doi:10.1111/j.1420-9101.2007.01490.x

Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck

M. I. TENAILLON,* F. AUSTERLITZ† & O. TENAILLON‡§

*UMR8120 de Génétique Végétale, INRA/Univ. Paris-Sud/CNRS/AgroParisTech, Ferme du Moulon, Gif-sur-Yvette, France
†Laboratoire Ecologie Systématique et Evolution, UMR 8079, CNRS/Univ. Paris-Sud/AgroParisTech, Orsay, France
‡Institut National de la Santé et de la Recherche Médicale (INSERM) U722, Faculté de Médecine Xavier Bichat, Paris, France
§Laboratoire Ecologie et Evolution des Microorganismes, Université Denis Diderot-Paris VII, Paris, France

Keywords:

coalescence;
demography;
polymorphism aggregation;
recombination.

Abstract

Genome wide patterns of nucleotide diversity and recombination reveal considerable variation including hotspots. Some studies suggest that these patterns are primarily dictated by individual locus history related at a broader scale to the population demographic history. Because bottlenecks have occurred in the history of numerous species, we undertook a simulation approach to investigate their impact on the patterns of aggregation of polymorphic sites and linkage disequilibrium (LD). We developed a new index (Polymorphism Aggregation Index) to characterize this aggregation and showed that variation in the density of polymorphic sites results from an interplay between the bottleneck scenario and the recombination rate. Under particular conditions, aggregation is maximized and apparent mutation hotspots resulting in a 50-fold increase in polymorphic sites density can occur. In similar conditions, long distance LD can be detected.

Introduction

Mutational hotspots and recombination hotspots are common features of diversity and recombination patterns. When averaged across 200-kb windows, rates of heterozygosity in the human genome show a 10-fold variation (Sachidanandam *et al.*, 2001). In yeast, Fay & Benavides (2005) also reported the hyper variability in noncoding regions at two loci, namely MLS1 and PDR10. Similarly, heterogeneity of recombination rates has been described in many model organisms such as yeast (Petes, 2001), maize (Dooner & Martinez-Ferez, 1997; Fu *et al.*, 2002) and mammals. Variation in mutation and recombination rate are strongly correlated in fruit flies, humans and plants (Przeworski *et al.*, 2000; Andolfatto & Przeworski, 2001; Aquadro *et al.*, 2001; Baudry *et al.*, 2001; Nachman, 2001; Tenaillon *et al.*, 2002). This pattern holds in noncoding regions suggesting that recombination itself is an important mutagenic agent (Lercher & Hurst, 2002). Consistently, Hellmann *et al.* (2003) demonstrate that regions harbouring less recombination also exhibit a reduced level of divergence between closely related species arguing for a neutral explanation for the link between diversity and recombination. In other words, variations in nucleotide diversity may reflect variations of mutation rate caused by underlying patterns of variable recombination. However, the association between recombination and diversity may not result from the mutagenic effect of recombination but rather may be mediated by a third factor that influence both recombination and diversity, such as natural selection (Nachman, 2001) which seems widely acting even in noncoding region (Andolfatto, 2005; Haag-Liautard *et al.*, 2007).

Interestingly, Reich *et al.* (2002) shed a new light into this debate by demonstrating that genome wide patterns of single nucleotide polymorphism variability in humans are primarily determined by differences in gene history as opposed to differences in local mutation rate. They suggested that variation of the mutation rate along the sequence is not the primary cause of mutational hotspots but rather that the past history of DNA sequences is the key factor for understanding local variation in mutation rate. Wang *et al.* (2002) provided evidences that differences in haplotype block lengths can arise in the absence of variation of the underlying recombination rate and also as a result of variation in population sizes. Altogether

Correspondence: Maud Tenaillon, INRA Univ. Paris-Sud/CNRS/AgroParisTech, UMR8120, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France.
Tel.: 33 (0) 1 69 33 23 34; fax: 33 (0) 1 69 33 23 80;
e-mail: tenaillon@moulon.inra.fr

these studies call for a deeper investigation of the role of evolutionary history in shaping both patterns of diversity and recombination.

The history of a sample of sequences can be simulated backward in time up to their most recent common ancestor (TMRCA) using the coalescent framework. This framework is particularly useful for simulating population genetics data. In contrast to forward simulations, it allows decoupling the neutral drift process according to which genealogies of a given sample of sequences are built, from the mutational process. Mutations are placed on each branch of the genealogies according to their length; long branches receive on average more mutations than shorter ones. Coalescent theory has been broadly used to help elucidate the history of numerous species including humans, drosophila and plants (Adams & Hudson, 2004; Voight *et al.*, 2005; Wright *et al.*, 2005; Bachtrog & Andolfatto, 2006; Thornton & Andolfatto, 2006). Under the neutral coalescent process with recombination, a given DNA sequence is a mosaic of segments, each segment being defined as a nonrecombining fragment, and crossing-overs define the borders between adjacent segments (Hudson, 1983; Hudson & Kaplan, 1985). Some of these segments will coalesce far back in time and hence have a large TMRCA while others will coalesce more recently and have a small TMRCA. In other words, a DNA sequence can be pictured as a mosaic of short segments bearing a high number of mutations interspersed by long segments bearing few mutations. A clustering of mutations and hence a mutational hotspot can therefore be obtained by simulating a neutral coalescent process with recombination, without involving heterogeneity in mutation rate along a sequence (Reich *et al.*, 2002).

