BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 449–455
doi:10.1093/bioinformatics/btq689

Genome analysis

Advance Access publication December 14, 2010

Alignment-free estimation of nucleotide diversity

Bernhard Haubold[1,*], Floyd A. Reed[1] and Peter Pfaffelhuber[2]

[1]Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Biology, Plön and [2]Abteilung für Mathematische Stochastik, Mathematical Institute, Albert-Ludwigs University, Freiburg, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Sequencing capacity is currently growing more rapidly than CPU speed, leading to an analysis bottleneck in many genome projects. Alignment-free sequence analysis methods tend to be more efficient than their alignment-based counterparts. They may, therefore, be important in the long run for keeping sequence analysis abreast with sequencing.

Results: We derive and implement an alignment-free estimator of the number of pairwise mismatches, $\widehat{\pi}_{\mathrm{m}}$. Our implementation of $\widehat{\pi}_{\mathrm{m}}$, `pim`, is based on an enhanced suffix array and inherits the superior time and memory efficiency of this data structure. Simulations demonstrate that $\widehat{\pi}_{\mathrm{m}}$ is accurate if mutations are distributed randomly along the chromosome. While real data often deviates from this ideal, $\widehat{\pi}_{\mathrm{m}}$ remains useful for identifying regions of low genetic diversity using a sliding window approach. We demonstrate this by applying it to the complete genomes of 37 strains of *Drosophila melanogaster*, and to the genomes of two closely related *Drosophila* species, *D.simulans* and *D.sechellia*. In both cases, we detect the diversity minimum and discuss its biological implications.

Availability: `pim` is written in standard C and its sources can be downloaded from http://guanine.evolbio.mpg.de/pim/.

Contact: haubold@evolbio.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 17, 2010; revised on November 23, 2010; accepted on December 9, 2010

1 INTRODUCTION

The study of sequence diversity within natural populations provides a powerful link between raw sequence data and functional knowledge. Such population genetic studies have traditionally been centered on a few transcribed loci. However, the exponential decay of sequencing costs over the past two decades has recently made it possible to carry out population genetics on a genome scale. In a comprehensive early example of such studies, Begun *et al.* (2007) surveyed the resequenced genomes of six *Drosophila simulans* lines with the closely related *D.yakuba* serving as outgroup. They described the large-scale patterns of within population diversity and between species divergence along the three major *D.simulans* chromosomes. While diversity declined toward the telomeres and centromeres, where recombination rates are low, divergence, and thus mutation rates, remained constant across the chromosomes, which is a classical hallmark of selection at linked sites (Begun and Aquadro, 1992; Charlesworth *et al.*, 1993).

Similar genome-wide studies of population genetic processes are poised to become the norm. There are currently several projects under way aimed at sampling 1000 genomes of single model species. This shift to genome-scale sequence comparison has sparked interest in new population-centered sequence analysis methods. These fall into two categories: alignment-based and alignment-free.

As to alignment-based methods, a number of recent studies have focused on the problem of estimating population mutation and recombination rates from aligned shotgun reads (Haubold *et al.*, 2010; Hellmann *et al.*, 2008; Jiang *et al.*, 2009; Johnson and Slatkin, 2006; Lynch, 2008, 2009). The main challenge in this work is to account for variable coverage and the errors introduced by the diverse chemistries of next-generation sequencing approaches.

In contrast, alignment-free methods of sequence comparison usually start from finished genomes or contigs. Given two homologous sequences, the most popular approaches are based on one of two kinds of measurements: the correlation between the frequencies of short 'words' such as 5mers found in each sequence (Chapus *et al.*, 2005), and the lengths of matches between sequences (Ulitsky *et al.*, 2006). In Bioinformatics, such methods have a long tradition due to their superior computational efficiency when compared with sequence alignment (Ferragina *et al.*, 2007; Martinez, 1983; Vinga and Almeida, 2003). However, until recently a disadvantage of alignment-free sequence comparison was that word frequencies or match lengths could not be transformed into mutation rates.

In order to combine the computational efficiency of alignment-free sequence comparison with the biological relevance of alignment-based divergence estimation, Haubold *et al.* (2009) have derived and implemented a moment-based estimator of the substitution rate from match lengths. This estimator can be used to rapidly compute meaningful phylogenies from closely related genomes ranging in size from 10^4 to 10^8 bases (Domazet-Lošo and Haubold, 2009). Here, we extend this work by deriving a maximum likelihood estimator of pairwise nucleotide diversity that can be computed without alignment. We explore the accuracy of this estimator using simulations. In addition, we survey the genetic diversity along the chromosomes of 37 strains of *D.melanogaster* and the genome-wide divergence between *D.simulans* and *D.sechellia*.

2 APPROACH AND DATA

2.1 Derivation of estimator ($\widehat{\pi}_{\mathrm{m}}$) without recombination

Our estimator of nucleotide diversity, $\widehat{\pi}_{\mathrm{m}}$, is based on the distribution of the lengths of exact matches between a pair of DNA sequences. Following the previous description of these exact

*To whom correspondence should be addressed.

matches (Haubold *et al.*, 2009), we label one member of such a pair *query*, the other *subject*. For each suffix in the query, we look up the length of the shortest prefix that is absent from the subject. Haubold *et al.* (2005) designated these absent prefixes SHortest Unique subSTRINGS (shustrings) and showed how to look them up efficiently using an enhanced suffix array (Abouelhoda *et al.*, 2002; Domazet-Lošo and Haubold, 2009).

