Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples

Simon Boitard,*[1] Christian Schlötterer,[2] Viola Nolte,[2] Ram Vinay Pandey,[2] and Andreas Futschik[3]

[1]Institut National de la Recherche Agronomique, Laboratoire de Génétique Cellulaire, Castanet-Tolosan, France

[2]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria

[3]Institute of Statistics and Decision Support Systems, University of Vienna, Vienna, Austria

***Corresponding author:** E-mail: simon.boitard@toulouse.inra.fr.

Associate editor: Ryan Hernandez

Abstract

Due to its cost effectiveness, next-generation sequencing of pools of individuals (Pool-Seq) is becoming a popular strategy for characterizing variation in population samples. Because Pool-Seq provides genome-wide SNP frequency data, it is possible to use them for demographic inference and/or the identification of selective sweeps. Here, we introduce a statistical method that is designed to detect selective sweeps from pooled data by accounting for statistical challenges associated with Pool-Seq, namely sequencing errors and random sampling among chromosomes. This allows for an efficient use of the information: all base calls are included in the analysis, but the higher credibility of regions with higher coverage and base calls with better quality scores is accounted for. Computer simulations show that our method efficiently detects sweeps even at very low coverage (0.5× per chromosome). Indeed, the power of detecting sweeps is similar to what we could expect from sequences of individual chromosomes. Since the inference of selective sweeps is based on the allele frequency spectrum (AFS), we also provide a method to accurately estimate the AFS provided that the quality scores for the sequence reads are reliable. Applying our approach to Pool-Seq data from *Drosophila melanogaster*, we identify several selective sweep signatures on chromosome X that include some previously well-characterized sweeps like the *wapl* region.

Key words: selective sweeps, next-generation sequencing, pooled DNA, Drosophila, allele frequency spectrum, hidden Markov model.

Research article

Introduction

The detection of genomic regions that have evolved under natural selection is an important topic in population genetics, which poses interesting theoretical challenges and holds great potential for medical and economic benefits. The case of hard sweeps, where a new mutant goes to fixation in a population due to strong directional selection, has received particular attention. Exploiting a typical pattern of reduced genetic diversity in the vicinity of the selected site, several methods were proposed to detect such events by screening the allele frequencies along the genome in a single population (Kim and Stephan 2002; Jensen et al. 2005; Nielsen et al. 2005; Boitard et al. 2009) and were applied to several species (Li and Stephan 2006; Williamson et al. 2007).

Today, the advent of next-generation sequencing (NGS) technologies provides a new dimension to such genome scans for selection. Genomes can be covered with very high density, and the ascertainment bias caused by SNP identification is becoming less important. Currently, the precise identification of individual genotypes, which requires a high sequencing coverage of each individual, remains very expensive for large samples. However, hard sweeps can also be detected when using only information concerning the genetic diversity of the sample along the genome. This information can be obtained also from experiments where the DNA from a pool of individuals is sequenced simultaneously (Pool-Seq). Although Pool-Seq is considerably cheaper than the sequencing of individuals, there are some methodological challenges associated with the analysis of the resulting data. For a discussion, see Futschik and Schlötterer (2010).

The new sequencing technologies have resulted in a dramatic cost reduction compared with classic Sanger sequencing, but the error rate per sequence is considerably higher. Even for diploid individuals, the distinction between sequencing errors and true SNPs is challenging when the coverage is not high enough. Similarly, for Pool-Seq, the identification of rare SNPs is difficult. One common strategy is thus to remove all singletons or doubletons from the analysis, because they might also result from sequencing errors. For the same reason, base calls with low-quality scores tend to be removed as well. Although it is possible to obtain unbiased estimates of genetic diversity using this approach, it is apparent that information is lost. In particular, the detection of selective sweeps could be compromised by this strategy because low-frequency alleles are an important signal to detect recent selective sweeps.

Hidden Markov models provide a natural framework to take both sequencing errors and unequal local coverage

into account. Here, we develop a Hidden Markov Model for detecting sweeps using pooled NGS data: This model extends the one investigated in Boitard et al. (2009) for classical sequencing data. As part of the model, we also introduce a version of the Expectation Maximization (EM) algorithm to estimate the allele frequency spectrum (AFS) using the information from all available genomic positions. Indeed, the estimated AFS is used to scan the genome for regions where the AFS is biased toward extreme allele frequencies. Our approach involves computing the likelihoods of the observed read information conditional on the number of derived alleles in the pool across genome positions. It takes into account the uncertainty concerning the true allele frequencies in the pool, which might typically be higher for sites with low-coverage or bad-quality scores.

Using computer simulations, we study the accuracy of this procedure for estimating the AFS, and its power for detecting selective sweeps. We then apply our approach to scan the X chromosome of *Drosophila melanogaster*, using two pooled samples of 97 flies sequenced at 100× coverage.

Materials and Methods

Accounting for Sequencing Errors and Chromosome Sampling at One Position

We consider here a sample of n chromosomes that have been subjected to Pool-Seq. We assume an infinite sites model and denote by Y_i, the number of derived alleles at genomic position i ($0 \leq Y_i \leq n$). With NGS of pools, Y_i is unobserved. We observe a collection of r_i reads, among which the observed number of derived alleles will usually differ from Y_i due to 1) the random sampling of reads from the n chromosomes and 2) the sequencing errors. Let $Z_{i,j}$ ($0 \leq j \leq r_i$) denote an indicator variable equal to 1, if the observed allele at read j is derived, and let $Z_i = (Z_{i,1}, \ldots, Z_{i,r_i})$. Let furthermore, $e_{i,j}$ be the probability for a sequencing error at read j. The conditional probability of the observed reads Z_i given Y_i is then equal to

$$\mathbb{P}(Z_i|Y_i) = \prod_{j:Z_{i,j}=1}\left((1-e_{i,j})\frac{Y_i}{n} + e_{i,j}\left(1-\frac{Y_i}{n}\right)\right) \times \prod_{j:Z_{i,j}=0}\left((1-e_{i,j})\left(1-\frac{Y_i}{n}\right) + e_{i,j}\frac{Y_i}{n}\right). \quad (1)$$