Many species have experienced one or several population bottlenecks in their evolutionary history. In humans, for instance, most studies agree on an origin of modern humans around 200 000 years ago in Africa through a bottleneck (Harding & McVean, 2004). Demographic scenarios, although still highly debated, involve subsequent population bottlenecks accompanying the human migration out of Africa, 30 000–87 500 years ago, in Asian and European populations (Marth *et al.*, 2003; Adams & Hudson, 2004). Similarly, the pattern of linkage disequilibrium (LD) in North American population of *Drosophila simulans* is compatible with a recent and severe bottleneck (Wall *et al.*, 2002). In *Drosophila melanogaster*, nucleotide variation in non-African populations is consistent with a recent bottleneck accompanying its dispersal out of Africa (Thornton & Andolfatto, 2006). Domesticated species have also been through bottlenecks. For example, in an extensive survey in maize, Wright *et al.* (2005) define as a most probable scenario a simple bottleneck occurring 7500 years ago with a strength measured by the ratio of the bottleneck duration over the population size during the bottleneck of 2.45.

In the present article, we are interested in understanding the extent to which demography, modelled as a simple bottleneck, can affect the spatial variation in coalescent times along a sequence and, as a consequence, the observed pattern of diversity and LD. In particular we focus on the aggregation of polymorphism, i.e. the spatial distribution of polymorphic sites along the sequence, and explore in which conditions bottlenecks affect long distance LD. We performed coalescent simulations using a range of bottleneck scenarios and introduced a statistical index [Polymorphism Aggregation Index (PAI)] to characterize the resulting level of aggregation. We showed that, for a given recombination rate, aggregation is the result of interplay between the recombination and the demographic scenarios. We define the conditions that maximized this aggregation and show that long distance LD is likely to occur in these conditions.

Methods

Model description

Coalescent simulations were used to model the impact of a bottleneck on sequence diversity and recombination as described in Tenaillon *et al.* (2004). Going backward in time, the model includes an instantaneous reduction of the present population size N_p at time t into a bottleneck population size, N_b, and an instantaneous increase of population size from N_b to the ancestral population size, N_a. The bottleneck itself was characterized by its duration, d, expressed in generations and by N_b. We set $N_a = N_p = 10^6$ and $N_b = 10^4$, so that population size was reduced by 100-fold during the bottleneck phase. Following this model, we simulated the evolution of 20 sequences (or haplotypes) subjected to intragenic recombination and mutation. We varied d by taking 10 values from 10^2 to 10^5 generations, c, the intragenic recombination rate per site per generation using six values from 0 to $10^{-7}$and t, using 15 values from 10^3 to 10^6 generations. Values for each parameter were equally distributed on a log scale, their range encompassing empirical values described in the literature, for instance for domesticated species, humans and *Drosophila*. This lead to a total of 900 scenarios explored. For the sake of comparison, we performed also a set of replicates without bottleneck. As an additional control, we also performed simulations under a simple population expansion. We modelled, going backward on time, an exponential size reduction from N_p to N_a starting at generation t and lasting d generations. We used the same range of parameters values for d, c and t, as in the bottleneck model.

The program from Tenaillon *et al.* (2004) was slightly modified in order to condition on the final number of polymorphic sites, S, rather than on the mutation rate, μ. This was achieved by determining the length (L) of the sequence to simulate under each scenario to obtain on average 100 polymorphic sites. The average length of the

resulting coalescent trees (for a given tree the length is determined by the sum of all branch lengths) obtained for each scenario was first estimated on 1000 replicates. The length of the sequence to simulate was then determined as:

$$L = \frac{S}{\text{average length} * \mu}.$$

The mutation rate, μ, was set to 10^{-9} mutation/site/generation and the number of polymorphic sites, S, to 100.

Polymorphism Aggregation Index

To quantify the aggregation of polymorphic sites along a sequence, we developed a statistical index, which we called the PAI. This index is based on the broken stick model, which was previously used by Goss & Lewontin (1996) as a method to test the occurrence of clustering. It focuses on the distance in base pairs between adjacent polymorphic sites. If the polymorphic sites were distributed at random along the sequence, the distribution of this distance would follow a broken stick model, in which a fixed quantity is randomly divided into a number of fragments. The mean and the variance of such a distribution are derived in Karlin & Taylor (1981):

$$\text{mean} = \frac{L}{(S+1)}, \text{variance} = \frac{2L^2}{(S+1)(S+2)} - \frac{L^2}{(S+1)^2}.$$

And therefore the coefficient of variation equals:

$$\text{CV} = \sqrt{\frac{S}{S+2}}.$$

We defined the PAI as:

$$\text{PAI} = \frac{\text{observed standard deviation}}{\text{observed mean}} / \sqrt{S/(S+2)},$$

so that under a null hypothesis of random distribution of polymorphic sites, the expectation of PAI equals 1. A value of PAI above 1 indicates that polymorphic sites are more aggregated than expected under the null hypothesis; conversely a PAI value below 1 indicates that polymorphic sites are more uniformly distributed than expected. We evaluated the relationship between PAI and S (the number of polymorphic sites) by computing the correlation between both variables.