In our analysis, we assume that the shustring in the query is homologous to the corresponding substring in the subject. Note that we do not need to know the position of the homologous substring. To express diversity as a function of shustring lengths, let π be the number of mismatches per site between a pair of sequences. Under the infinite sites model, where each mutation affects a hitherto unchanged site, π is a proxy for the rate of mutation per site. In the absence of recombination, the probability of observing a shustring of length $X=x$ is therefore the probability of observing a mutation terminating a shustring preceded by $x-1$ constant positions :

$$\begin{aligned} P\{X=x\} &= \pi(1-\pi)^{x-1} \\ &\approx \pi e^{-\pi(x-1)}. \end{aligned} \tag{1}$$

We generalize this to calculate the probability of the shustring lengths found at every position in a query,

$$P\{X_i=x_i\forall i\}\approx\prod_i \pi e^{-\pi(x_i-1)}, \tag{2}$$

where we ignore dependencies between shustrings at different positions, and where X_i is the (random) shustring length at position i. Equivalently,

$$\log(P\{X_i=x_i\forall i\})\approx\sum_i \log\pi-\pi(x_i-1). \tag{3}$$

To find the value of π that maximizes this expression, we differentiate with respect to π

$$0=\frac{\partial}{\partial\pi}\sum_i \log\pi-\pi(x_i-1)=\sum_i \frac{1}{\pi}-(x_i-1), \tag{4}$$

which leads to our maximum likelihood estimator of π,

$$\widehat{\pi}_{\mathrm{m}}=\frac{|Q|}{\sum_i X_i-1}, \tag{5}$$

where $|Q|$ is the length of the query sequence. Notice that $\widehat{\pi}_{\mathrm{m}}$ is approximately the inverse of the average shustring length. As a consequence of the ergodic theorem [e.g. Durrett (2010), Section 6], $\widehat{\pi}_{\mathrm{m}}\approx\pi$ for large $|Q|$.

2.2 $\widehat{\pi}_{\mathrm{m}}$ with rcombination

Here, we derive the basic property that $\widehat{\pi}_{\mathrm{m}}$ is downwardly biased in the presence of recombination. For this, we use the approximation that the single nucleotide polymorphisms (SNPs) are generated by a Poisson point process in the absence of recombination and tend to cluster in the presence of recombination. Figure 1 shows 100 SNPs distributed along a single pair of sequences of arbitrary length simulated with or without recombination. In the absence of recombination, the time to the most recent common ancestor is constant along the sequences and hence the SNPs are distributed according to a constant rate Poisson process (Fig. 1A). In contrast, with a rate of recombination that is similar to the mutation rate, which is the case in most eukaryotes, the time

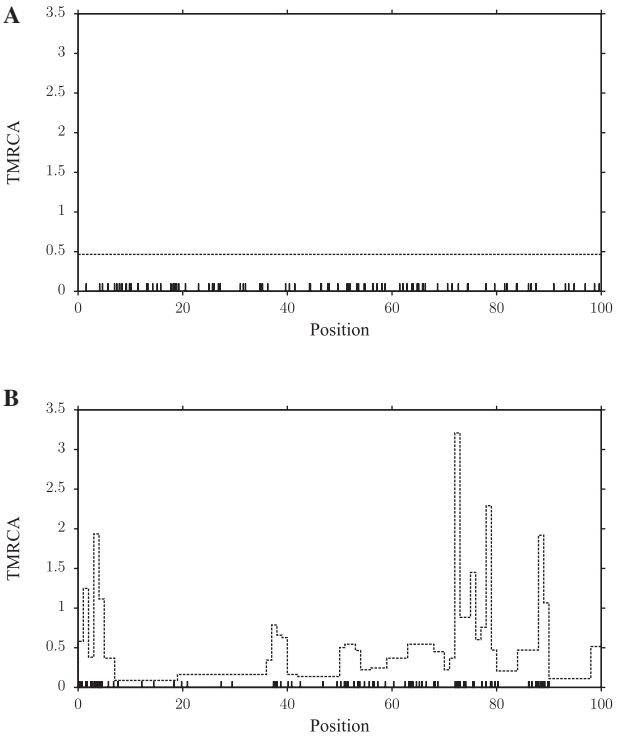

Fig. 1. The effect of recombination on the Time to the Most Recent Common Ancestor (TMRCA, dashed line) and the distribution of mutations (solid vertical lines) along a sequence. A pair of sequences of arbitrary length was simulated with 100 mutations. (**A**) No recombination; (**B**) with recombination between 100 sequence segments at a rate equal to the rate of mutation ($\rho=100$).

to the most recent common ancestor varies widely along the sequence (Fig. 1B). As a consequence, young, long segments contain fewer SNPs than old segments, which have had more time to both mutate and recombine, and SNPs cluster along the sequence (Wiuf and Hein, 1999).