In this equation, Y_i/n should be interpreted as the probability that read j is taken from the subset of derived alleles in the pool. Because the sequencing is performed on one single pool, this probability is the same for all reads j. It is equal to Y_i/n because we assume that reads are sampled uniformly from each of the n chromosomes. Indeed, we do not account here for the possible biases arising from unequal concentration or quality among individuals, or allele specific amplification. Note that the influence of unequal concentration or quality among individuals on allele frequency estimation are expected to be low for large sample sizes, as shown by the derivations of Futschik and Schlötterer (2010). The sequencing error probabilities, $e_{i,j}$, can be deduced from the PHRED scores, $Q_{i,j}$, provided with the sequenced data, using the relation $e_{i,j} = 10^{-Q_{i,j}/10}$. As Illumina PHRED scores are known to be biased (Dohm et al. 2008), we include a discussion concerning the effects of inaccurate quality scores, as well as possible strategies to cope with the problem. (See the section on the real data application.)

Estimation of the AFS

Let $p = (p_0, \ldots, p_n)$ be the AFS in a region of length L covered by the reads, that is, p_j is the probability of observing j copies of the derived allele among the n chromosomes at a given genomic position. The likelihood of this spectrum given the observed reads is

$$\mathcal{L}(Z_1, \ldots, Z_L|p) = \prod_{i=1}^{L}\left(\sum_{Y_i=0}^{n}\mathbb{P}(Z_i|Y_i)\mathbb{P}(Y_i|p)\right) = \prod_{i=1}^{L}\left(\sum_{Y_i=0}^{n}\mathbb{P}(Z_i|Y_i)p_{Y_i}\right). \quad (2)$$

As this likelihood involves a summation over the unobserved variables Y_i, we propose to maximize it using an EM strategy. Our algorithm starts from an arbitrary initial value of p and iteratively computes new values of p that increase the current likelihood. If p^c is the current value of the AFS, the next value is given as

$$p_j^{c+1} = \frac{1}{L}\sum_{i=1}^{L}\frac{\mathbb{P}(Z_i|Y_i=j)p_j^c}{\sum_{k=0}^{n}\mathbb{P}(Z_i|Y_i=k)p_k^c}. \quad (3)$$

The EM iterations are thus based on the conditional probabilities computed using equation (1). This algorithm is similar to the *EM-AFS* strategy independently proposed in Li (2011). More details about the algorithm are provided in the Supplementary Material online.

What we denote by p here is an estimate of the allele frequency probabilities in a random sample of size n from the population. Inference based on coalescent theory, as the derivations of Nielsen et al. (2005) used in our sweep detection model, generally involve this quantity. Note, however, that the shape of this sample AFS can be expected to resemble the shape of the AFS in a population of size N. Indeed, it also permits to come up with an estimate of the population allele frequency distribution in terms of an approximate continuous model. A natural way to estimate the parameters of the continuous model would be via maximum likelihood. In a Bayesian context, Gompert and Buerkle (2011) use a continuous parametric model to come up with a prior distribution for p in their hierarchical model.

Detection of Selective Sweeps

Since the allele frequency pattern in the vicinity of a selected allele differs from the one under neutrality, such local variation in allele frequencies can be used to detect past

selection events (Kim and Stephan 2002; Nielsen et al. 2005). To detect selective sweeps from Pool-Seq data, we extend the Hidden Markov Model (HMM) approach that we initially developed for completely sequenced data (Boitard et al. 2009).

We assume that each site i is associated with a hidden state X_i, which can take three different values: "Selection," for the sites that are very close to a swept site; "Neutral," for the sites that are far away from any swept site; and "Intermediate," for the sites in between. These three values are associated with different frequency spectra (the "Selection" spectrum is more skewed toward low and high allele frequencies than the "Intermediate" spectrum, and even more than the "Neutral" one). The hidden states X_i form a Markov chain along the genome with a per site probability q of switching state, so that close sites tend to be in the same hidden state. The observed variables are Z_i, containing the site frequencies taken from the pooled sequence reads. After computing suitable emission probabilities, the Viterbi algorithm is used to predict the most likely hidden states X_i from the observed states, and thus detect the swept regions. Combining equation (1) with the AFS in hidden state X_i leads to the emission probabilities

$$\mathbb{P}(Z_i|X_i) = \sum_{Y_i=0}^{n} \mathbb{P}(Z_i|Y_i) p_{Y_i}^{X_i}. \quad (4)$$

We prefer, however, to only consider those sites for which two different alleles are observed among the reads (i.e., where $\sum_{j=1}^{r_i} Z_{i,j} > 0$), and run the HMM using the emission probabilities

$$\mathbb{P}\left(Z_i|X_i, \sum_{j=1}^{r_i} Z_{i,j} > 0\right) = \frac{\mathbb{P}(Z_i|X_i)}{1 - \mathbb{P}\left(\sum_{j=1}^{r_i} Z_{i,j} = 0\right)}. \quad (5)$$

These emission probabilities account also for the fact that an observed polymorphism could be due to a sequencing error.

Notice that the approach in Boitard et al. (2009) would not be applicable here, as it assumes equal coverage at each position and also that the true numbers of derived alleles Y_i are known.