Linkage disequilibrium descriptors

We studied the impact of a bottleneck on the patterns of LD. LD was measured by r^2 (Hill & Robertson, 1968) between all pairs of informative polymorphic sites with frequency > 5%. Subsequently, we plotted r^2-values against the base-pair distance between sites. We fitted by least-squares estimation using a fitting recursion algorithm (Press *et al.*, 1992) the observed values to the expectation of r^2 (Hill & Weir, 1994; Hudson, 2003):

$$\text{E}(r^2) = \frac{1}{1 + (4Nc * (\text{dist}))} + \frac{1}{n},$$

where N is the population size (N_p), c is the per site recombination rate, dist the distance between sites, and n (=20) the sample size. This fit provided with an estimate of $4Nc$. The ratio of the χ^2 statistic over the number of pairwise comparisons was used to measure the goodness of fit of the observed LD decay to the expectation of r^2. An elevated ratio indicates a poor fit to the theoretical expectation. Given the impact of the number of polymorphic sites on the χ^2 statistic, conditioning on the number of polymorphic site rather than on the mutation rate was particularly important. It allowed us to compare contrasted scenarios leading however to a similar number of polymorphic sites and therefore to a similar number of pairwise comparisons.

Simulations outputs

We performed 1000 replicates for each of the 900 bottleneck scenarios. Each replicate resulted in a number of coalescent trees equal to the number of segments generated through recombination events. Each tree was characterized by its length (mean and variance of the branch lengths) and its coalescent time in units of generations (mean and variance of the branch coalescent times). To summarize the information contained in the resulting trees for a given simulation, we averaged among all segments the mean and variance of the total tree length and TMRCA. Going backward in time, we also determined the average number of segments for which the TMRCA predated the bottleneck, i.e. segments that 'survived' the bottleneck as opposed to the segments that coalesced during or before the bottleneck phase. Similarly, we estimated the average number of haplotypes per segment among the 20 initial haplotypes that 'survived' the bottleneck. These descriptive statistics as well as PAI values were further averaged over 1000 replicates for each scenario. Concerning the expansion scenarios, we determined the average PAI value calculated among 100 replicates for the 900 scenarios explored.

Because the simulations were highly time consuming, we were able to study the decay of LD over distance as well as the fit to the r^2 expectation only for 10 of the 1000 replicates for each parameter set. In fact, the decay of LD required the determination of r^2 values of on average 4950 pairs of informative sites and the fitting recursion was computationally very intensive.

Results

Aggregation results from an interplay between recombination and the bottleneck severity

To study the conditions in which aggregation occurred, we computed the PAI. PAI values were averaged across

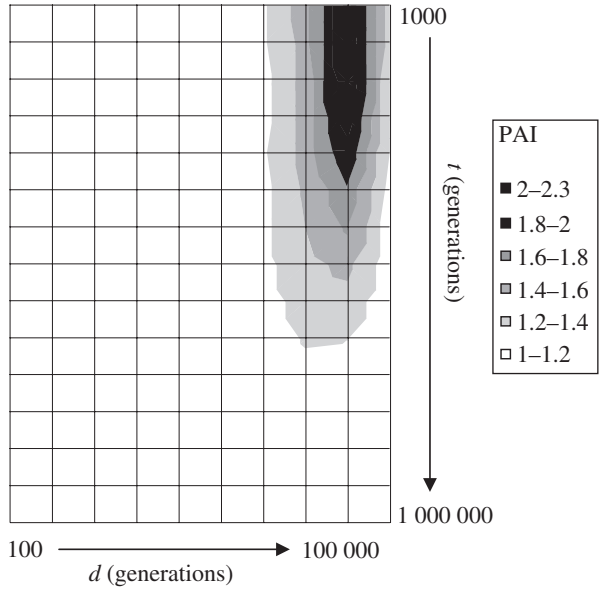

Fig. 1 Polymorphism Aggregation Index (PAI) as a function of bottleneck duration (d) and time to the bottleneck (t). $c = 10^{-8}$. PAI reaches its maximum value for a long (d = 46 415 generations) and recent (t = 1637 generations) bottleneck.

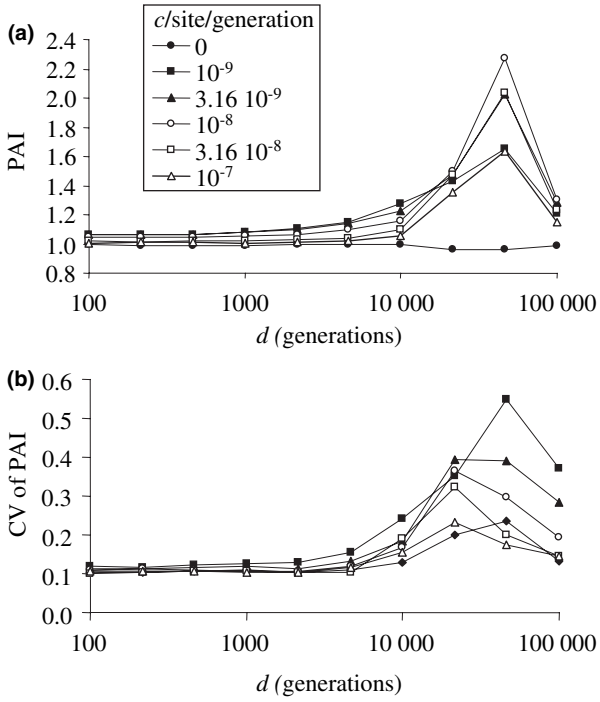

Fig. 2 Polymorphism Aggregation Index (PAI) (a) and its corresponding coefficient of variation (b) as a function of bottleneck duration (d) and recombination (c). t = 1637 generations. For a given bottleneck scenario, aggregation is maximized for an intermediate values of recombination rate.

