince, Haiti (ID#: H, 29 and H, 25). All flies were maintained at 25 °C in vials on a standard cornmeal diet.

Libraries and sequencing of south-east US and Caribbean lines

All lines were subjected to full-sibling inbreeding for at least five generations before we collected 15–20 females from each line for library preparation. DNA was extracted using an Epicentre MasterPure kit (Madison, WI, USA) and cleaned with the Zymo Quick-gDNA Miniprep kit (Irvine, CA, USA). Illumina sequencing libraries were prepared according to Dunham & Friesen (2013) with the exception that DNA was sheared with dsDNA Shearase Plus (Zymo: Irvine, CA, USA) and cleaned using Agencourt AMPure XP beads (Beckman-Coulter: Indianapolis, IN, USA). Fragment size selection was also carried out using beads instead of gel electrophoresis. Libraries were visualized in an Agilent Bioanalyzer 2100 and quantified using the Kapa Biosystems Library Quantification Kit, according to manufacturer's instructions. Libraries were loaded into an Illumina flow cell v.3 and run on a HiSeq 2000 for 2 × 100 cycles. Library quality control and initial sequencing were performed at the USC NCCC Epigenome Center's Data Production Facility (University of Southern California, Los Angeles, CA, USA). Additional sequencing to achieve at least 5× genomewide average coverage for all lines was performed at the USC UPC Genome and Cytometry Core (University of Southern California, Los Angeles, CA, USA), in an Illumina HiSeq 2500 following the same run format. Sequencing data are available under NCBI BioProject number, PRJNA274815.

Sources of other sequenced populations

We used the 35 isogenic lines from Winters, CA, USA and 33 isogenic lines from Raleigh, NC, USA described in Campo *et al.* (2013). Raleigh lines were a subset of the Drosophila Genetic Reference Panel (DGRP) (Mackay *et al.* 2012). The 10 isofemale lines from Oku, Cameroon, were sequenced as a part of the Drosophila Population Genetic Panel (DPGP-2 African Survey) (Pool *et al.* 2012). The Cameroon population in the DPGP African Survey was chosen based on previous colonization history of *D. melanogaster* (David & Capy 1988). Additionally, it is one of the African populations least affected by recent non-African admixture (Pool *et al.* 2012). Sequencing reads for 20 isofemale lines from Montpellier, France were downloaded via the Bergman laboratory webpage (Bergman & Haddrill 2015). Pooled sequencing data sets for other European populations are available, but are not suitable for some of the types of analysis described below. Map locations of all populations used in this study are reflected in Fig. 1.

Mapping

For each fly line, the raw sequencing reads were trimmed by quality using the SOLEXAQA package (ver. 1.12) with default parameters and all trimmed reads <25 bp were discarded (Cox *et al.* 2010). The quality trimmed reads were then mapped to the *D. melanogaster* reference genome (FLYBASE version 5.41) using BOWTIE 2 (ver. beta 4) with the 'very sensitive' and '-N = 1' parameters (Langmead & Salzberg 2012). Following mapping, the GATK (ver. 1.1-23, DePristo *et al.* 2011) IndelRealigner tool was used to perform local realignments around indels and PCR and optical duplicates were identified with the MarkDuplicates tool in the Picard package (http://picard.sourceforge.net).

SNP calling, phasing and filtering

SNP variants were identified in all lines simultaneously using the GATK UnifiedGenotyper (ver. 2.1-8) tool with all parameters set to recommended default values. The raw SNP calls were further filtered following the GATK best practices recommendations (Auwera *et al.* 2013) resulting in 4 021 717 SNP calls. We then used BEAGLE to perform haplotype phasing as well as impute missing data (Browning & Browning 2007, 2009). SNPs were further filtered using VCFtools (http://vcftools.sourceforge.net/) for 5% minor allele frequency and biallelic sites resulting in 1 047 913 SNPs across the major chromosomal regions: 2L (222 464 SNPs), 2R (192 120 SNPs), 3L (212 601 SNPs), 3R (268 701 SNPs) and X (152 027 SNPs) to be considered for further analysis.