For our exploration of the mathematical properties of $\widehat{\pi}_{\mathrm{m}}$ with recombination, we assume that the SNPs arise at (random) positions $Z_0<Z_1<\ldots$. Clearly, if $Z_k<i<Z_{k+1}$, we find $X_i=Z_{k+1}-i+1$. We denote by $W=Z_{i+1}-Z_i$ the waiting time between two SNPs (which by homogeneity does not depend on i). Any random site in the DNA is located between two SNPs separated by some distance $\widetilde{W}$. The distribution of $\widetilde{W}$ is given by the size-biased distribution of W, since there is a higher chance to pick a site between two distant SNPs [see e.g. Kallenberg (1984)]:

$$\mathbb{P}[\widetilde{W}=w]=\frac{w\mathbb{P}[W=w]}{\mathbb{E}[W]}.$$

Moreover, given $\widetilde{W}$, X is uniformly distributed between 1 and $\widetilde{W}$. Therefore,

$$\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X|\widetilde{W}]]=\mathbb{E}\Big[\frac{\widetilde{W}}{2}\Big]=\frac{\sum_{w=1}^{\infty} w^2\mathbb{P}[W=w]}{2\mathbb{E}[W]}=\frac{\mathbb{E}[W^2]}{2\mathbb{E}[W]}.$$

Assume $\pi\ll 1$ between the two sequences. This means that $\mathbb{E}[W]\approx 1/\pi$. Under recombination, SNPs occur in clusters and hence, $\mathbb{V}[W]>1/\pi^2$, since $1/\pi^2$ is the approximate variance of W without recombination. As a consequence, for large $|Q|$, and again using the

ergodic theorem,

$$\widehat{\pi}_{\mathrm{m}} = \frac{1}{\frac{1}{|Q|}\sum x_i - 1} \approx \frac{1}{\mathbb{E}[X]} = \frac{2\pi}{\pi^2 \mathbb{V}[W] + 1} < \pi.$$

This explains why $\widehat{\pi}_{\mathrm{m}}$ is downwardly biased under recombination. Moreover, it shows that the size of the bias is determined by the deviation of $\mathbb{V}[W]$ from $1/\pi^2$. However, the function that relates $\mathbb{V}[W]$ to the rate of recombination is unknown.

2.3 Algorithm and implementation

We have implemented the computation of $\widehat{\pi}_{\mathrm{m}}$ based on a publicly available software library for text indexing (Manzini and Ferragina, 2002). The resulting program, `pim`, can compute $\widehat{\pi}_{\mathrm{m}}$ for two closely related, unaligned DNA sequences in FASTA format. In addition, it efficiently carries out the sliding window analyses described in the Section 3. The software is written in standard C, runs on the UNIX command line and its documented source code is freely available from

`http://guanine.evolbio.mpg.de/pim/`

2.4 Simulations

We simulated samples of DNA sequences according to the following protocol:

(1) Use the coalescent simulation program `ms` (Hudson, 2002) to generate a set of haplotypes conditioned on the number of segregating sites and the rate of recombination.

(2) Convert these haplotypes to DNA sequences using our program `ms2dna`, which is freely available as part of our collection of sequence analysis tools, `biobox`:

`http://guanine.evolbio.mpg.de/bioBox/`

(3) Calculate $\widehat{\pi}_{\mathrm{m}}$ and in many cases also π from the simulated sequences. π was computed using the program `gd`, which is also part of `biobox`.

2.5 Raw data and data analysis

At the time of our analysis, the Drosophila Population Genomics Project had released version 1.0 of its dataset. This contained the following 37 strains collected in Raleigh, NC, USA: RAL-301_1, RAL-303_1, RAL-304_1, RAL-306_1, RAL-307_2, RAL-313_1, RAL-315_1, RAL-324_1, RAL-335_2, RAL-357_1, RAL-358_1, RAL-360_1, RAL-3 62_2, RAL-365_1, RAL-375_1, RAL-379_1, RAL-380_2, RAL-391_2, RAL-399_1, RAL-427_1, RAL-437_1, RAL-486_1, RAL-514_1, RAL-517_1, RAL-555_1 , RAL-639_1, RAL-705_1, RAL-707, RAL-714_1, RAL-730_1, RAL-732_1, RAL-765_1, RAL-774_1, RAL-786_1, RAL-799_1, RAL-820_1, RAL-852_1. We downloaded the corresponding aligned sequence data from `http://www.dpgp.org`.

For the comparison between *D.simulans* and *D.sechellia*, we used the data released as part of the 12 *Drosophila* species genome sequencing project (Drosophila 12 Genomes Consortium, 2007) at:

`http://rana.lbl.gov/drosophila/caf1/all_caf1.tar.gz`

Tajima's D (Tajima, 1989) was computed using a program by Guillaume Achaz:

`http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html`

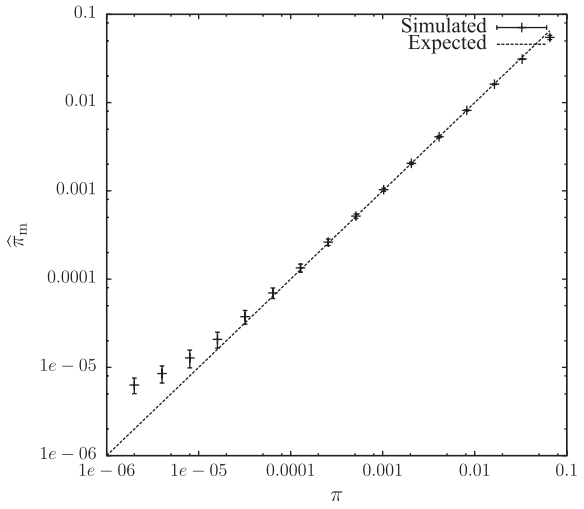

Fig. 2. Our estimator of the number of pairwise mismatches, $\widehat{\pi}_{\mathrm{m}}$, as a function of the true number of pairwise mismatches, π. Each point represents the average ±SD of 10^3 iterations with sequence pairs of length 500 kb each.