1000 replicates for each bottleneck scenario and a given recombination rate value. Figure 1 presents the resulting averaged PAI values as a function of bottleneck duration (d) and time to the bottleneck (t) for $c = 10^{-8}$ (Fig. 1). Our results show that PAI values are close to 1, i.e. the value expected for a random distribution of polymorphic sites, in a wide range of demographic scenarios. More importantly, they show that recent (from 1000 to 19 307) and long bottlenecks (d = 46 415 generations) are favourable to a highly aggregated distribution of polymorphic sites (PAI values above 1.6). Among these 'aggregation' scenarios, PAI reaches a maximum value for bottleneck duration of 46 415 generations and a time to the bottleneck of 1637 generations (called thereafter the 'highest aggregation scenario'). This observed pattern holds only when recombination was not null as shown in Fig. 2a. In fact, considering a given bottleneck scenario, there is an intermediate value of recombination, $c = 10^{-8}$, that maximizes the aggregation (Fig. 2a). As shown in Fig. 2b, the coefficient of variation of the PAI increased as the mean PAI value increased. However, the coefficient of variation remains below 1 and therefore its level of variation did not affect our interpretations. In contrast to bottleneck scenarios, none of the expansion scenarios exhibited an elevated PAI. Average PAI values ranged from 0.97 to 1.06 across the 900 scenarios. Hence, elevated PAI values resulted from an interplay between recombination and the bottleneck scenario.

To further study the coupled effect of recombination and bottleneck on PAI, we plotted the distributions of PAI values (Fig. 3) for the 'highest aggregation scenario', a bottleneck scenario with the same conditions (d = 46 415 and t = 1637) without recombination and a 'nonbottleneck scenario' with the same recombination rate ($c = 10^{-8}$). As expected, the null distribution centred

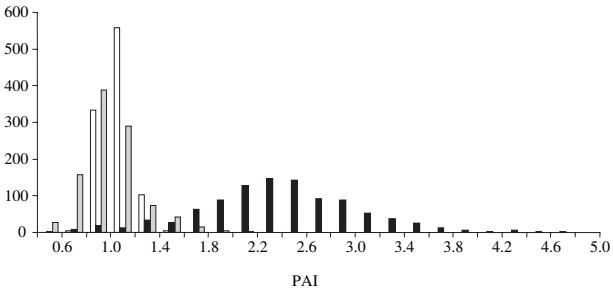

Fig. 3 Distribution of Polymorphism Aggregation Index (PAI) values under a scenario without bottleneck $c = 10^{-8}$ (open bars), a scenario with a bottleneck, d = 46 415, t = 1637, $c = 10^{-8}$ (black bars) and a scenario with a bottleneck (d = 46 415, t = 1637) but no recombination (grey bars). The null distribution centred on 1 is obtained in the absence of recombination. A slight increase in aggregation is observed with recombination even in the absence of demography. In the presence of recombination, the bottleneck strongly increases the average PAI value as well as its variance.

on an average PAI value of 0.99 (standard deviation 0.11) was obtained without recombination. A slight increase in aggregation was observed with recombination even in the absence of demography (average PAI value of 1.05, standard deviation 0.24). Finally, under the 'highest aggregation scenario', the PAI distribution was strongly skewed towards high PAI values (average PAI value of 2.37) and its variance was also considerably larger.

To obtain a better qualitative description of our data, we further investigated the consequences of the bottleneck on the heterogeneity of polymorphism along simulated sequences. We simulated sequences containing on average 500 polymorphic sites under two contrasted scenarios: the 'highest aggregation scenario' and the 'nonbottleneck scenario' with the same recombination rate (10^{-8}). Because we conditioned the simulations on an expected final number of polymorphic sites of 500, the length of the simulated sequences was 57 334 bp for the 'highest aggregation scenario' and 5733 bp for the 'nonbottleneck scenario'. For each scenario, we estimated the variation of the number of polymorphic sites along the simulated sequences by counting the number of polymorphic sites contained in nonoverlapping windows of 1000 and 100 bp for the 'highest aggregation scenario' and the 'nonbottleneck' scenario respectively. Considering the differences in sequence length between the two scenarios, the chosen window sizes allowed to divide the sequences in an equal number of windows.

Patterns of aggregation in these two scenarios were qualitatively very different: the variation in the density in polymorphic sites along the simulated sequence was much lower in the 'nonbottleneck scenario' (Fig. 4a–c) as compare to the 'highest aggregation scenario' (Fig. 4d–f). In this last scenario, the number of polymorphic sites ranged from 0 to 67, with a lower bound always at 0 (i.e. all replicates contained a window without polymorphic sites) and an average upper bound of 59.6 across 10 replicates. The corresponding average lower and upper bounds in the 'nonbottleneck scenario' were respectively 1.5 and 21.9. For each simulation, we measured the increase in the density of polymorphic sites as the ratio between the upper bound (i.e. maximum number of polymorphic sites observed in a 1 kb window) and the median number of polymorphic sites along the sequence. In the 'highest aggregation scenario', the average increase estimated among 10 replicates was around 50. While in the 'nonbottleneck scenario', the corresponding average increase was around 2.5-fold (maximum increase, 3.625).

Primary causes of aggregation

A potential problem with the use of PAI could arise if it happens to be positively correlated with the number of polymorphic sites, i.e. if regions with fewer polymorphic sites aggregate significantly less than regions with higher number of polymorphic sites. To verify that PAI was independent from the level of variation, we estimated the correlation between PAI and the number of polymorphic sites in nonbottleneck scenarios. We found no correlation between the two parameters for various levels of recombination (data not shown).