A natural method for obtaining allele frequency spectra is to first estimate the AFS under the state "Neutral" by applying the EM algorithm presented above to the whole genome data. An approximate AFS for the other hidden states can then be obtained by adequately modifying the neutral AFS using the method described in Nielsen et al. (2005).

Simulations

We used MSMS (Ewing and Hermisson 2010) to simulate genomic samples under a coalescent model with mutation, recombination, and constant population size. Our considered models involved both neutrality and a selective sweep at a single locus. From the genomic samples obtained from MSMS, pooled NGS data were simulated using our own MATLAB code: For each site, a number of reads r_i was simulated (independently of the other sites) from a Poisson distribution with expected value λ, the coverage of the sequencing experiment. For each read j ($1 \leq j \leq r_i$), we then simulated the allele type $Z_{i,j}$ from a Bernoulli distribution with parameter Y_i/n. (Recall that Y_i is the derived allele frequency at site i in the pool.) Next, a sequencing error probability $e_{i,j}$ was generated by drawing from the empirical distribution of PHRED scores, observed in our NGS data set (supplementary fig. S3, Supplementary Material online). Finally, sequencing errors were introduced by drawing from a Bernoulli distribution with parameter $e_{i,j}$. This leads to a simulated NGS sample, which can then be used for the AFS estimation and the detection of selective sweeps.

Table 1. Selective Sweep Detection Power.

Sample Size	$n = 25$	$n = 50$	$n = 100$	$n = 200$
Pooled NGS data	0.91	0.90	0.91	0.90
Sequence data—all sites	0.89	0.88	0.87	0.87
Sequence data—segregating sites	0.57	0.60	0.64	0.62

Detection power for pools of $n = 25$ to $n = 200$ chromosomes of length $L = 100$ kb simulated under a constant population size coalescent model with $\theta = 0.003$, $\rho = 0.003$, and $\alpha = 500$. NGS data sets were simulated with an expected coverage, $\lambda = 100$.

For detecting selective sweeps, we adapted the approach taken in Boitard et al. (2009) to pooled NGS samples instead of complete sequence data. When analyzing our sample, we identify a selective event as soon as the state "Selection" has been predicted for at least one site. For evaluating the detection power under a given selective scenario, we simulate several samples under this scenario (500 in this study, because of the high computational cost of the analysis) and look at the percentage of scenarios for which a sweep window is detected that also includes the true position of the selected site. The output of the analysis depends on the switching probability q used in the transition matrix of the HMM. In order to calibrate this probability, we preliminarily simulated 500 neutral samples under the same demographic scenario and analyze them with different values of q. We select a value of q such that selection is detected in 5% of these neutral samples, which means that we have a false positive rate of 5%. For the results shown in table 1, the selected value of q was around 2.10^{-4}, with little variation for different sample sizes.

For the comparison between pooled NGS data and complete sequence data, we applied the HMMs described in Boitard et al. (2009).

Analysis of Drosophila Chromosome X

We looked for selective sweeps on the X chromosome of *Drosophila melanogaster* using two samples of 97 female flies, both taken from the F1 generation derived from 5,000 flies, which were collected in November 2009 at the Kahlenberg, Austria. These flies were adapted to lab conditions during 2 days before they reproduced to form the F1 generation. Two samples of 97 females were subjected to Pool-Seq.

Genomic DNA was extracted from 97 individuals, which were homogenized with a Ultraturrax T10 (IKA-Werke, Staufen, Germany) and purified with the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). Genomic

DNA was sheared with a S2 device (Covaris, Inc., Woburn, MA) and used to prepare paired-end genomic libraries with the Paired-End DNA Sample Preparation Kit (Illumina, San Diego, CA) following the manufacturer's instructions. Sequencing was performed with an Illumina GAIIx sequencer.

Reads were trimmed to remove low-quality bases and mapped with bwa (version 0.5.7) (Li and Durbin 2009) against the *D. melanogaster* reference genome (version 5.18) and Wolbachia (NC_002978.6). We used the following mapping parameters: -n 0.01 (error rate), -o 2 (gap opening), -d 12 and -e 12 (gap length) disabling the seed option. The alignment files were converted to the Sequence Alignment/Map (SAM) format using the bwa module sampe enabling a local alignment procedure (Smith–Waterman), whenever one of the reads of the pair could not be mapped with global alignment. The SAM files were filtered for reads mapped in proper pairs with a minimum mapping quality of 20 using SAMtools (Li et al. 2009). The filtered SAM files were converted into the pileup format. We used RepeatMasker 3.2.9 (www.repeatmasker.org) to create a gff file to mask simple sequence repeats and transposable elements of the *D. melanogaster* genome version 5.34. Finally, indels together with five flanking nucleotides (on both sides) were masked in the alignments of each population if the indel was present in at least one population and supported by at least two reads.

The expected coverage was 100× for sample 1 and 87× for sample 2.

We also explored recalibration of the read qualities using GATK (DePristo et al. 2011) before creating the pileup file. This software estimates the sequencing error probabilities based on reads from sites that are assumed to be nonpolymorphic. Consequently, a list of true polymorphic sites is needed. We included in this list: 1) the transposable element positions reported by RepeatMasker (see above), 2) the positions flanking indels (5 bp upstream and downstream), and 3) the positions with more than two copies of the minor allele.

The statistical analysis (both for AFS estimation and for genome scans for selection) was based on folded sequence data, so we did not require for SNPs the ancestral alleles to be known. The allele labels 0 and 1 have thus been chosen arbitrarily. We used the folded likelihood

$$\mathbb{P}_f(Z_i|Y_i) = \frac{1}{2}\mathbb{P}(Z_i|Y_i) + \frac{1}{2}\mathbb{P}(1 - Z_i|Y_i). \qquad (6)$$

Polymorphic sites with three different alleles were also used in the analysis. They were converted into SNPs by removing the least frequent allele, which we considered to be most likely due to a sequencing error.