Population structure analysis

We used VCFtools (Danecek *et al.* 2011) to calculate F_{ST} via the Weir and Cockerham estimates (1984) as a proxy for genetic distance between all our populations. F_{ST} values can be inflated when the amount of intrapopulation variation is too low relative to the between population divergence (Cruickshank & Hahn 2014). That bias might only be a problem when comparing either different species, or highly divergent populations of the same species such as the case with our African vs. non-African populations. Therefore, we also calculated D_{XY} between all pairs of populations and compared it to our F_{ST} estimates. We first generated a single population-specific reference file by updating the *D. melanogaster* reference genome (FLYBASE version 5.41) with population-specific SNPs using the GATK FastaAlternateReferenceMaker tool (DePristo *et al.* 2011). We then calculated D_{XY} in 10K windows for each pair of populations using an accessory script, calculate-dxy.pl, from the PoPoolation software package (Kofler *et al.* 2011).

Additionally, we used the R package SNPRelate (Zheng *et al.* 2012) to perform principal component analysis (PCA). We did PCA with all populations and then removed the Cameroon population for another PCA to investigate North American patterns further without the influence of the African population.

ADMIXTURE (Alexander *et al.* 2009) estimates ancestry of a given set of unrelated individuals in a model-based manner from large autosomal SNP genotype data sets. The program outputs the proportion of ancestral population for each individual. To run the program, a prior belief number of ancestral populations (K) must be provided. We used a cross-validation procedure of ADMIXTURE to propose the number of ancestral populations (K). Optimal K values will have lower cross-validation error compared to other values. We ran a fivefold cross-validation on a plink file (.ped) which was generated using a custom PERL script from the Variant Calling File (VCF). Linkage disequilibrium can affect the results of ADMIXTURE thus the marker set used for this analysis was further filtered to include only autosomal markers that were at least 250 bp apart resulting in a total of 234 497 SNPs.

Chromosome painting

We utilized the software Chromopainter (Lawson *et al.* 2012) to estimate which parts of the genome each North American individual were contributed by European or African ancestors. We ran Chromopainter for 60 iterations to estimate parameters of the algorithm and then ran Chromopainter with the estimated parameters to obtain the final results as recommended in the user manual. Additionally, we implemented hierarchical clustering in R (heatmap.2 with standard options in the gplots library) to examine the similarity of Chromopainter results across each chromosomal region between all the North American individuals. To compare the statistical significance of proportion of African ancestry in different populations, we calculated the proportion as predicted by Chromopainter of each individual in all populations and then applied the Wilcoxon rank sum test in R to look for significant differences in predicted African ancestry proportions between populations.

Linkage disequilibrium analysis

To look at linkage disequilibrium decay over genomic distance, measures of D′ were estimated using VCFtools (Danecek *et al.* 2011) in 10 000 bp windows across the genome.

Results

Investigating population structure by principal component analysis

To explore initial relationships between populations, we performed PCA on the 1 047 913 quality-filtered SNPs using the R package SNPRelate. The first principal component represented the separation between African and non-African populations and the second principal component was the variation within the Cameroon population (Fig. 2). Upon closer inspection of the non-African cluster (Fig. 2), the first principal component could also be a proxy to how genetically close each non-African population is to the Cameroon population, with the Caribbean population located the closest. The non-African populations were approximately grouped into two subclusters of Caribbean and non-Caribbean. There were, however, a few Caribbean fly lines that clustered close to and within the non-Caribbean group. The four Caribbean lines that clustered with the US populations were collected from locations on islands closest to the US and Caribbean border (i.e. Freeport, Grand Bahamas-west and Bullock's Harbor, Berry Islands). Along with these four Caribbean lines, the sequenced fly lines from locations in the south-east United States were interspersed with fly lines from Raleigh, indicating a potential east coast US admixture zone. The Raleigh population clustered very closely with the Winters, but both Raleigh and Winters visually appeared to still be distinct populations. The 20 French lines appeared dispersed in the non-Caribbean cluster, which supports the notion that there is much European influence in North American populations.

Upon inspection of additional principal components (Fig. S1, Supporting information), principal components 3 and 4 explained variation within the Cameroon population indicating there was much diversity in the African population, which is known from previous study (Pool *et al*. 2012). We believe that the variation in the Cameroon population may indicate the mixed history of African ancestry for this population rather than influence from non-African admixture because the non-African proportion has been estimated to be on average just under 4% (Pool *et al*. 2012). Additional PC's did not explain large portions of the variation in the data, and thus, we did not look at additional PCs (Fig. S2, Supporting information).