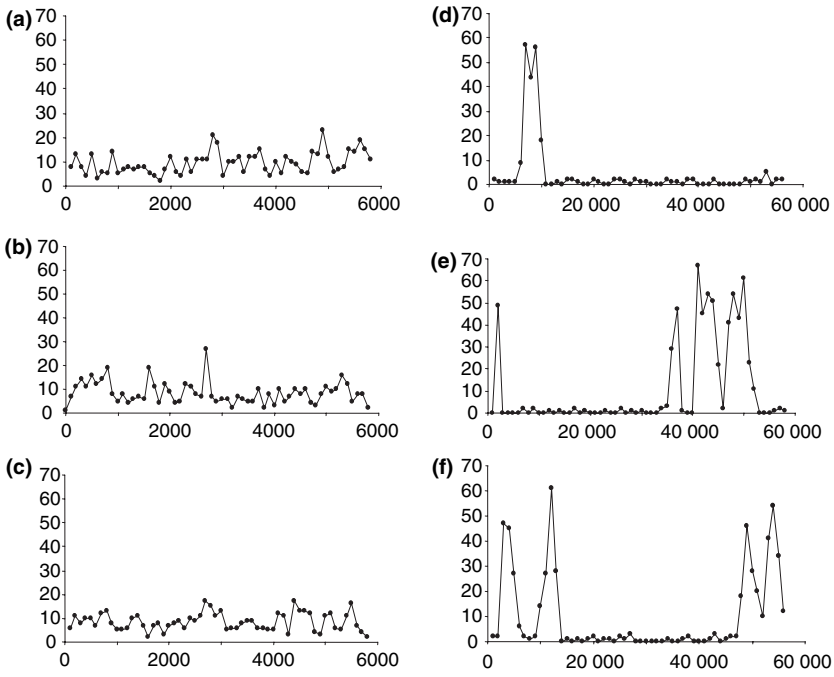

Fig. 4 Frequency distribution of the number of polymorphic sites along a simulated sequence. (a–c) Three simulations under a scenario without bottleneck $c = 10^{-8}$. (d–f) Three simulations under a scenario with a bottleneck ($d = 46\,415$, $t = 1637$, $c = 10^{-8}$).

We estimated several descriptive statistics from the coalescent trees to determine which factor explained the largest part of the variation in polymorphism aggregation. These analyses considered correlations between descriptive statistics and PAI for all scenarios except those involving a null recombination. We considered first the average number of haplotypes per segment among the initial 20 haplotypes that survived the bottleneck. In simulations for which the aggregation was low (around 1), the number of surviving haplotypes ranged between 0.014 and 19.06 depending on the bottleneck severity. More intense bottlenecks tended to reduce the number of surviving haplotypes. For higher aggregation values (> 1.5), the haplotype number was comprised between 0.937 and 2.3. Overall, the number of surviving haplotypes explained poorly the aggregation (data not shown).

In contrast, we obtained a positive correlation as estimated by the Pearson correlation coefficient (r) between PAI values and both the average variance in tree time and tree length (r = 0.71 and 0.77 respectively). A linear regression model was fitted to the data and the resulting R^2 values, 0.51 and 0.59 respectively, indicated a good fit to the data. Finally, we measured the square of the coefficient of variation in tree length and correlated it to the PAI values. This correlation was the strongest of all (r = 0.84). These results are in accordance with the expectation of the bottleneck causing an increase in the variance of tree length and TMRCA among segments. The resulting sequences are composed of a mosaic of both: short segments with a 'long' history (old TMRCA), therefore bearing a high number of mutations, and long segments with a recent history (recent TMRCA) bearing few mutations. This pattern enhances the aggregation of polymorphic sites as measured by PAI.

Impact of bottleneck on recombination estimates and LD patterns

We studied the impact of the bottleneck on LD patterns by measuring the deviation of fit of the observed pairwise LD measure, r^2, as a function of distance to the theoretical prediction of LD decay over distance. Figure 5 presents the deviation of fit as measured by the ratio of the χ^2 statistic over the number of pairwise comparisons as a function of bottleneck duration (d) and time to the bottleneck (t) for $c = 10^{-8}$. For any given value of recombination rate, the fit was fairly good for a wide range of demographic scenario. However, for values of recombination exceeding 10^{-9}, and for fairly recent (t comprised between 1000 and 7196 generations) and long bottlenecks (d from 4641 to 21 544 generations) the deviation from the model increased, reaching a maximum value of 0.108 for a bottleneck duration of 10 000 generations and a time to the bottleneck of 1000 generations (Fig. 5). Elevated χ^2 values were obtained when long distance LD was observed, i.e. when a number of pairwise comparisons resulted in r^2 values of

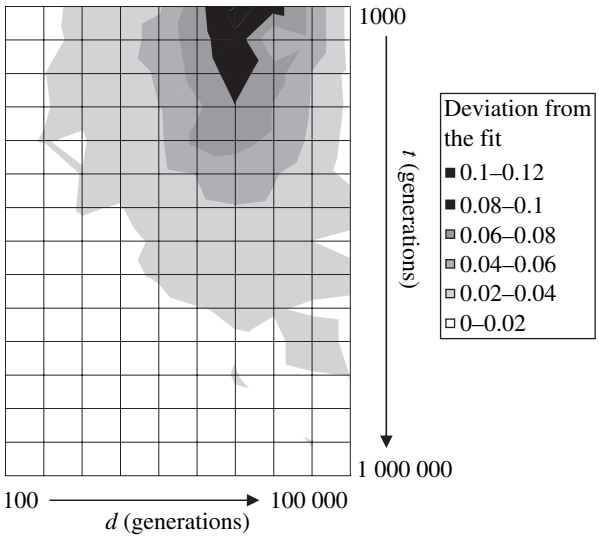

Fig. 5 Deviation from the fit to the theoretical linkage disequilibrium decay over distance as measured by χ^2 values averaged among the 10 first replicates as a function of bottleneck duration (d) and time to the bottleneck (t). $c = 10^{-8}$. Elevated χ^2 values indicate a poor fit of the simulated data to the theoretical curve and are observed under scenarios involving long and recent bottlenecks.

