LETTERS

Positive Selection Versus Demography: Evolutionary Inferences Based on an Unusual Haplotype Structure in Drosophila simulans

Humberto Quesada,*† Sebastián E. Ramos-Onsins,† Julio Rozas,† and Montserrat Aguadé†

*Departamento de Bioquímica, Genética e Imunología, Facultad de Ciencias, Universidad de Vigo, Vigo, Spain; and †Departamento de Genética, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

Coalescent simulations were used to investigate the possible role of population subdivision and history in shaping nucleotide variation in a recombining 88-kb genomic fragment of Drosophila simulans displaying an unusual large-scale haplotype structure. The multilocus analysis, based on summary statistics using specific demographic null models under recombination, indicates that the observed levels of linkage disequilibrium differed significantly from the values expected under different bottleneck and population admixture scenarios. These results indicate that demography alone may not account for the observed pattern of variation and support the previous claim that the data are better described by a model in which an adaptive mutation has not yet gone to fixation.

Uncertainty about the demographic history of populations can hinder genome-wide scans for selection (Harr et al. 2002; Glinka et al. 2003; Orengo and Aguadé 2004; Akey et al. 2004; Haddrill et al. 2005; Ometto et al. 2005; Schmid et al. 2005). Although demographic events apply to the whole genome and selective events are locus specific, the large variance of nucleotide diversity in stationary panmictic populations may generate large differences among loci just by chance. Indeed, different regions on a recombining chromosome have different genealogies whose sizes can differ considerably by genetic drift. These local fluctuations of variation may be amplified by the superimposed effect of demographic events and recombination, which, in the absence of selection, may mimic the pattern expected under natural selection. The theory of coalescence provides the framework to develop robust statistical tests and therefore to obtain the probability of empirical data under different evolutionary scenarios (Hudson 1990).

Drosophila simulans has historically served as an important model system in molecular and evolutionary genetics, although its demographic history is still far from being well understood (Lachaise et al. 1988; Powell 1997). In a recent study, Quesada et al. (2003) detected an unusual haplotype structure in a recombining 88-kb genomic fragment in an African population of this species. They found a core region of up to 38 kb with a major haplotype at intermediate frequency. This unusual haplotype structure gradually vanished from the core region until disappearing, thus supporting a recent (~6500 years ago [yra]) and incomplete selective sweep (Rozas et al. 2001; Quesada et al. 2003). The observed pattern is incompatible with a panmictic population in mutation-drift equilibrium. Here we analyze these data (Quesada et al. 2003) under 2 different nonequilibrium demographic scenarios in an effort to ascertain whether nonselective processes can explain the presence of 2 subsets of sequences (Parsch et al. 2001; Rozas et al. 2001). The first model considers population subdivision and subsequent admixture, which appears to be the most relevant scenario to African populations (Hamblin and Veuille 1999). The second model envisions a population of constant size that experiences a recent bottleneck with a few lineages surviving.

Demography and selection can both affect patterns of linkage disequilibrium (LD) in the genome (Przeworski 2002; Haddrill et al. 2005). To investigate the effects caused by population admixture and bottlenecks on nucleotide variation, we determined whether the overall LD between polymorphic sites, as measured by the $Z_{N_S}$ statistic (Kelly 1997), departs from that expected under each explicit demographic scenario. The $\Psi$ multilocus summary statistic ($\Psi = \sum_{i=1}^{n} (O_i - E_i)$) (Quesada et al. 2003), where $O_i$ is the observed $Z_{N_S}$ value in region $i$, $E_i$ is the $Z_{N_S}$ value expected under the corresponding null hypothesis, and $n$ is the number of regions ($n = 11$ in our case; sequence data from Quesada et al. 2003), was used for hypothesis testing. A positive and a negative $\Psi$ value indicate, respectively, an excess or a deficit of observed LD with respect to the value expected under the corresponding null demographic model. Expected $Z_{N_S}$ values for each region and the empirical distribution of $\Psi$ were obtained by neutral coalescent simulations (1,000 replicates) with recombination. The estimate of the population recombination parameter $(C_M = 0.0368)$ was obtained from the comparison of physical and genetic maps and assuming $N = 2 \times 10^6$ and $c = 0.92 \times 10^{-8}$ (Rozas et al. 2001; $C = 2Nc$ in Drosophila given that males do not recombine). Because summary statistics may be sensitive to assumptions about recombination rates under certain demographic models (Thornton 2005), 2 additional $C$ values were considered in the simulations. The first, $C_L$, which constitutes the lower bound of $C_M$, is based on the minimum number of recombination events ($R_M$) in the sample (Hudson 1987). It is defined as the lowest value of $C$ for which the right tail (5%) of the $R_M$ distribution (obtained by coalescent simulations) contains values equal or higher than the observed value of $R_M$ ($C_L = 0.0284$; Quesada et al. 2003). The second, $C_H$, represents the highest estimate of $C$ in the 3R chromosome of Drosophila melanogaster ($C_H = 0.0650$; see Hey and Kliman 2002).