3 RESULTS

In the following sections, we apply $\widehat{\pi}_{\mathrm{m}}$ to simulated data with and without recombination and to two empirical datasets: 37 genomes of *D.melanogaster* and the genomes of the two closely related *Drosophila* species *D.simulans* and *D.sechellia*.

3.1 Simulations

3.1.1 Without recombination We started our exploration of the properties of $\widehat{\pi}_{\mathrm{m}}$ by simulating pairs of 500 kb sequences with a fixed number of mismatches. Figure 2 shows that when the mutation rate is close to the inverse of the sequence length, $\widehat{\pi}_{\mathrm{m}}$ overestimates π, while for mutation rates >0.03 it underestimates π. However, for population genetically relevant intermediate values of π, the theory closely approximates the simulations.

3.1.2 With recombination The population recombination rate is expressed as $\rho = 2N_{\mathrm{e}}c$, where N_{e} is the effective number of chromosomes in the population and c is the probability of a recombination affecting a given stretch of sequence per generation. As explained in Section 2, $\widehat{\pi}_{\mathrm{m}}$ is based on the distances between pairs of mutations. Recombination causes these distances to deviate from uniformity (Fig. 1), which makes $\widehat{\pi}_{\mathrm{m}}$ sensitive to recombination. In agreement with these theoretical considerations, Figure 3 shows that even low rates of reciprocal recombination lead to strong underestimation of π due to an 'excess' of long, young tracts of low diversity. This means that $\widehat{\pi}_{\mathrm{m}}$ in its current form is not generally suitable as a global measure of the number of pairwise mismatches. However, as we demonstrate in the next paragraph, it is still useful as an indicator of local fluctuations in π.

3.1.3 Sliding window analysis We began our study of the local properties of $\widehat{\pi}_{\mathrm{m}}$ with a sliding window analysis of one pair of 500 kb sequences without recombination. In Figure 4A, the average value of $\widehat{\pi}_{\mathrm{m}}$ is 1.007×10^{-2}, which is close to the average value of π, 1.001×10^{-2}. In addition, $\widehat{\pi}_{\mathrm{m}}$ closely tracks the true diversity (Pearson's correlation $r = 0.76$; root mean square error,

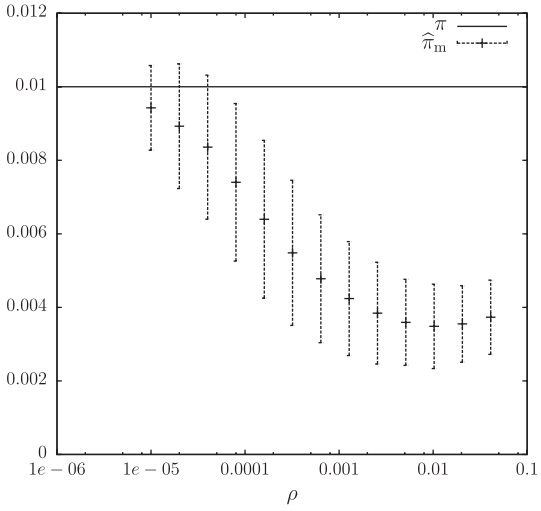

Fig. 3. The new estimator of the number of pairwise mismatches, $\widehat{\pi}_{\mathrm{m}}$, as a function of the population recombination rate, ρ. Each point represents the average $\pm$SD of 10^3 iterations with sequence pairs of length 500 kb each and $\pi = 0.01$.

RMSE $= 7.5 \times 10^{-4}$). As expected from the simulations shown in Figure 3, the addition of recombination at a rate of $\pi = \rho$ leads to an overall underestimation of π with an average $\widehat{\pi}_{\mathrm{m}}$ of 5.632×10^{-3} compared with the average π of 7.407×10^{-3}. This average underestimation of π can also be observed in Figure 4B where the RMSE increases sixfold (RMSE $= 4.5 \times 10^{-3}$), even though the local correlation is not decreased compared to the case without recombination ($r = 0.82$). In 1000 repetitions of this sliding window analysis $<r> \pm \mathrm{SD} = 0.72 \pm 0.06$ and $<\mathrm{RMSE}> = 9.2 \times 10^{-4} \pm 7.8 \times 10^{-5}$, while with recombination $<r> = 0.78 \pm 0.05$ and $<\mathrm{RMSE}> = 4.5 \times 10^{-3} \pm 3.1 \times 10^{-4}$.

3.2 Genetic diversity in a sample of *D.melanogaster*

We carried out a sliding window analysis of $\widehat{\pi}_{\mathrm{m}}$ along the five chromosomal arms of the 37 *D.melanogaster* strains contained in the sample from Raleigh, NC, USA. It is important to realize that in an alignment-free sliding window analysis sequence coordinates can only refer to a single query sequence. As a result, of the $\binom{37}{2}$ possible pairwise comparisons only 36 comparisons between an anchor sequence and the other members of the sample can be carried out. We arbitrarily used the first sequence in the dataset, RAL-301_1, as this anchor query, but apart from numerical detail our results are unaffected by the choice of anchor sequence [see (Haubold *et al.*, 2009) for a more thorough discussion of shustring-based metrics]. Figure 5 shows the distribution of the $\widehat{\pi}_{\mathrm{m}}$ values compared with the corresponding π values computed from the aligned original dataset. As expected from Figure 3, the two distributions clearly differ with $\widehat{\pi}_{\mathrm{m}}$ having a mean $\pm$ SD of 0.0041 ± 0.0001 and π a mean of 0.0056 ± 0.0017. However, the tails of the two distributions are remarkably similar with almost identical ranges ($0.00085 \leq \widehat{\pi}_{\mathrm{m}} \leq 0.001, 0.00097 \leq \pi \leq 0.001$).