1 (Fig. 6). Hence, we suggest the idea that past population history could affect patterns of LD through the generation of long distance LD.

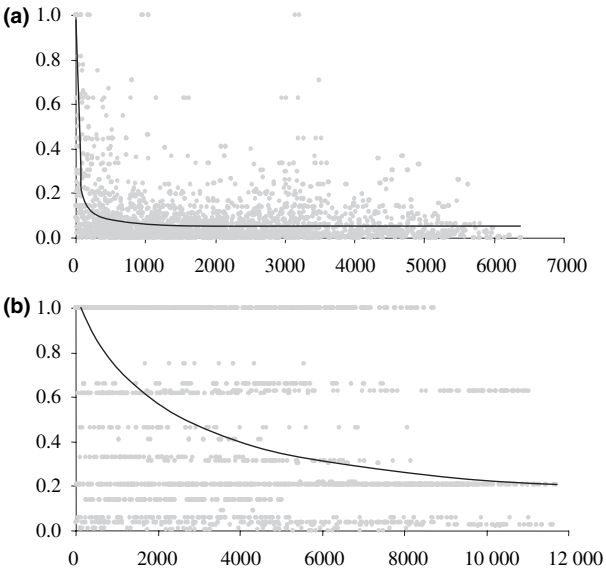

Fig. 6 Linkage disequilibrium decay over distance as measured by r^2 values calculated among all pairs of informative sites (Y-axis) plotted against the distance in bp. (a) Scenario without bottleneck ($c = 10^{-8}$); (b) scenario with a bottleneck (d = 10 000, t = 1000, $c = 10^{-8}$). Data points (in grey) are fitted to the expectation of r^2 (black line).

Discussion

Patterns of polymorphism are shaped by the history of populations. Consequently, one can use summary statistics on polymorphism data to infer some elements of this history. Thus, a growing number of studies use allelic frequency spectrum statistics such as Tajima's D (Tajima, 1989), Fay and Wu H (Fay & Wu, 2000), Fu and Li D (Fu & Li, 1993) or statistics based on LD such as the haplotype diversity and number (Fu, 1997), to infer the existence of population bottlenecks or expansions. The distribution of polymorphic sites positions along a sequence has not been investigated along those lines. Coalescent theory, however, predicts that a neutral coalescent with recombination gives rise to aggregation patterns that deviate from a random distribution of polymorphic sites. To characterize the aggregation, we developed a new index, the PAI. The calculation of PAI is straightforward, and, more importantly, it is independent of the number of polymorphic sites. PAI can therefore be used among other summary statistics to describe the patterns of polymorphism. A simple software, available upon request, has therefore been developed to allow users to both calculate PAI on a sequence alignment and to compare the observed value to the simulated distribution of PAI values under a neutral coalesent process with recombination. In the present article, we used it to study more specifically the effect of a bottleneck on aggregation patterns.

Aggregation of polymorphisms along a sequence is classically interpreted as the result of selective constraints acting at the gene level (i.e. conserved exons vs. variable noncoding regions) or as the result of variation in mutation rate. Looking at noncoding DNA, several studies have concluded that variation in nucleotide diversity results in part from local sequence properties associated with base composition: CpG dinucleotides, GC content, presence of simple repeats, poly(A/T) and poly(R/Y) (Wolfe *et al.*, 1989; Hwang & Green, 2004; Hellmann *et al.*, 2005). In addition phenomena including selection, gene density (Payseur & Nachman, 2002) and gene expression contribute to these patterns of variation (Surralles *et al.*, 2002; Hoede *et al.*, 2006). Finally, variation in recombination rate has also been proposed to explain much of this variation through a mutagenic effect of recombination (Lercher & Hurst, 2002). However, as common local or genomic factors may affect both mutation and recombination, whether this association results from a direct causal effect is still debated.

Using a simulation approach in which all these effects are excluded (i.e. evolution of neutral sequences with homogenous mutation and recombination rates), we showed that, in the presence of recombination, bottlenecks could promote a strong aggregation signal leading to apparent mutation hotspots. The intensity of these hotspots can be as dramatic as a 50-fold increase in the number of polymorphic sites per kilobase, as compared to a 2.5-fold increase in the absence of bottleneck. Striking differences in patterns emerge as shown in Fig. 4. By splitting up the sequence in small fragments with decoupled history, recombination allows the coexistence along the sequence of long-history segments that survived the bottleneck with short-history segments that coalesced during the bottleneck. The former segments accumulate many mutations while the latter do not. Therefore, aggregation results from interplay between demography and recombination. This effect is maximized for long and recent bottlenecks (Fig. 1) and intermediate recombination rates (Fig. 2). On one hand, long and recent bottlenecks favour the complete coalescence of most lineages during the bottleneck phase while they allow the survival of few lineages coalescing far back in time. These conditions maximize the variance in coalescent times and tree lengths. On the other hand, recombination modulates the size of the segments. In the absence of recombination, TMRCA of the whole sequence will either occur during or after the bottleneck: all nucleotides of the sequence share a common history and no aggregation is observed (Fig. 3). Elevated recombination rate will ultimately split up the sequence in single nucleotides (segments) and lead to a complete decoupling of the history of two adjacent sites and hence an absence of aggregation pattern. Finally, intermediate levels of recombination allows for the existence at the time of the bottleneck of long segments coalescing during the bottleneck and long segments surviving the bottlenecks (next split up in smaller segments before the TMRCA). The contrast between the number of polymorphic sites in each type of fragments generates a strong aggregation signal (Fig. 3).