As the African diversity may have been masking patterns in the non-African populations, we removed the Cameroon population and performed a second PCA using non-African populations (Fig. S3, Supporting information). The first principal component in this second PCA explained the variation within the North American populations preserving the same pattern of 'more' African and 'less' African as shown in the previous analysis (Fig. 2). The second principal component separated the French population from the North American populations. Clustering patterns of the second PCA were similar to those in the first PCA with the French population forming a distinct cluster and being closest to the group containing Winters, Raleigh and south-east US populations. The third and fourth principal components accounted for more variation within the North American populations (Fig. S4, Supporting information).

Genetic differentiation between populations

To quantify the level of genetic differentiation, we calculated Weir and Cockerham (1984) F_{ST} between all pairs of populations per SNP and averaged the F_{ST} estimates per chromosomal region. We also calculate D_{XY} values to confirm our African vs. non-African F_{ST} comparisons.

We found a consistent pattern in which Cameroon was highly differentiated from all cosmopolitan populations, but was closest to the Caribbean population as

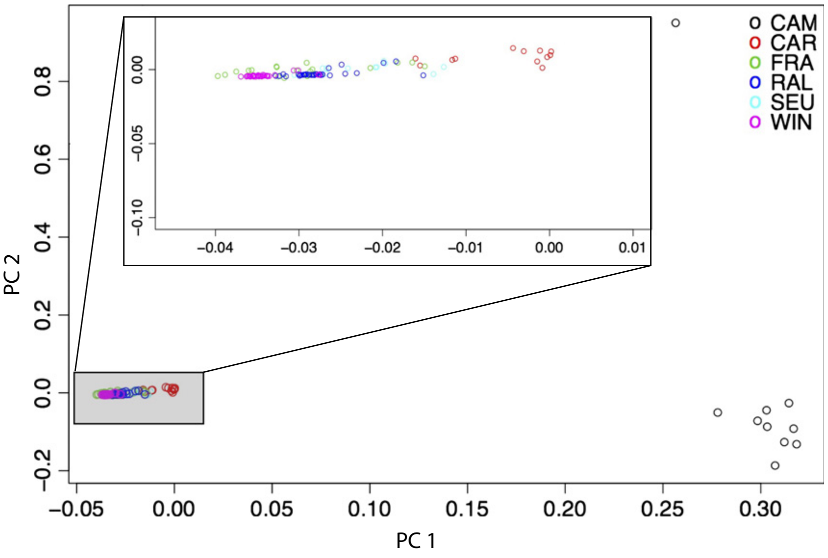
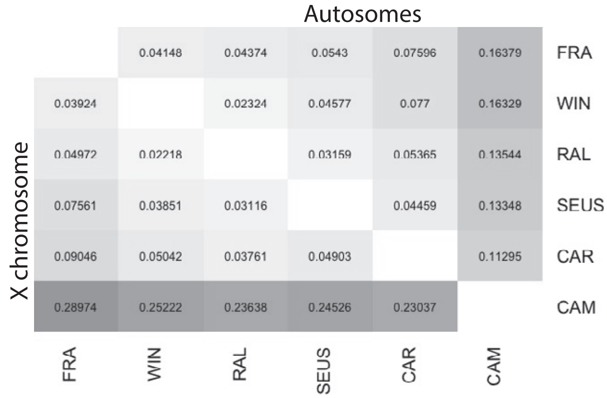


Fig. 2 First and second principal components (PC) from principal components analysis with populations from Cameroon (CAM), Caribbean Islands (CAR), France (FRA), Raleigh (RAL), south-east US (SEU) and Winters (WIN). Population structure of individuals in the grey highlighted box is magnified in secondary enlarged plot.

also indicated in Yukilevich *et al.* (2010) (Fig. 3). The French and Winters populations were the most differentiated from the Cameroon lines. The greatest differentiation between the Cameroon population and the non-African populations was on the X chromosome (Fig. 3). D_{XY} values confirmed our African vs. non-African F_{ST} values (Table S3, Supporting information).

The French population was the least genetically differentiated from the Winters and Raleigh populations (Fig. 3). Interestingly enough, the Caribbean population was slightly more differentiated from the Winters population than from the French population in the 2L and 3R chromosomal regions (Tables S1 and S2, Supporting information), perhaps indicating a slightly larger European influence in the Caribbean than the west coast US.