In our subsequent analysis, we concentrated on the lower tail of the distribution and located the global minimum of $\widehat{\pi}_{\mathrm{m}}$ at the telomeric tip of chromosome 3L (Fig. 6A). Figure 6B zooms into this region with a 10 kb sliding window to reveal the global

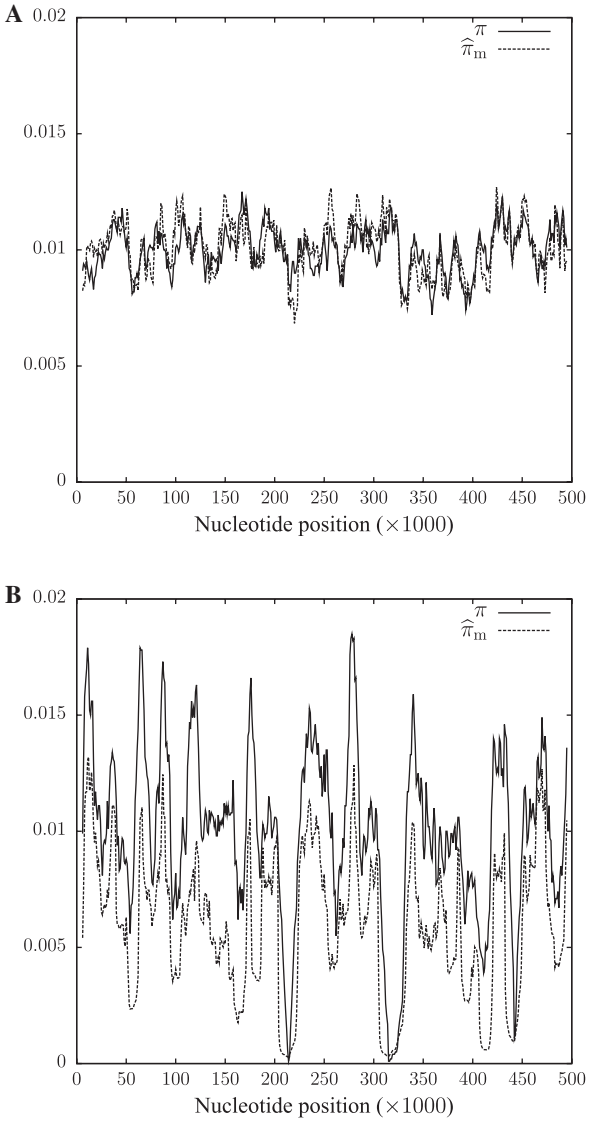

Fig. 4. Sliding window analysis of the number of pairwise mismatches, π and $\widehat{\pi}_{\mathrm{m}}$, along a single pair of 500 kb DNA sequences with $\pi = 0.01$. Window size was 10 kb. (**A**) No recombination, Pearson's correlation between π and $\widehat{\pi}_{\mathrm{m}}$: $r = 0.76$; (**B**) with recombination ($\rho = 0.01$), $r = 0.82$.

diversity minimum with $\widehat{\pi}_{\mathrm{m}} = 7.8 \times 10^{-4}$ centered on nucleotide 199 799. Significantly, the lowest of the corresponding π values was exactly centered on the homologous region located using $\widehat{\pi}_{\mathrm{m}}$. We analyzed 5 kb around this point from all 37 strains and identified 37 segregating sites in the region. Of these, 31 were singletons. Since a completed selection event (selective sweep) is characterized by an excess of singletons, we quantified the deviation from neutrality by computing a statistic commonly used to test the null hypothesis of neutral evolution called Tajima's D, D_{T} (Tajima, 1989). It is proportional to the difference between the average number of pairwise differences and the normalized number of SNPs in a sample of DNA sequences. Under neutrality, these two quantities are equal and hence the neutral expectation of D_{T} is zero, while after a complete selective sweep D_{T} is expected to be transiently negative.

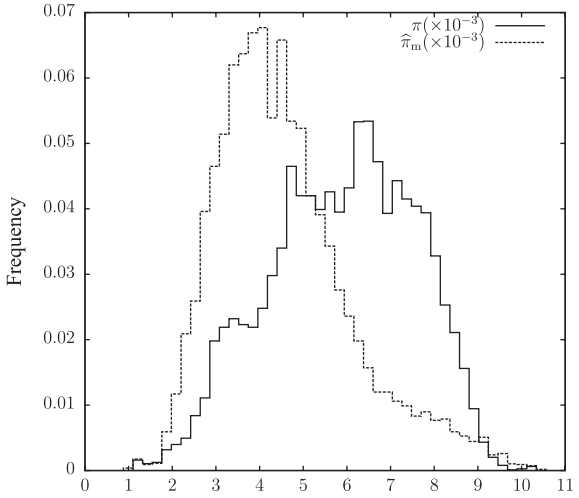

Fig. 5. Distribution of the $\widehat{\pi}_{\rm m}$ values obtained in the 150 kb sliding window analysis of 37 complete *D.melanogaster* genomes compared with the corresponding distribution of π values.