Just like contrasted genealogy lengths between segments along the sequence affect polymorphism patterns, they also affect underlying LD patterns. Previous studies have described the effect of bottlenecks on patterns of LD by using an estimate of the effective population size (N), based on both LD ($4Nc$) and an estimate of c (recombination rate) for each locus. Low estimates of the effective population size indicate high levels of LD and vice versa. Exploring a few bottleneck conditions, Wall *et al.* (2002) showed that recent and severe bottlenecks cause a genome-wide increase in LD. Scaling $4Nc$ by $4N\mu$, estimated from observed levels of variability and an estimate of the mutation rate (μ), provides with a measure of the extent to which LD levels in a population conform to the predictions of the standard neutral model (Andolfatto & Przeworski, 2000). Using this approach, Haddrill *et al.* (2005) showed that recent bottlenecks increase LD more strongly than they reduce levels of variability, resulting in depleted $4Nc/4N\mu$ ratios that could account for patterns observed in non-African *Drosophila melanogaster* populations. A complementary approach to depict LD levels is to study the decay of LD over distance. In humans, Frisse *et al.* (2001) studied the decay of LD within and between 1-kb fragments

separated by ~8 kb and showed that LD declines with distance at a rate roughly four times faster in the Hausa than in the Italian or the Chinese sample, consistent with the idea that non-African populations went through a recent bottleneck. A similar pattern was observed in maize where the rate of LD decay in inbred lines is slower than in a combined sample of inbred lines and landraces (Tenaillon *et al.*, 2001), the latter one being subjected to a less severe (recent) bottleneck (Yamasaki *et al.*, 2005). Although reporting statistic on LD decay (such as the average r^2 value among pairs for a given distance between sites) brings valuable information, most of these studies also miss some information on the qualitative LD patterns. For instance, long distance LD ($r^2 = 1$) is often observed but not commented (see Fig. 1 in Frisse *et al.*, 2001). Exploring the homogeneity of the LD decay under a wide range of bottleneck scenarios, we emphasize how and when do bottlenecks affect LD and in particular long distance LD.

Interestingly, we obtained a poor fit of the LD decay over distance for bottleneck conditions quite similar to the ones maximizing aggregation. In other words, long and recent bottlenecks tend to favour the occurrence of long distance LD (Figs 5 and 6). This may be due to the simultaneous coalescence of noncontiguous segments during the bottleneck. Hence, correlations in coalescent times are increased by coalescence during the bottleneck, which may generate long distance LD (McVean, 2002). In contrast, some DNA stretches will survive the bottleneck and thus be characterized by long genealogies allowing more opportunities to both mutate and recombine (Reich *et al.*, 2002) leading to numerous small segments bearing a high number of polymorphic sites. One would therefore expect to find recombination hotspots around the apparent mutational hotspots generated by bottlenecks. Indeed, looking at the density of segments along the simulated sequence, we found a strong correlation between the density in polymorphic sites and the number of segments created by recombination (data not shown). However, we were not able to detect associated recombination hotspots using PHASE v2.1.1 (Li & Stephens, 2003). In fact in the 'highest aggregation scenario' only very few haplotypes, in most cases only two, survived the bottleneck which makes it impossible to detect recombination. In addition, when relaxing the aggregation pattern by simulating weaker bottlenecks, a higher number of haplotypes survived but we were limited in our power to detect recombination by the restricted number of simulated sequences. Further investigations would be necessary to test for a positive correlation between diversity and recombination along the genome under particular demographic scenarios.

To conclude, our results suggest that apparent mutational hotspots can arise in the absence of underlying sequence properties or genomic features but simply as a result of a past bottleneck. Although this effect is maximum for recent and long bottlenecks characterized by a size reduction of a 100-fold, for long bottleneck durations the effect is significant in a broad range of time to the bottleneck (Fig. 1). Because bottlenecks can affect significantly patterns of nucleotide variation within a species while population expansions do not, our measure of aggregation, PAI, could be used among other summary statistics to discriminate between contrasted demographic scenarios. In addition, genome scans will help identify the determinants of polymorphism aggregation, whole genome increase of PAI would favour a bottleneck explanation while local increase of PAI would favour alternative explanations such as selection or population admixture that could also modulate the aggregation patterns. For instance, PAI could help detect the loss of diversity associated with strong selective sweeps. The target (selected) region would appear as a cold spot of mutation, which, in turn, would inflate the PAI for the genomic region surrounding it. Selective sweeps of small effect could also generate aggregation within the target region as the selective process mimics a local bottleneck. Finally, because population admixture brings highly divergent segments into the recipient population, it should create an aggregation pattern detectable using PAI as introgressed segments will be characterized by a high apparent mutation rate.

Acknowledgments

This study was supported in part by the Agence National de la Recherche: ANR-05-JCJC-0067-01 to MIT. We thank E. Della-Chiesa for helpful discussion and Jean-Luc Jannink for useful comments on the manuscript as well as two anonymous reviewers.

References

Adams, A.M. & Hudson, R.R. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**: 1699–1712.

Andolfatto, P. 2005. Adaptative evolution of non-coding DNA in *Drosophila*. *Nature* **437**: 1149–1152.

Andolfatto, P. & Przeworski, M. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**: 257–268.

Andolfatto, P. & Przeworski, M. 2001. Regions of lower crossing over harbour more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158**: 657–665.