For computational reasons, the AFS estimation was based on only 10% of the sites from the pileup file. This subset was selected at random and included about 2 million sites, which was largely sufficient to estimate the AFS in a pool of 200 chromosomes (see simulation results).

An important tuning parameter of the selection scans based on our HMM is the transition probability q between neutral and selected states. The larger the q, the less evidence is required for a transition to selection and the more sweep candidates will be detected. In our real data application, the transition probabilities were based on the genetic locations, which were deduced from physical locations using Marey maps (Fiston-Lavier et al. 2010). The probability of switching state between two consecutive SNPs is then given by qd, where d was the genetic distance between the two SNPs.

To avoid a high rate of false positives, it is important to choose a small enough value for q. A natural strategy is to simulate sequences under a neutral scenario with realistic demography and estimates for mutation and recombination. Based on such simulations, $q = q_{sim}$ can be chosen such that the probability of falsely detecting selection on any segment of a given length is controlled and kept below a certain threshold α, as already explained in the "Simulations" section. However, a simulated scenario will always involve some simplifications or biases compared with the real demography, and the real background scenario will usually be unknown. To account for this uncertainty, a conservative approach is to choose a value of q lower than q_{sim}. In the extreme case, if the simulated model were completely unrealistic, it would actually make sense to choose $q = 0$ so that no false positives will be obtained.

Even with a good estimation of the population demography, reliable neutral simulations are difficult to design and extremely time intensive, because they must account for the variation of recombination rate along the whole analyzed region (from 1 to 4 cM per Mb in our case, plus a large region with no recombination). Besides, in the specific case of Pool-Seq, the scaled recombination rate cannot be estimated from the data because haplotype information is not available. Consequently, we decided to choose q_{sim} using a very simple model and to select a value of q considerably lower than q_{sim}.

More specifically, we simulated neutral samples with length 100 kb under a model with constant population size. We chose $\theta = 0.003$, which is consistent with the AFS estimated from our data, and $\rho = \theta$ as suggested by the results of Haddrill et al. (2005) for non-African populations of *D. melanogaster*. The analysis of these samples using the folded AFS likelihood described in equation (6) led to $q_{sim} = 4.10^{-5}$. With this value, we only detected a sweep signal in 1% of the simulated samples with length 100 kb, which suggests only one false positive signal every 10 Mb. In order to take into account the discussed uncertainties about the true model, we decided to work with the value $q = 10^{-10}$ which is considerably below that obtained via simulations.

For a more detailed discussion concerning the choice of transition probabilities, see Boitard et al. (2009).

Results

Accuracy of the AFS Estimation

In order to investigate the accuracy of our AFS estimation procedure, we simulated reads from 100 pools of sequences

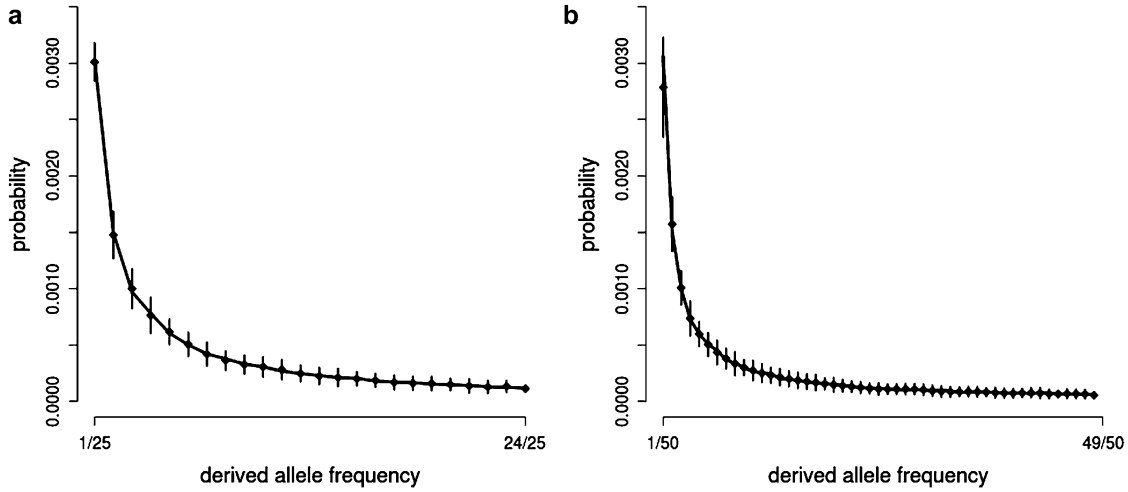

FIG. 1. AFS estimation. Pools of $n = 25$ (a) and $n = 50$ (b) chromosomes of length $L = 100$ kb were simulated under a constant population size coalescent model with $\theta = 0.003$ and $\rho = 0.003$. Solid lines show the AFS extracted from the complete sequence information and averaged over 100 simulated samples (it closely fits the AFS expected from coalescent theory). Diamonds and error bars represent the average estimated AFS and the average absolute deviation respectively using the same 100 samples. The estimates were obtained from pooled NGS data with 100× expected coverage using the EM algorithm.