Key words: Drosophila, selective sweep, coalescent simulations, bottleneck, population admixture.

E-mail: hquesada@uvigo.es.

Mod. Biol. Evol. 23(9):1643–1647. 2006
doi:10.1093/molbev/ms6031
Advance Access publication June 6, 2006

© The Author 2006. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
Simulations were conditioned on the number of segregating sites (S). Random data sets of DNA fragments as long as the surveyed fragment (88 kb) were generated in the simulations. The simulation program is available from the authors upon request. The P-value of the test (2-tailed test) was obtained as the proportion of computer replicates with \( \Psi \) values more extreme than the observed value.

To test for admixture, we considered a simple model where an ancestral population split into 2 subpopulations at time \( t_1 \), and admixture occurred at time \( t_2 \) (fig. 1). Times \( t_1 \) and \( t_2 \) (measured in \( 4N \) generations) were obtained from extant nucleotide variation. Under an admixture model, most nucleotide differences between the 2 subsets of sequences would have accumulated after the population split. Thus, the split time (\( t_1 = 1.04 \)) was estimated from the average silent nucleotide divergence between the 2 subsets of sequences and the estimated silent mutation rate per base pair and year (\( 1.4 \times 10^{-8} \); Rozas et al. 2001). Although \( t_1 \) might be overestimated by our method, computer simulations using \( t_1 = 1.04 \) yielded an average silent nucleotide divergence between the 2 most divergent subsets of sequences close to that observed (results not shown).

Similarly, recombination events between both subsets of haplotypes would have occurred after admixture. In this case, the time since admixture \( (t_2 = 0.003) \) was estimated as the average coalescent time required to account for the number of recombination events observed between the 2 sequence subsets (Quesada et al. 2003). An elevated level of LD is a typical signature of recent population admixture. However, as can be seen in figure 2, observed levels of LD were consistently and significantly much lower than expected under admixture. This conclusion is robust even to an error of at least one order of magnitude in the estimated time since admixture and also after considering very unequal effective sizes of both subpopulations (up to 1:99).

Incorporating some migration between subpopulations also resulted in highly significant departures from the admixture hypothesis, despite that the homogenizing effect of migration leads to a less extreme subdivision and, thus, to less power to reject the null admixture model (figs. 2 and 3). Similarly, the correlation in the genealogies of nearby segregating sites (and LD) decreases as \( C \), or the time since admixture, increases, and thus, there will be less useful information for making demographic inferences. However, varying the recombination rate did not have any substantial effect on the results. Only when considering a very ancient admixture event \( (t_2 \text{ two orders of magnitude higher than estimated}) \), and an unrealistically high recombination rate for this region, the admixture model was not rejected (fig. 2). Moreover, using \( t_1 \) values lower than 1.04 led to a decrease in LD and therefore made the rejection of the admixture model more unlikely. However, in neither of these more unlikely scenarios did simulated data show any remarkable haplotype structure \( (P < 0.05, \text{using the } \Psi \text{ statistic}) \).

A recent population bottleneck, occurring at different times after the last glacial maximum \( (~20,000 \text{ yrs}) \), cannot account for the observed data either. As the power of \( Z_{SS} \) decreases rapidly with the age of the bottleneck (Depaulis et al. 2003), we restricted our simulations to the time range with the highest power (0.00125–0.025 in \( 4N \) generations; fig. 3 in Depaulis et al. 2003). We simulated a population of initial effective size \( N_1 \), crashing to size \( N_2 \) at time \( h_1 \), and growing to the current effective population size \( N_0 \) at time \( t_0 \) (fig. 1). The severity of the bottleneck \( S_b \) is determined by the reduction in population size and its duration (Fay and Wu 1999). Furthermore, the distribution of summary statistics is not affected by the specific values of these 2 variables for bottlenecks of the same severity (Fay and Wu 1999; Orenco and Aguadé 2004). Severities ranging from weak \( (S_b = 0.005) \) to intermediate \( (S_b = 0.1) \) were considered to allow some lineages to survive the bottleneck, with \( N_b/N_1 \) values varying from \( 1 \times 10^{-1} \) to \( 5 \times 10^{-3} \) and times measured in \( 4N \) generations. Simulation results reveal that the observed level of LD is in all cases much lower than that expected under a bottleneck scenario (fig. 4), a pattern similar to that observed under the admixture model. LD values compatible with a bottleneck model were only observed after relatively weak bottlenecks \( (S_b \leq 0.01) \) when assuming the upper bound of \( C \) (fig. 4). These nonsignificant values can be attributed to the fact that \( Z_{SS} \) has a lower statistical power than under a more severe bottleneck scenario (Depaulis et al. 2003). Indeed, none of these simulated samples had a major haplotype class compatible with the observed haplotype structure \( (P < 0.05, \text{ comparing the observed and simulated number of haplotypes, normalized by the number of segregating sites } \text{[Przeworski 2002])}. \)

The power of coalescent-based tests is not independent of the mutation population parameter \( \theta \) and, therefore, the statistical power might vary as a function of the true value of \( \theta \) for a given sample size and \( S \) (Markovtsova et al. 2001), we performed a prospective analysis to test this effect on our data. However, conditioning coalescent simulations on the estimated value of \( \theta \) instead of on \( S \) did not have a substantial effect on the results. Indeed, similar or more extreme \( P \)-values were obtained when simulations were conditioned on \( \theta \). This observation

---

**FIG. 1.—Demographic models.** Left panel: Population subdivision. Right panel: Population bottleneck. See text for symbols and details.
is consistent with computer simulations indicating that the difference in the rejection probability due to using $S$ instead of $\theta$ in coalescent-based tests is substantially reduced in regions with moderate or high rates of recombination, as those studied here (SE Ramos-Onsins, unpublished data).