	FRA	WIN	RAL	SEUS	CAR	CAM	
		0.04148	0.04374	0.0543	0.07596	0.16379	FRA
	0.03924		0.02324	0.04577	0.077	0.16329	WIN
	0.04972	0.02218		0.03159	0.05365	0.13544	RAL
	0.07561	0.03851	0.03116		0.04459	0.13348	SEUS
	0.09046	0.05042	0.03761	0.04903		0.11295	CAR
	0.28974	0.25222	0.23638	0.24526	0.23037		CAM

Fig. 3 Average F_{ST} values between populations for chromosome X (lower diagonal) and all autosomes (upper diagonal). Shades of grey illustrate the degree of genetic differentiation with larger F_{ST} values being darker and smaller F_{ST} values being lighter.

Admixture patterns

From our cross-validation procedure, it was determined that the optimal number of ancestral populations for ADMIXTURE was $K = 2$ (Fig. S5, Supporting information). According to the ancestral proportions (Fig. 4a), it appears that the North American lines are a composite of European and African ancestry. Furthermore, the proportion of African-like markers is higher in Caribbean individuals and decrease in proportion with increasing latitude (Fig. 4b). We additionally performed the same analysis for the X chromosome, but found almost no African/European admixture in the non-African populations (Fig. S6, Supporting information).

Genomewide African and European influences

While results from ADMIXTURE are useful in understanding how populations are structured and point towards approximate the influences of African and European ancestors, we cannot determine the pattern of influence across a genome using that algorithm. Therefore, we used Chromopainter to estimate the ancestry of all the North American fly lines across the genome. The most striking result from visualizing the local ancestry of all genomes (Fig. 5) was that larger chunks of African or European ancestry seemed to be retained in telomeric and centromeric regions known to have low recombination (Comeron *et al.* 2012).

When we clustered individual genomes by genomic inheritance patterns, the patterns of individuals within one population clustered more with each other than with other populations except for chromosomal region 2R where Caribbean and south-east US individuals

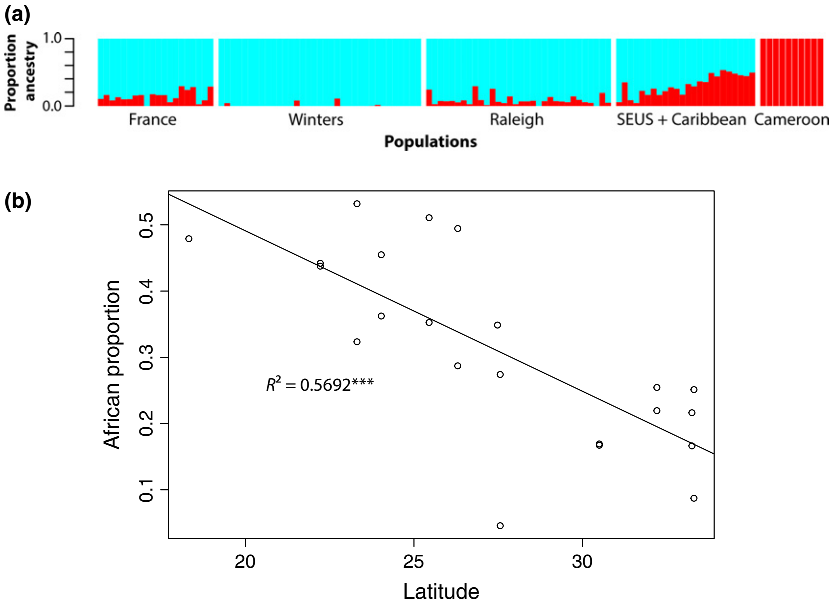


Fig. 4 (a) ADMIXTURE results of quality and LD filtered autosomal markers for two ancestral populations ($K = 2$). (b) Relationship between latitude and proportion of African ancestry of south-east US and Caribbean individuals. Asterisks on the $R^2 = 0.5692$ correspond with $F = 26.42$ and a significance of $P < 0.0001$.