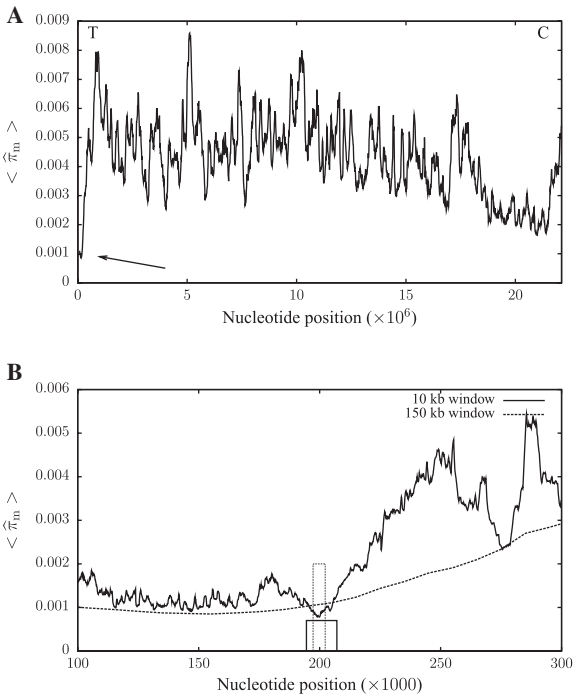

Fig. 6. Sliding window analysis of average $\widehat{\pi}_{\rm m}$, $<\widehat{\pi}_{\rm m}>$, when comparing chromosome 3L of *D.melanogaster* strain RAL-301_1 in turn to all the remaining 36 strains in the sample. (**A**) Windows of length 150 kb were plotted every 10 kb; *T*, telomere; *C*, centromere; the *arrow* indicates the global minimum of $\widehat{\pi}_{\rm m}$. (**B**) Zoom into the global minimum; the dotted, tall box indicates the 5 kb region subjected to the analysis of nucleotide frequencies (see text); the solid box shows the position of the gene *Enhancer of bithorax* [*E(bx)*], which encodes the NURF301 subunit of the NUcleosome Remodelling Factor, a key regulator of innate immunity.

In our case, $D_{\rm T}$ is significantly negative ($D_{\rm T} = -2.54$; $P = 0.00013$). Since 9 of the 31 singletons were due to strain RAL-427_1, we recalculated $D_{\rm T}$ in its absence and found a still significant value of $D_{\rm T} = -2.42$ ($P = 0.00045$).

The *D.melanogaster* genome sample is released in aligned form. This allowed us to check the accuracy of $\widehat{\pi}_{\rm m}$ by comparison to its alignment-based analogue, π. However, it does does not make a compelling case for *alignment-free* analysis and we therefore go on to demonstrate *bona fide* alignment-free genome comparison in the next section.

3.3 Comparison between *D.simulans* and *D.sechellia*

Drosophila simulans has diverged from *D.sechellia* for only approximately 1 million years (Drosophila 12 Genomes Consortium, 2007). In the Comparative Assembly Freeze 1 of the 12 *Drosophila* genome sequencing project, the genome of *D.simulans* was released in fully assembled form while the genome of *D.sechellia* consisted of 14 730 contigs. The combined size of these two datasets was 308 982 892 bp. We analyzed these data with *D.simulans* serving as query in a 150 kb sliding window analysis of divergence from *D.sechellia*. This took 22 min and 44 s on an AMD Opteron 2.3 GHz processor. In order to compare this to the time requirement of the equivalent alignment-based analysis, we aligned these two genomes using the fast genome aligner `MUMmer` (Kurtz *et al.*, 2004) on the same machine, which took 1 h, 32 min, 42 s.

Again, we focus our analysis of the sliding window scan on regions of exceptional similarity between the two genomes and Figure 7A shows the sliding window analysis for chromosome 2L, which contains the global minimum of $\widehat{\pi}_{\rm m}$ at position 13 965 000. In contrast to the diversity minimum (Fig. 6A), the divergence minimum is far from either the centromere or the telomere. In Figure 7B, we zoom into this region and show in addition to the 150 kb sliding windows 10 kb sliding windows, which reveal a number of conserved regions. The most highly conserved of these regions intersects with the protein-coding gene *pickpocket* (*ppk*), which is involved in larval locomotion and development (Ainsley *et al.*, 2008).

To investigate whether the *D.sechellia* 12 genomes *ppk* reference allele is typical for *D.sechellia*, we compared it to the allele from an unrelated *D.sechellia* line (SynA, provided by E. Earley and C. Jones). The nucleotide divergence between the two *D.sechellia* lines is 6.4×10^{-3} (14 unambiguous SNPs in a 2198 bp alignment), which is higher than expected for *D.sechellia* in general (Legrand *et al.*, 2009), but is still much lower than typical of divergence between *D.sechellia* and *D.simulans* (Fig. 7). This confirms that there is no reason to suspect that the *ppk* allele we identified in the *D.sechellia* reference genome is the result of an assembly error.

4 DISCUSSION

Much has recently been written about the impressive speed of next-generation sequencing equipment (Shendure and Ji, 2008), which is driving the current generalization of population genetics to population genomics (Begun *et al.*, 2007). However, technology change in sequence analysis algorithms has been equally remarkable. A substantial portion of these advances is based on the progressive abstraction of suffix trees (Gusfield, 1997) to enhanced suffix arrays (Abouelhoda *et al.*, 2002) and, more recently, to highly memory efficient FM indexes (Ferragina *et al.*, 2008).