Tests using specific demographic null models under recombination are emerging as an alternative to null stationary panmictic models (Glinka et al. 2003; Orengo and Aguadé 2004; Bauer DuMont and Aquadro 2005; Haddrill et al. 2005; Ometto et al. 2005; Wright and Gaut 2005; Beisswanger et al. 2006; Pool et al. 2006). The models used in this study, like in other studies, are simple but likely are the most relevant for African $D. simulans$ populations. The present analysis using a relatively large range of parameter values allows us to conclude that demography alone is unlikely to account for the observed haplotype structure. By contrast, previous experimental studies for this genomic region reveal a pattern that corresponds very well with the outline predicted under an incomplete selective sweep: the strong haplotype structure dissipates with distance to the core region, polymorphism decreases in the structured...
domain, and there is an increased frequency of derived variants (Parsch et al. 2001; Rozas et al. 2001; Quesada et al. 2003; Meiklejohn et al. 2004). The new evidence provided here indicates that a major role of demographic effects may be disregarded, thus supporting the previous claim that the data are better described by a model in which directional selection has acted recently on this region (Quesada et al. 2003; Meiklejohn et al. 2004).

Acknowledgments

This research has been partially performed using the Centre de Supercomputació de Catalunya facilities. This work was supported by grants BMC2001-2909 and BFU2004-02253 from Comisión Interdepartamental de Ciencia y Tecnología, Spain; 2001SGR-00101 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Catalonia, Spain; and by special support (Distinció per la Promoció de la Recerca Universitària) from Generalitat de Catalunya to M.A.

Literature Cited


Fig. 3.—Admixture test results (II). Observed and expected $Z_{nS}$ values for each region surveyed. Observed and expected $Z_{nS}$ values are depicted as closed and open circles, respectively. Expected $Z_{nS}$ values for each region were obtained by neutral coalescent simulations (1,000 replicates) with recombination ($C_M$) under a simple admixture model with some (lower panel) or no migration (upper panel) allowed and considering unequal subpopulation sizes. $t_2$: admixture time in $4N$ generations.

FIG. 3.—Admixture test results (II). Observed and expected $Z_{nS}$ values for each region surveyed. Observed and expected $Z_{nS}$ values are depicted as closed and open circles, respectively. Expected $Z_{nS}$ values for each region were obtained by neutral coalescent simulations (1,000 replicates) with recombination ($C_M$) under a simple admixture model with some (lower panel) or no migration (upper panel) allowed and considering unequal subpopulation sizes. $t_2$: admixture time in $4N$ generations.

Fig. 4.—Bottleneck test results. The multilocus $\Psi$ statistic and its empirical distribution was obtained as indicated in the text, assuming different recombination rates ($C_L$, $C_M$, $C_U$) and bottleneck severities ($S_b$). The duration of the bottleneck in $4N$ generations is fixed to 0.0005. The conservative criterion of equal sizes for the ancestral and derived populations is used here. The $P$-values were obtained as in the admixture model. Times are calculated assuming $N = 2 \times 10^6$ and 10 generations per year. Open symbols: $P < 0.05$. Closed symbols: $P > 0.05$.  

Fig. 4.—Bottleneck test results. The multilocus $\Psi$ statistic and its empirical distribution was obtained as indicated in the text, assuming different recombination rates ($C_L$, $C_M$, $C_U$) and bottleneck severities ($S_b$). The duration of the bottleneck in $4N$ generations is fixed to 0.0005. The conservative criterion of equal sizes for the ancestral and derived populations is used here. The $P$-values were obtained as in the admixture model. Times are calculated assuming $N = 2 \times 10^6$ and 10 generations per year. Open symbols: $P < 0.05$. Closed symbols: $P > 0.05$.  

Fig. 4.—Bottleneck test results. The multilocus $\Psi$ statistic and its empirical distribution was obtained as indicated in the text, assuming different recombination rates ($C_L$, $C_M$, $C_U$) and bottleneck severities ($S_b$). The duration of the bottleneck in $4N$ generations is fixed to 0.0005. The conservative criterion of equal sizes for the ancestral and derived populations is used here. The $P$-values were obtained as in the admixture model. Times are calculated assuming $N = 2 \times 10^6$ and 10 generations per year. Open symbols: $P < 0.05$. Closed symbols: $P > 0.05$.


Diethard Tautz, Associate Editor

Accepted June 1, 2006