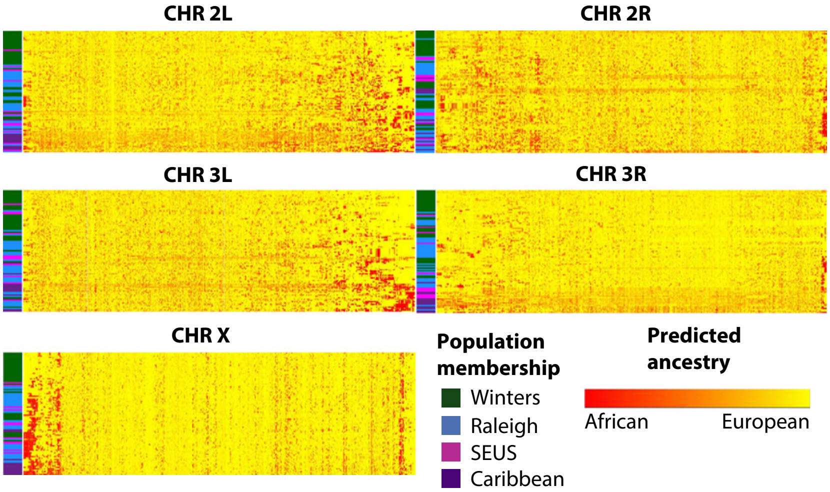


Fig. 5 Painted chromosomal regions heatmap with hierarchical clustering of individuals. Each row in heatmap represents one individual. Population membership of individual designated by vertical bar to the right of each chromosomal heatmap (Green: Winters, CA, Blue: Raleigh, NC, Pink: South-east US, Purple: Caribbean). Red represents SNPs that are most similar to the Cameroon donor population; Yellow represents SNPs that are most similar to the French donor population.

seem to be evenly dispersed among Winters and Raleigh populations. Chromosome X appeared to be the least influenced by African ancestry (Fig. 5).

Individuals from the Caribbean populations and some from the south-east US had a larger percentage of African-painted alleles compared to Raleigh and Winters (Wilcoxon, $W = 84$, $P < 0.0001$), which was visually apparent in the chromosomal regions of 2L and 3R (Wilcoxon, $W = 185$ and $W = 206$, both $P < 0.0001$) (Fig. 5). The long stretches of the African-painted SNPs in these chromosomal regions coincided with the locations of common cosmopolitan inversions, In(2L)t and In(3R)P (Corbett-Detig & Hartl 2012).

Overall, the expected proportion of probable African ancestry ranged between 3.6% (Winters, CA) to 47% (Caribbean Islands) for the painted genomes. On average over the whole genome, the estimated percentage of African ancestry was highest in the Caribbean population at 24.75%, which is expected if these flies were established by African ancestors in a more recent time (David & Capy 1988; Yukilevich *et al.* 2010). The lowest percentage of African ancestry was in the Winters population at 8.68% (Wilcoxon, $W = 0$, $P < 0.0001$). Raleigh and south-east US populations had 14% and 15.6% of predicted African ancestry, which is consistent with previous findings (Duchen *et al.* 2013; Bergland *et al.* 2014). In summary, populations had decreasing African ancestry with respect to geographical distance from the Caribbean Islands in all genomic areas. Out of all the chromosomes, the X had the lowest estimated percentage of African-inherited alleles for all North American populations (Fig. S7, Supporting information).

Linkage disequilibrium patterns

Elevated levels of linkage disequilibrium (LD) can be an indicator of recent admixture in populations because inherited ancestral tracts have not had sufficient time to be broken down by recombination (Loh *et al.* 2013). We calculated D′ as a measure of LD and averaged the absolute value of D′ to get approximate LD levels in our populations across different genomic regions. We found that on average Cameroon and France have lower LD values than North American populations (Fig. 6). Out of all the North American samples, the Caribbean had one of the lowest LD values on most chromosomal regions except the X.

Discussion

Caribbean flies most likely established by African ancestors

Although all non-African populations pairwise F_{ST} values were high throughout the genome when compared to the African sample, the Caribbean population had on average the lowest values. With the Caribbean population located closest in the first PC analysis to the Cameroon population and the highest percentage of predicted African ancestry out of all the North American samples we analysed, these pieces of evidence do seem to further support the migration event of west African flies to the Caribbean islands via the transatlantic slave trade (David & Capy 1988).