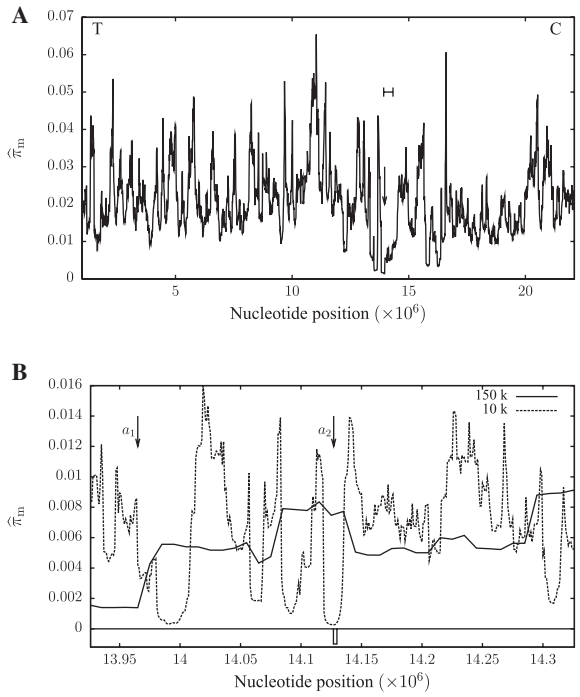

Fig. 7. Sliding window plot of $\widehat{\pi}_m$, when comparing chromosome 2L of *D.simulans* to the genome of *D.sechellia*. (**A**) Windows of length 150 kb were plotted every 10 kb; *T*, telomere; *C*, centromere; the *arrow* indicates the global minimum of $\widehat{\pi}_m$. (**B**) Zoom into the global minimum indicated by the *interval mark* in (A); a_1, arrow in (A) indicating the global minimum of 150 kb windows; a_2, global minimum of 10 kb windows; the *solid box* shows the position of the gene *pickpocket* (*ppk*).

Here we point out that the distances between SNPs can quickly be looked up using suffix tree and that these distances are connected to the number of mismatches by Equation (1). This suggests our new estimator of pairwise genetic diversity, $\widehat{\pi}_m$, given in Equation (5).

Our implementation of this estimator is based on an enhanced suffix array. The efficiency of the resulting software is best judged by comparing it to an alignment tool that uses the same data structure, such as `MUMmer` (Kurtz *et al.*, 2004). As we report in the Section 3, `pim` is four times faster than `MUMmer` when applied to the genomes of *D.sechellia* and *D.simulans*. This comparison ignores that the `MUMmer` result would still need to be processed to yield the desired diversity values. It should also be borne in mind that alignment-free approaches are bound to outperform alignment-based methods as the latter need to carry out far more steps than the former. The important question about an alignment-free estimator such as $\widehat{\pi}_m$ is therefore how useful it is for quantifying genetic diversity in real genomes.

We show through simulations that for mutation rates greater than the inverse of the sequence length and less than 0.03, our estimator is very precise (Fig. 2). This is the range of mutation rates relevant for analyzing genomes sampled from populations or closely related species.

Recombination causes fluctuations in the time to the most recent common ancestor along a chromosome, which results in clustering of mutations (Fig. 1). As a consequence, the aggregate shustring length increases, which leads to an underestimation of π in the presence of even low levels of recombination (Fig. 3) and for biologically relevant recombination rates ($\rho \approx \pi$), the effect on $\widehat{\pi}_m$ is rather strong. While this makes $\widehat{\pi}_m$ unsuitable for the estimation of global π in recombining genomes, our sliding window analyses of simulated sequences demonstrate that $\widehat{\pi}_m$ accurately tracks local fluctuations in nucleotide diversity (Fig. 4) and is reasonably robust to the choice of window length (Figs 6B and 7B; see also Supplementary Material).

We have compared the results of our alignment-free estimator to the corresponding alignment-based measure wherever possible. The sequences published by the Drosophila Population Genomics Project allowed this comparison to be made for a substantial empirical dataset, as the newly released genomes are distributed aligned with respect to the reference genome. Such a comfirms that $\widehat{\pi}_m$ tends on average to underestimate π, while preserving the range of true diversity values (Fig. 5). However, there is a reassuring correspondence between the empirical distributions of π and $\widehat{\pi}_m$ (Fig. 5) and the simulation with recombination (Fig. 3): we surmise that the regions where $\widehat{\pi}_m \approx \pi$ are those of exceptionally low recombination, like the subtelomeres, and those of exceptionally high recombination. In between these two extremes, $\widehat{\pi}_m$ is less than π and in Section 2.2 we show that the difference between π and $\widehat{\pi}_m$ is determined by the degree to which the variance in inter-SNP distance ($\mathbb{V}[W]$) exceeds its theoretical minimum of $1/\pi^2$.

In agreement with this interpretation of our statistic, we correctly identified the global diversity minimum at the telomeric tip of chromosome 3L (Fig. 6A). This region is also known to have reduced genetic diversity and recombination in *D.simulans* (Begun *et al.*, 2007). The region of lowest diversity intersects the gene *Enhancer of bithorax*, *E(bx)* (Fig. 6B). Not only is this gene among the most monomorphic in the genome, it also contains a significant excess of rare mutations, as expressed by the highly significant rejection of the standard neutral model ($P = 0.00013$).