under neutrality. We considered four different pool sizes ($n = 25$, 50, 100, and 200), took $\lambda = 100$ as the expected coverage, and $L = 100$ kb as sequence length. For further details, see the section on Materials and Methods. Under this setup, we compared the AFS estimated from pooled NGS data with our EM algorithm to the AFS computed under the assumption that the complete genetic information of the pool were available. As shown in figure 1 for $n = 25$ or 50, we found that our estimation procedure was essentially unbiased and had a small average absolute deviation. The main difference between the results obtained for $n = 25$ and $n = 50$ was a slight underestimation of the singleton frequency estimated when $n = 50$. This is likely due to the lower per chromosome coverage in this case, which implies that it is more difficult to decide whether observed singletons are true or result from sequencing errors. Results obtained for $n = 100$ and 200 have been very similar to those obtained with $n = 50$. Overall, our simulations show that accurate estimates of the AFS can be obtained from NGS data from pools with low per chromosome coverage (0.5×), despite of the sequencing errors. We also point out that the accuracy of the estimates will increase with sequence length, suggesting a very high accuracy when estimating the AFS at a whole genome scale. Notice, however, that the simulations were performed under the assumption that the probabilities of sequencing errors are known. As discussed below, inaccurate or biased error probabilities result in biased AFS estimates.

Selective Sweep Detection Power

Next, we simulated reads under a selective sweep scenario. The simulation parameters were the same as above, except that one selected locus with selection intensity $\alpha = 2\,N\,s = 500$ was placed in the middle of the 100 kb segment. This value of α corresponds to rather weak selection, compared with the distribution of selection intensities for sweeps identified in *D. melanogaster* (Li and Stephan 2006). For each simulated sample, we detected selection using either complete sequence data and the method in Boitard et al. (2009) or Pool-Seq data and the method presented here. As our new method extends that in Boitard et al. (2009), it should be of interest to compare the power of the two methods that make use of different amounts of information. For the analysis of the complete sequence data, we used either all sites or only segregating sites. Recall that the lower density of segregating sites (i.e., the larger probability of allele count 0) in swept regions is used as an additional sweep signal when using all sites.

As shown in table 1, the detection power with pooled data using only segregating sites was similar to that obtained with sequence data and all sites. At first sight, it might be surprising that the power was even slightly better with pooled samples. A closer look reveals, however, that the estimated sweep windows were usually slightly larger with pooled data than with classical sequencing data, and consequently had a higher probability of including the true selected site. The slight gain in detection power is thus associated with a slight loss in accuracy of localizing the sweep. Nevertheless, it is surprising that the results for pooled samples were considerably better than those for error-free classical separate sequencing when in both cases only segregating sites are used. A possible explanation is that with NGS sequencing data many singletons are sequencing errors at nonpolymorphic sites. Since a high proportion of singletons serves as a signal of selection such sequencing errors seem to increase the sensitivity of our test without causing an excess of false positives.

Application to Real Data in Drosophila

Using our new approach for Pool-Seq data, we estimated the AFS and searched for selective sweeps on the X chromosome of an Austrian *Drosophila melanogaster*

population. We analyzed a pool of 97 female flies from this population that has been sequenced at 100× coverage (for more details, see Materials and Methods). The sweep regions found with our scan are listed in table 2. Most of these regions were between 10- and 40-kb long, suggesting that hitchhiking mapping from Pool-Seq data identifies narrow intervals containing only a few genes that may have undergone recent selective sweeps. A few longer regions (up to 400 kb) were detected close to the centromere (supplementary fig. S1, Supplementary Material online). This is due to the fact that the recombination rate is very low close to the centromere, which increases the hitchhiking effect of positive selection.

Some of the detected regions were already identified by previous studies as sweep candidates in Europe. For instance, region 10 corresponds to the *wapl* region identified in Beisswanger et al. (2006). This region had a very high confidence index, within the top 10 of table 2. Interestingly, the size of the sweep window inferred by our method is similar to the one previously reported (Beisswanger et al. 2006) (74 vs. 60.5 kb). Region 16, which is located around the gene *unc-119*, was detected in Glinka et al. (2006).

Apart from these well-characterized regions, we also detected some narrow sweep windows with a high confidence score. The high confidence region 14 contains only a single gene, *Ca-alpha1T*, which is predicted to encode a Calcium channel (http://flybase.org). Four regions encompass only two annotated genes in *D. melanogaster*. One of them, region 28, contains the gene *Shaker (Sh)*, which encodes a voltage-dependent potassium channel and has been shown to affect sleeping behavior and lifespan (Cirelli et al. 2005).

The AFS estimated from our sample had an unusual pattern, showing a reduced proportion of extreme allele counts (fig. 2*a*). To investigate potential causes, we first considered the nonnegligible proportion of tri-allelic SNPs (about 1% of all sites). In our analysis, the least frequent allele of tri-allelic SNPs has been removed systematically (see Materials and Methods). We therefore reestimated the AFS after having removed all tri-allelic SNPs, but this resulted essentially in the same AFS pattern (data not shown). Hence tri-allelic SNPs cannot explain the observed deficit in low-frequency alleles. We also studied the influence of the coverage per site, by estimating the AFS using only positions with a specific coverage, and obtained again similar patterns. Varying the initial AFS that is used as starting point for the EM algorithm also had little influence on the finally estimated AFS, even when we started the algorithm from the AFS of an expanding population, which is characterized by an excess of small minor allele counts. We observed different estimated AFS patterns, however, when we restricted the analysis to base calls characterized by a specific range of PHRED scores. In particular, the restriction to base calls with PHRED score greater than 35 resulted in an estimated AFS with no deficit in extreme allele counts (fig. 2b), which is a reasonable neutral background AFS. Computer simulations show that the estimation bias observed when using all base calls is not caused by the low quality of many base calls in itself, but rather arises from a discrepancy between the PHRED scores provided by Illumina and the exact sequencing error probabilities. This observation is consistent with previous results (Dohm et al. 2008), showing that Illumina scores tend to be too pessimistic. The resulting overestimated probabilities for sequencing errors affected in particular those sites with low minor allele frequencies.