African and European admixture in North America

Recently, admixed populations exhibit more linkage disequilibrium than older long-established populations (Loh *et al.* 2013). This is because newer populations, which are a combination of genetic material from older base populations, have not gone through enough generations for recombination to break down LD blocks. We do detect higher LD in the North American populations

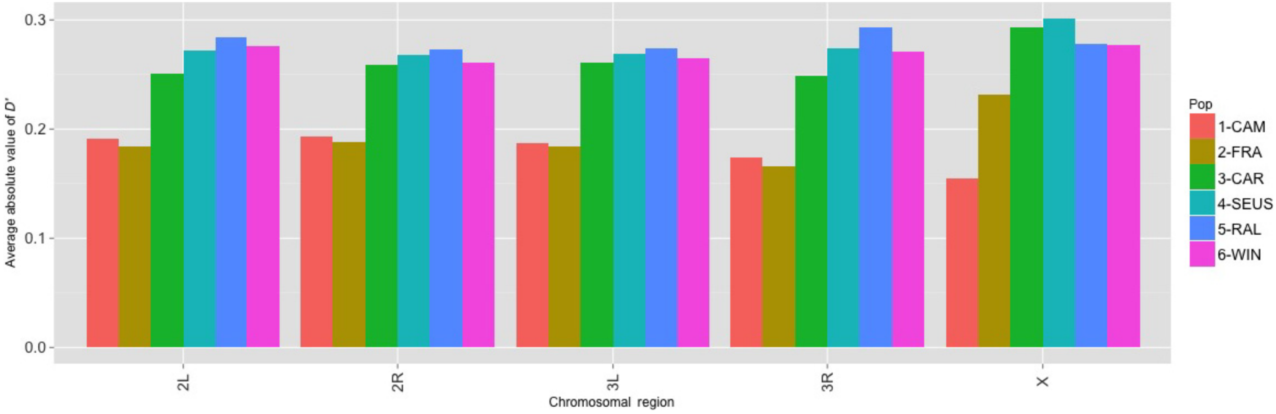


Fig. 6 Average |*D*′| as a measure of linkage disequilibrium by population and chromosome.

than in our African and European samples. Although this is a common signature of admixture, higher LD values can also result from other demographic events such as a population bottleneck or positive selection. However, previous studies have already established the existence of admixture in some North American populations, particularly Raleigh, (Duchen *et al.* 2013; Bergland *et al.* 2014) which would support that elevated LD in our case is most likely due to admixture.

We are able to extend the admixture scenario in North America with our 23 sequenced genomes from the south-east US and Caribbean islands. It has been postulated that American *D. melanogaster* are more genetically variable than European *D. melanogaster* due to admixture from the Caribbean islands (Caracristi & Schlötterer 2003). Our results from ADMIXTURE (Fig. 4) and chromosome painting (Fig. 5) clearly show a clinal pattern of African introgression into the United States, which supports the notion that these African alleles in the US are originating from the Caribbean Islands. Furthermore, the PCA groupings (Fig. 2) also illustrate that the border between the south-east US and Caribbean Islands is where fly populations are experiencing the most admixture. Previous work on premating reproductive isolation has shown distinct phenotype and male courtship patterns of American and Caribbean populations (Yukilevich & True 2008a,b). A more recent study indicating the potential effects of admixture show that postmating reproductive traits such as lowered fertility and increased sperm toxicity tolerance towards genetically unfamiliar males are affected particularly in locations corresponding to this admixture hot zone at the border of the south-east US and Caribbean islands (Kao *et al.* 2014).

According to our ADMIXTURE and Chromopainter analysis, the X chromosome was the least affected by African and European admixture with non-African populations having more similar X chromosomes than our Cameroon population (Fig. 5 and Fig. S6, Supporting information). The reduced African ancestry in the sex chromosome could be attributed to sex-biased gene flow from Africa due to mating preferences with African flies having strong preferences to mate within their populations (Wu *et al.* 1995; Hollocher *et al.* 1997), which has been previously demonstrated (Yukilevich & True 2008a). With this partial sexual isolation based on mating preferences, the X chromosome is more likely to be more cosmopolitan than African.

Westward expansion of Drosophila melanogaster

Our analysis of the Winters, CA genomes revealed that the Winters population is more related to the French population than to the other US populations. There appears to be very little to no African ancestry in the genomes from Winters, CA. Either there was a separate colonization event in the west or when *D. melanogaster* arrived in North America with European settlers, it quickly expanded to the west (Campo *et al.* 2013). The latter explanation may be more plausible given that the first sighting of *D. melanogaster* was in the mid-19th century (David & Capy 1988), which was when the United States was in the midst of active westward expansion with the rapid construction of a transcontinental railway to transport supplies out to early settlers in the west (Billington & Ridge 2001) with flies likely following the expansion (Keller 2007).