Aquadro, C.F., Bauer DuMont, V. & Reed, F.A. 2001. Genome-wide variation in the human and fruitfly: a comparison. *Curr. Opin. Genet. Dev.* **11**: 627–634.

Bachtrog, D. & Andolfatto, P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* **174**: 2045–2059.

Baudry, E., Kerdelhue, C., Innan, H. & Stephan, W. 2001. Species and recombination effects on DNA variability in the tomato genus. *Genetics* **158**: 1725–1735.

Dooner, H.K. & Martinez-Ferez, I.M. 1997. Recombination occurs uniformly within the *bronze* gene, a meiotic recom-

bination hotspot in the maize genome. *Plant Cell* **9**: 1633–1646.

Fay, J.C. & Benavides, J.A. 2005. Hypervariable noncoding sequences in *Saccharomyces cerevisiae*. *Genetics* **170**: 1575–1587.

Fay, J.C. & Wu, C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J. & Di Rienzo, A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**: 831–843.

Fu, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915–925.

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

Fu, H., Zheng, Z. & Dooner, H.K. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl Acad. Sci. USA.* **99**: 1082–1087.

Goss, P.J.E. & Lewontin, R.C. 1996. Detecting heterogeneity of substitution along DNA and protein sequences. *Genetics* **143**: 589–602.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Charlesworth, B. & Keightley, P. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**: 82–85.

Haddrill, P.R., Thornton, K.R., 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.

Harding, R.M. & McVean, G. 2004. A structured ancestral population for the evolution of modern humans. *Curr. Opin. Genet. Dev.* **14**: 667–674.

Hellmann, I., Ebersberger, I., Ptak, S.E., Paabo, S. & Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72**: 1527–1535.

Hellmann, I., Prufer, K., Ji, H., Zody, M.C., Paabo, S. & Ptak, S.E. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**: 1222–1231.

Hill, W.G. & Robertson, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**: 226–231.

Hill, W.G. & Weir, B.S. 1994. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.* **54**: 705–714.

Hoede, C., Denamur, E. & Tenaillon, O. 2006. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. *PLoS Genet.* **2**: 1697–1701.

Hudson, R.R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**: 183–201.

Hudson, R.R. 2003. Linkage disequilibrium and recombination. In: *Handbook of Statistical Genetics* (D. J. Balding, M. Bishop & C. Cannings, eds), pp. 662–680. John Wiley & Sons, Sussex.

Hudson, R.R. & Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.

Hwang, D.G. & Green, P. 2004. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA.* **101**: 13994–14001.

Karlin, S. & Taylor, H.M. 1981. *A Second Course in Stochastic Processes*. Academic press, New York.

Lercher, M.J. & Hurst, L.D. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**: 337–340.

Li, N. & Stephens, M. 2003. Modelling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**: 2213–2233.

Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., Kholodov, M., Phan, L., Czabarka, E., Murvai, J., Cutler, D., Wooding, S., Rogers, A., Chakravarti, A., Harpending, H.C., Kwok, P.Y.R., Sherry, S.T. 2003. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl Acad. Sci. USA.* **100**: 376–381.

McVean, G. 2002. A genealogical interpretation of linkage disequilibrium. *Genetics* **162**: 987–991.

Nachman, M.W. 2001. Single nucleotide polymorphisms and recombination in humans. *Trends Genet.* **17**: 481–485.

Payseur, B.A. & Nachman, M.W. 2002. Gene density and human nucleotide polymorphism. *Mol. Biol. Evol.* **19**: 336–340.

Petes, T.D. 2001. Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* **2**: 360–369.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Przeworski, M., Hudson, R.R. & Di Rienzo, A. 2000. Adjusting the focus on human variation. *Trends Genet.* **16**: 296–302.

Reich, D.E., Shaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S. & Altshuler, D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**: 135–142.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J.M., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. & Altshuler, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.

Surralles, J., Ramirez, M.J., Marcos, R., Natarajan, A.T. & Mullenders, L.H. 2002. Clusters of transcription-coupled repair in the human genome. *Proc. Natl Acad. Sci. U.S.A.* **99**: 10572–10574.

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

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. & Gaut, B.S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp mays L.). *Proc. Natl Acad. Sci. U.S.A.* **98**: 9161–9166.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S.M., Doebley, J. & Gaut, B.S. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Genetics* **162**: 1401–1413.

Tenaillon, M.I., U'Ren, J., Tenaillon, O. & Gaut, B.S. 2004. Selection versus demography: a multilocus investigation of

the domestication process in maize. *Mol. Biol. Evol.* **21**: 1214–1225.

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

Voight, B.F., Adams, A.M., Frisse, L.A., Qian, Y., Hudson, R.R. & Di Rienzo, A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. U.S.A.* **102**: 18508–18513.

Wall, J.D., Andolfatto, P. & Przeworski, M. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**: 203–216.

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. & Jin, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination and mutation. *Am. J. Hum. Genet.* **71**: 1227–1234.

Wolfe, K.H., Sharp, P.M. & Li, W.-H. 1989. Mutations differ among regions of the mammalian genome. *Nature* **337**: 283–285.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D. & Gaut, B.S. 2005. The effects of artificial selection on the maize genome. *Science* **308**: 1310–1314.

Yamasaki, M., Tenaillon, M.I., Bi, I.V., Schroeder, S.G., Sanchez-Villeda, H., Doebley, J.F., Gaut, B.S. & McMullen, M.D. 2005. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**: 2859–2872.

Received 11 September 2007; revised 12 November 2007; accepted 23 November 2007