As an alternative to filtering with respect to quality scores, we recalibrated the quality scores using GATK (DePristo et al. 2011) before estimating the AFS. However, this correction had little effect on the AFS pattern. A reason for this might be that GATK uses monomorphic positions for the recalibration. Since we provided a list of SNPs with at least two copies of the minor allele in our sample, the remainder of the sequence contained singletons, which were a mixture of sequencing errors and true singletons. Our failure to distinguish true polymorphism from sequencing errors may have negatively affected our efforts to recalibrate the quality scores.

Since base calls with PHRED score greater than 35 provide a more reliable estimate of the AFS, we also performed a scan for selection using only these base calls, and compared the results with those obtained using all base calls. The signals obtained with the two strategies were generally consistent (fig. 3). Among the 32 sweeps detected with all base calls, 24 were confirmed using only high-quality base calls. The proportion of sweeps detected with both strategies increased with the confidence index. Among the 15 sweeps detected using all base calls with a confidence index greater than 20, 13 were confirmed using only high-quality base calls.

In order to see whether the sweep windows that were not confirmed when only using high-quality base calls are false positives or rather false negatives, we sequenced (at 87× coverage) a further independent sample of 97 flies from the same population. Due to the random sampling of different flies in the two pools and to the random differences of coverage and base quality scores along the genome inherent to NGS technology, we do not expect to find the same false positives in the two samples. The new sample provided a very similar estimated AFS (not shown), which suggests that no major experimental problem affected either of the samples. Furthermore, most sweep windows detected using all base calls were detected again using the second sample (29 over 32, see also supplementary fig. S2, Supplementary Material online). In particular, 7 of the 8 sweep windows that were not confirmed using only high-quality base calls were detected using the second sample, and can thus be seen as false negatives in the analysis focusing on high-quality base calls. This suggests that sweep detection based on all sites is more reliable than expected given the pronounced bias in global AFS estimation. A possible explanation for this consistency is that the bias in AFS estimation is homogeneous along the genome and does not affect the detection of true local variation in the AFS along the genome.

Table 2. Selective Sweeps Detected on Chromosome X in *Drosophila melanogaster*.

Region	Start[a]	End[a]	Length[a]	CI[b]	Genes within the Window
1	19	460	441	Inf	CG17636, RhoGAP1A, CG17707, SP71, CG3038, CG2995, cin, CG13377 CG13376, ewg, CG3777, CG13375, CG12470, Or1a, CG32816, y, ac, sc l(1)sc, pcl, ase, Cyp4g1, Exp6, CG13373, CG18275, CG32817, CG18166 CG3176, CG18273, CG3156, CG17896, CG17778, svr, arg, elav, CG4293, Appl
2	530	669	139	Inf	su(s), CG13367, Roc1a, Suv4-20, skpA, sdk, CG13362, CG13361, CG5254 CG5273, RpL22, fz3
3	1,046	1,144	98	Inf	eIF4E-7, CG34320, CG11378, CG11384, CG11379, CG14627, CG14626 CG11380, CG14625, CG11381, CG14624, CG11382, CG11398, CG3638 CG11403, A3-3
4	1,179	1,312	133	Inf	CG32812, DAAM, CG18091, fs(1)N, CG11409, CG11412, CG11418, Tsp2A CG12773, CG11417, png, CG14770, CG3056, SNF1A, CG3719, CG32813 CG11448, futsch
5	1,338	1,369	31	6.8	futsch, Gr2a, CG14785, CG14786, CG14787, l(1)G0431, O-fut2, CG14777 CG32808, CG14778, pck, CG14780, Rab27
6	1,373	1,408	35	33.7	CG14782, sta, Nmdar2, CG14795, CG32810
7	1,456	1,484	28	5.7	no gene
8	1,658	1,693	35	28.1	Adar, CG32806
9	1,728	1,809	81	33.8	CG14801, CG14812, deltaCOP, CG14814, MED18, CG14815, CG14803 CG14816, CG14804, CG14817, CG14805, CG14818, CG14806, trr mRpL16, arm, CG32803, CG32801, Edem1, mip130, CG17766
10	1,995	2,069	74	33.1	csw, ph-d, ph-p, CG3835, Pgd, bcn92, wapl, Cyp4d1, CG3630, CG3621 Cyp4d14
11	2,092	2,118	26	19.8	Mct1, CG18031, msta, Vinc, CG14052
12	3,662	3,681	19	12.7	Tlk
13	5,766	5,784	18	7.9	CG3033, mof, CG3016, CG16721
14	6,023	6,061	38	30.6	Ca-alpha1T
15	7,028	7,054	26	27.7	no gene
16	7,152	7,191	39	14.4	CG1958, CG1677, CG2059, unc-119
17	7,336	7,419	83	32.4	CG11368, CG32719
18	7,821	7,848	27	31.1	CG10777, CG10778, RpS14a, RpS14b, CG1530, l(1)G0193, CG1531, CG15332
19	10,358	10,383	25	16.3	CG17255, CG2889, CG2887, PPP4R2r, CG32687
20	11,371	11,407	36	31.3	Cyp4g15, CG1749, Spase25, CG33235, CG32666
21	11,441	11,499	58	32.4	CG32666, CG1572, PGRP-SA, RpII215, CG11699, l(1)G0237, CG11697 CG11696, e(y)2, CG11695, nod, CG1561, rho-4, CG2533
22	11,868	11,893	26	15	cac, gd, tsg, CG18130, fw
23	13,098	13,123	25	15	sno, REG, mew
24	14,937	14,953	16	6.6	hiw, CG5541
25	15,696	15,716	20	14.7	PGRP-LE, sd, CG8509
26	15,824	15,846	22	16.5	Ranbp16, Stim, CG8924, CG8928, CG15603, CG15604
27	17,743	17,764	21	17.1	CG15814, CG6506, CG32554, CG32557, CG6762, Arp8, CG6769, mnb
28	17,924	17,956	32	32.2	Sh, CG15373
29	18,539	18,559	20	13.2	l(1)G0003, CG6540, CG6617, Ing3, CG6659, fu, CG6696
30	19,455	19,479	24	17.5	Grip84, car, Tao-1, CG14218, CG14204
31	20,978	21,009	31	19.6	CG11566, stg1, unc, CG15445, CG34120
32	21,234	21,266	32	15.5	waw, bbx, slgA, Hlc, mst