Conclusions and implications

Understanding the origins and genomic patterns of North American *D. melanogaster* will be useful for researchers working with populations from this area of the world especially with the emerging public sequencing data becoming available (Mackay *et al.* 2012; Remolina *et al.* 2012). Our genome analyses of south-east US and Caribbean fly populations in relation to other North American populations and to their African and European ancestral populations further elucidate the history of *D. melanogaster* colonization of North America. We reveal clinal patterns of African ancestry from the Caribbean Islands to the south-east United States illustrating African and European admixture maintained in those populations, which is likely influencing populations that lie farther north on the east coast of the United States.

Our results do not conflict with previous clinal patterns found in many other studies (Sezgin *et al.* 2004; Schmidt & Paaby 2008; Paaby *et al.* 2010, 2014). However, it does call into question the driving force behind the generation of clines at least for east coast US populations. While it makes sense that clinally varying selection could produce such geographical patterns, the recent studies on African and European admixture (Duchen *et al.* 2013; Bergland *et al.* 2014; present study) make it likely that steady gene flow from Caribbean populations could also be partly responsible for the generation of east coast US clinal patterns. Care must be taken to interpret results emanating from some Floridian populations as it seems that admixture could affect certain phenotypes (Kao *et al.* 2014). Investigations of actual migration between populations along east coast United States and Caribbean Islands would further illuminate the influence of African and European admixture on clinal patterns.

Acknowledgements

We would like to thank Alan Bergland, Annalise Paaby and Peter Ralph for useful discussions pertaining to this work and comments on previous versions of this manuscript as well as Joseph Dunham for aid with sequencing library construction. We also thank Roman Yukilevich for bestowing upon us his south-east US and Caribbean fly collection this work was based on. This work was made possible by two NIH grants, GM102227 and MH091561.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Auwera GA, Carneiro MO, Hartl C *et al.* (2013) *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Bergland AO, Tobler R, Gonzalez J, Schmidt P, Petrov D (2014) Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. *bioRxiv*. doi: 10.1101/009084.

Bergman CM, Haddrill PR (2015) Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. *F1000 Research*, **4**, 31.

Billington RA, Ridge M (2001) The transportation frontier. In: *Westward Expansion: A History of the American Frontier*, 6th edn. *Abridged*, pp. 279–289. University of New Mexico Press, Albuquerque, New Mexico.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**, 1084–1097.

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, **84**, 210–223.

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

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

Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Comeron JM, Ratnappan R, Bailin S (2012) The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1002905.

Corbett-Detig RB, Hartl DL (2012) Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genetics*, **8**, e1003056.

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

David J, Capy P (1988) Genetic variation of *Drosophila melanogaster* natural populations. *Trends in Genetics: TIG*, **4**, 106–111.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Duchen P, Živković D, Hutter S, Stephan W, Laurent S (2013) Demographic inference reveals african and european admixture in the North American *Drosophila melanogaster* population. *Genetics*, **193**, 291–301.

Dunham JP, Friesen ML (2013) A cost-effective method for high-throughput construction of illumina sequencing libraries. *Cold Spring Harbor Protocols*, **2013**, 820–834.

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.

Kao JY, Lymer S, Hwang SH, Sung A, Nuzhdin SV (2014) Postmating reproductive barriers contribute to the incipient sexual isolation of US and Caribbean *Drosophila melanogaster*. *bioRxiv*. doi: 10.1101/007765.

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

Kofler R, Orozco-terWengel P, De Maio N *et al*. (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**, e1002453.

Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genetics*, **2**, e166.

Loh P-R, Lipson M, Patterson N *et al*. (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, **193**, 1233–1254.

Mackay TFC, Richards S, Stone EA *et al*. (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.

Paaby AB, Blacket MJ, Hoffmann AA, Schmidt PS (2010) Identification of a candidate adaptive polymorphism for *Drosophila* life history by parallel independent clines on two continents. *Molecular Ecology*, **19**, 760–774.

Paaby AB, Bergland AO, Behrman EL, Schmidt PS (2014) A highly pleiotropic amino acid polymorphism in the *Drosophila* insulin receptor contributes to life-history adaptation. *Evolution*, **68-12**, 3395–3409.