Our interspecific genome comparison identified a highly conserved gene (*ppk*), which may be closely tied to a behavioral peculiarity of *D.sechellia*: on its native islands, the Sychelles, it feeds on the toxic fruit of the non-native Indian mulberry (*Morinda citrifolia*), while its sister species *D.simulans* and most other fruit files avoid this plant (Dworkin and Jones, 2009). It is therefore likely that *D.sechellia* has recently undergone a shift in its breeding environment (Lemeunier and Ashburner, 1984). This biological background makes it intriguing that the region that is most similar between *D.sechellia* and *D.simulans*, and is thus a candidate for recent introgression, contains a gene that affects larval foraging behavior.

Population genomics projects tend to be based on a well-curated reference genome that serves as the subject sequence to which new reads are aligned. This is certainly the approach of the Drosophila Population Genomics Project. Given such data, there might appear to be no case for alignment-free methods when dealing with samples of genomes drawn from closely related individuals. However, resequencing is often only a first step on the way toward full *de novo* assembly. Moreover, our divergence scan of *D.simulans* versus *D.sechellia* shows that unaligned genomes from closely related species can be compared efficiently and used to rapidly identify diversity outliers for further investigation. There are many such

genome pairs both among eukaryotes and bacteria, and given the current rate of sequencing, this number is set to increase.

ACKNOWLEDGEMENTS

We are grateful to the members of the Drosophila Population Genomics Project for freely distributing their data, and to Corbin D. Jones and Eric J. Earley for sharing unpublished *D.sechellia* sequence data. We have also benefited from discussions with R. Guy Reeves and comments on the manuscript by Vanessa L. Reed.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Freiburg Initiative for Systems Biology (grant 0313921 to P.P.). F.A.R. is supported by funds from the Max-Planck-Society.

Conflict of Interest: none declared.

REFERENCES

Abouelhoda,M. *et al.* (2002) The enhanced suffix array and its applications to genome analysis. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics.* Vol. of Lecture Notes in Computer Science, Springer, pp. 449–463.

Ainsley,J. *et al.* (2008) Sensory mechanisms controlling the timing of larval developmental and behavioral transitions require the drosophila DEG/ENaC subunit, pickpocket1. *Dev. Biol.,* **322**, 46–55.

Begun,D. and Aquadro,C.F. (1992) Levels of naturally occuring DNA polymorphism are correlated with recombination rates in *Drosophila melanogaster*. *Nature,* **356**, 519–520.

Begun,D. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.,* **5**, e310.

Chapus,C. *et al.* (2005) Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol. Biol.,* **5**, 63.

Charlesworth,B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics,* **134**, 1289–1303.

Domazet-Lošo,M. and Haubold,B. (2009) Efficient estimation of pairwise distances between genomes. *Bioinformatics,* **25**, 3221–3227.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, **450**, 203–218.

Durrett,R. (2010) *Probability—Theory and Examples*. 4th edn. Cambridge University Press, Cambridge, UK.

Dworkin,I. and Jones,C.D. (2009) Gentic changes accompanying the evolution of host specialization in *Drosophila sechellia*. *Genetics*, **181**, 721–736.

Ferragina,P. *et al.* (2007) Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, **8**, 252.

Ferragina,P. *et al.* (2008) Compressed text indexes: from theory to practice. *ACM J. Exp. Algorithms*, **13**, 1.12:1–1.12:31.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge.

Haubold,B. *et al.* (2005) Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics*, **6**, 123.

Haubold,B. *et al.* (2009) Estimating mutation distances from unaligned genomes. *J. Comput. Biol.*, **16**, 1487–1500.

Haubold,B. *et al.* (2010) `mlRho`: a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.*, **19**, 277–284.

Hellmann,I. *et al.* (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.*, **18**, 1020–1029.

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

Jiang,R. *et al.* (2009) Population genetic inference from resequencing data. *Genetics*, **181**, 187–197.

Johnson,P. and Slatkin,M. (2006) Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res.*, **16**, 1320–1327.

Kallenberg,O. (1984) An informal guide to the theory of conditioning in point processes. *Int. Stat. Rev.*, **52**, 151–164.

Kurtz,S. *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.

Legrand,D. *et al.* (2009) Species-wide genetic variation and demographic history of drosophila sechellia, a species lacking population structure. *Genetics*, **182**, 1197–1206.

Lemeunier,F. and Ashburner,M. (1984) Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (*Sophophora*). iv. the chromosomes of two new species. *Chromosoma*, **89**, 343–351.

Lynch,M. (2008) Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genomic-sequencing projects. *Mol. Biol. Evol.*, **25**, 2409–2419.

Lynch,M. (2009) Estimation of allele frequencies from high-coverage genome sequencing projects. *Genetics*, **182**, 295–301.

Manzini,G. and Ferragina,P. (2002) Engineering a lightweight suffix array construction algorithm. In *ESA '02: Proceedings of the 10th Annual European Symposium on Algorithms* . Springer, London, UK, pp. 698–710.

Martinez,H.M. (1983) An efficient method for finding repeats in molecular sequences. *Nucleic Acids Res.*, **11**, 4629–4634.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

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

Ulitsky,I. *et al.* (2006) The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol.*, **13**, 336–350.

Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison—a review. *Bioinformatics*, **19**, 513–523.

Wiuf,C. and Hein,J. (1999) Recombination as a point process along a sequence. *Theor. Popul. Biol.*, **55**, 248–259.