[a] In kilobases, along the X chromosome.
[b] Confidence Index: Maximum of $-\log(1-q_i)$ over the window, where q_i is the posterior probability of hidden state "Selection."

Discussion

Our aim has been to provide a new statistical method for estimating the AFS and detecting selective sweeps that can be used with experimental setups where a sample of individuals is sequenced in a single pool. As argued in Futschik and Schlötterer (2010), this experimental design is a cost-effective alternative to sequencing of individuals for population genetic analysis based on allele frequencies. Often fairly large samples are sequenced at low individual coverage using this approach. The analysis of NGS data from pools leads to new challenges, and existing methods for classical sequencing cannot directly be applied. Obviously, the per site coverage should be taken into account, and sites with high coverage should be more influential than sites with low coverage. Also, sampling of reads from the pool leads to an additional level of randomness that needs to be considered.

A major methodological challenge for the analysis of NGS data at low coverage arises from sequencing errors, because such designs do not provide enough redundancy to distinguish sequencing errors reliably from true low-frequency variants. So far, most theoretical studies on the subject, for both individual sequencing and Pool-Seq, have considered a simple approach where sites with minor allele count/frequency below a given threshold are omitted (Achaz 2009; Jiang et al. 2009; Futschik and Schlötterer 2010; Lee et al. 2011). This strategy is also currently popular for population genetic studies based on Pool-Seq data, see for instance Rubin et al. (2010). In contrast, our method uses all sites, but accounts for the

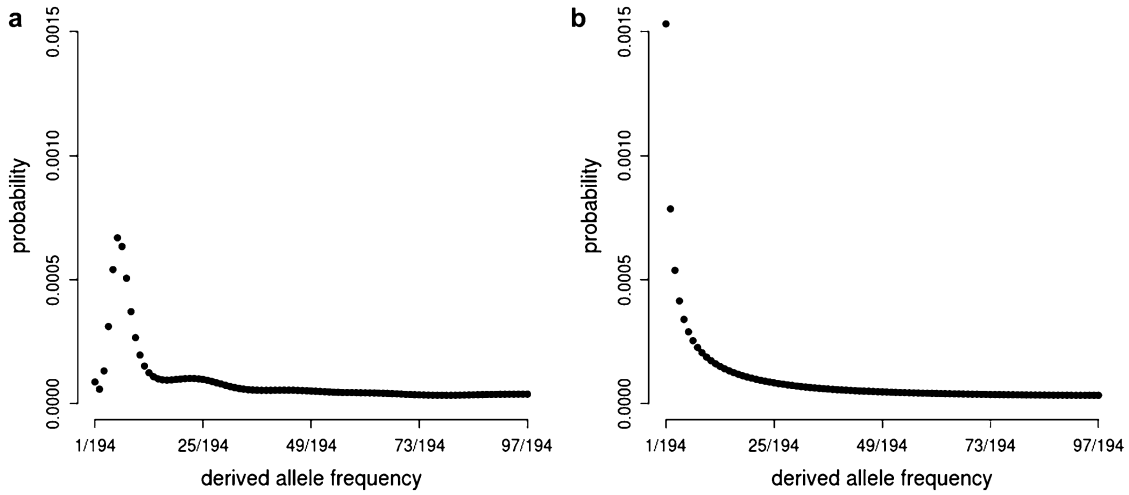

FIG. 2. AFS in *Drosophila melanogaster*. Estimated from all base calls (*a*) or only those with PHRED score greater than 35 (*b*). As we consider the folded AFS, the probabilities for allele frequencies 98/194 to 193/194 (not shown) can be deduced by symmetry from those for allele frequencies 1/194 to 96/194.

probability that a base call arises from a sequencing error. In principle, sequencing error probabilities could be deduced from the quality scores provided by the sequencing machine. It is known, however, that the Illumina PHRED scores, for instance, are biased. We discuss this point below.

First, we applied our method to simulated data, where we assumed accurate sequencing error probabilities. The obtained estimates of the AFS in pools from $n = 25$ to $n = 200$ chromosomes using Pool-Seq data at 100× expected coverage were not biased and highly accurate (fig. 1). This implies for instance that the frequency of singletons in a pool of 200 chromosomes can be reliably estimated using pooled NGS data at this coverage, despite of sequencing errors. We then evaluated, again for n from 25 to 200 and 100× expected coverage, the power of detecting a selective sweep event using pooled data. Our method provided very similar levels of power to that for individual sequencing of the entire pool (table 1). These promising results for sweep detection indicate that Pool-Seq data provide a rich source of information and may be suitable for