Pool JE, Corbett-Detig RB, Sugino RP *et al*. (2012) Population genomics of sub-saharan *Drosophila melanogaster*: African diversity and Non-African admixture. *PLoS Genetics*, **8**, e1003080.

Remolina SC, Chang PL, Leips J, Nuzhdin SV, Hughes KA (2012) Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution*, **66**, 3390–3403.

Schmidt PS, Paaby AB (2008) Reproductive diapause and life-history clines in North American populations of *Drosophila melanogaster*. *Evolution*, **62-5**, 1204–1215.

Sezgin E, Duvernell DD, Matzkin LM, Duan Y, Zhu CT, Verelli BC, Eanes WF (2004) Single-locus latitudinal clines and their relationship to temperate adaptation in metabolic genes and derived alleles in *Drosophila melanogaster*. *Genetics*, **168**, 923–931.

Wu CI, Hollocher H, Begun DJ, Aquadro CF, Xu Y, Wu ML (1995) Sexual isolation in *Drosophila melanogaster*. a possible case of incipient speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 2519–2523.

Yukilevich R, True JR (2008a) Incipient sexual isolation among cosmopolitan *Drosophila melanogaster* populations. *Evolution*, **62**, 2112–2121.

Yukilevich R, True JR (2008b) African morphology, behavior, and pheromones underline incipient sexual isolation between US and Caribbean *Drosophila melanogaster*. *Evolution*, **62**, 2807–2828.

Yukilevich R, Turner TL, Aoki F, Nuzhdin SV, True JR (2010) Patterns and processes of genome-wide divergence between North American and African *Drosophila melanogaster*. *Genetics*, **186**, 219–239.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**, 3326–3328.

J.Y.K. helped to conceptualize the project, prepared the genomic libraries for sequencing, coordinated data analysis among authors, performed PCA, Chromopainter, F_{ST} and LD analysis, wrote and organized manuscript. A.Z. performed the ADMIXTURE analysis, wrote ADMIXTURE section in Materials and Methods, helped to summarize the *D*xy analysis and edited the manuscript. M.P.S. aligned and called SNPs for NGS reads, provided code for PCA analysis, wrote alignment and SNP calling in Materials and Methods, performed *D*xy analysis and edited the manuscript. S.V.N. helped to conceptualize the project and edited the manuscript. D.C. conceptualized the project, sequenced the genomic libraries and edited the manuscript.

Data accessibility

South-east United States and Caribbean Islands population genomic sequences: All Illumina data from this study are available under the NCBI BioProject PRJNA274815. DPGP Cameroon population genomic sequences: SRX058148-SRX058159. French population genomic sequences: NCBI SRA: ERR705945-ERR705964. Winters population genomic sequences: Illumina data for this population can be accessed under NCBI BioProject PRJNA74721. DGRP Raleigh population genomic sequences: http://dgrp2.gnets.ncsu.edu/data.html. GATK variants files, filtered vcf files used for analyses plus code and input files for F_{ST}, LD, PC, ADMIXTURE and Chromopainter analyses: Dryad doi:10.5061/dryad.446sv.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 First four principal components of PCA including populations from Cameroon, Caribbean Islands, France, Raleigh, southeast US, and Winters reveal that most variation explained is within the Cameroon population.

Fig. S2 Percentage of variation explained over all principal components.

Fig. S3 First and second principal components (eigenvectors) of PCA with the Cameroon population removed, but including populations from Caribbean Islands (CAR), France (FRA), Raleigh (RAL), southeast US (SEU), and Winters (WIN).

Fig. S4 First four principal components of PCA with the Cameroon population removed, but including populations from Caribbean Islands, France, Raleigh, southeast US, and Winters.

Fig. S5 Cross validation results for ADMIXTURE analysis to determine the optimal number of ancestral populations.

Fig. S6 ADMIXTURE X-chromosome analysis for number of groups $K = 2$.

Fig. S7 Expected proportion of African ancestry for each population by chromosomal region.

Table S1 Average F_{ST} values between populations for chromosome 2 divided by regions 2L (below diagonal) and 2R (above diagonal).

Table S2 Average F_{ST} values between populations for chromosome 3 divided by regions 3L (below diagonal) and 3R (above diagonal).

Table S3 Average D_{XY} values between populations for autosomes (above diagonal) and X chromosome (below diagonal).

Appendix S1 Table of eigenvalues and eigenvectors from principal component analysis.