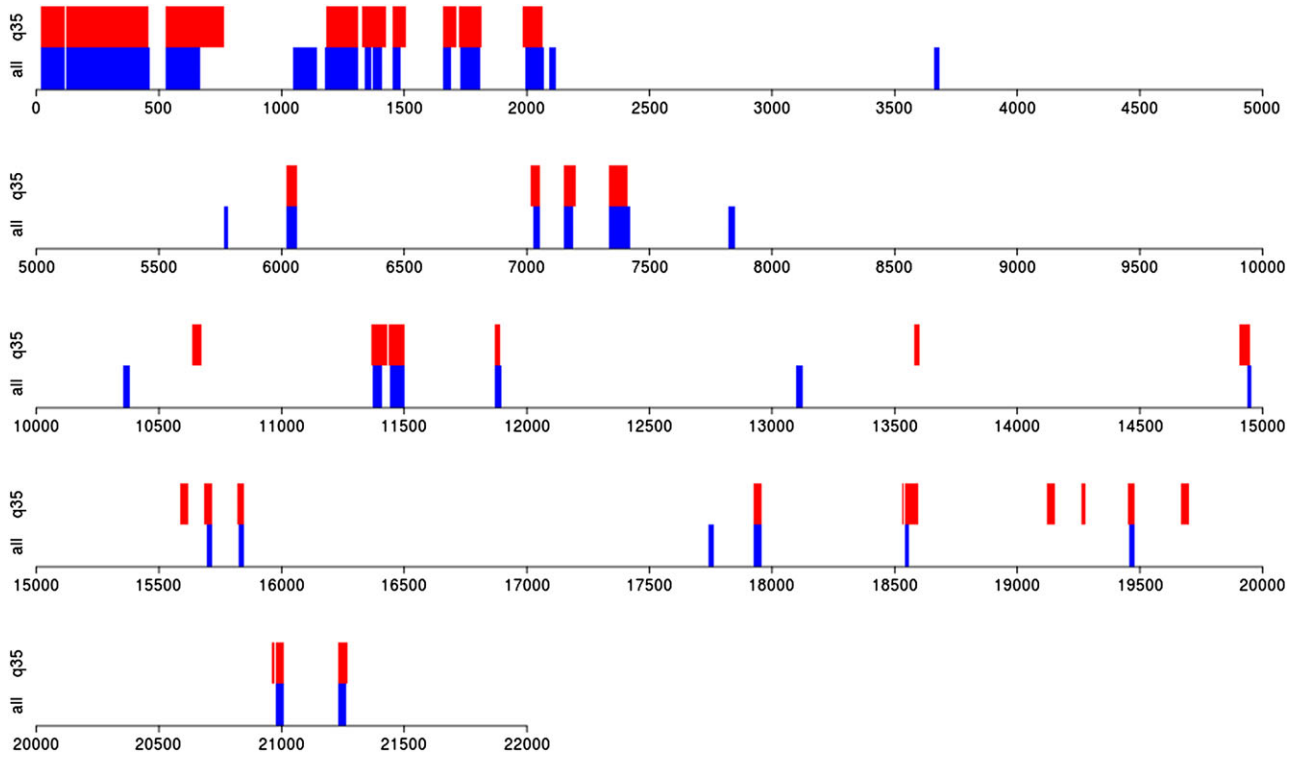

FIG. 3. Selective sweeps detected on the X chromosome of *D. melanogaster*. We used either all base calls or base calls with PHRED score greater than 35. The *x* axis labels permit to read off the physical position of the sweep window (in kilobases).

the inference of demographic scenarios such as population bottlenecks or expansions.

For applications to real data, the issue of inaccurate PHRED scores needs to be addressed. Unfortunately, no reliable approach on how to deal with biased quality scores in the context of Pool-Seq has been described so far. Although several models dealing with Pool-Seq data and including sequencing error probabilities have been proposed for SNP selection (Bansal 2010; Li 2011; Wei et al. 2011) and population genetic parameter estimation (Li 2011), only a few (Bansal 2010; Li 2011) take advantage of PHRED scores to determine these sequencing error probabilities. Nevertheless, they did not evaluate the influence of this strategy in the context of real reads and quality scores. When analyzing two Pool-Seq samples of *D. melanogaster*, we obtained underestimates of the probabilities of extreme allele counts. To spot potential biases in the estimates of sequencing error probabilities, we proposed to obtain repeated estimates of the AFS, by using base calls with different ranges of attached PHRED scores. If the estimated AFS are different and given a sufficient amount of reads for each individual estimate, this suggests that at least some of the obtained estimates are biased. Our results also indicate that recalibrating the PHRED scores using GATK or other similar software can be difficult, if the populations involved have not been extensively studied so that a large proportion of the SNPs in the genome is already known. For nonmodel organisms, another possible strategy might be to sequence individually a small number of individuals at high coverage in order to recalibrate the quality scores, and a large pool of individuals at low coverage for further analysis. Alternatively, one could include a known SNP-free DNA fragment in all sequencing runs and evaluate the sequencing error probabilities using this fragment, as done in Druley et al. (2009).

Fortunately, our sweep detection method turned out to be relatively insensitive to incorrect error probabilities. Indeed, we identified 32 selective sweep signatures, most of which were confirmed when using only high-quality base calls (PHRED score more than 35) and when analyzing an additional sample from the same population. One of the regions with the strongest evidence for selection was the *wapl* region, which was already identified as a sweep region in Europe (Beisswanger et al. 2006). A natural question is whether the signals our HMM is looking for, might have been caused by phenomena other than selective sweeps. One possibility are local fluctuations in the mutation parameter that may arise for instance from variable levels of purifying selection among codon positions or coding/noncoding sequences. Although we do not take the density of segregating sites as a signal by itself, sequencing errors will lead to an increase in the proportion of sites with low numbers of derived alleles in windows where θ is small, as observed in our simulations. This is due to the fact that the classification between sequencing errors and correct reads is not perfect. Notice, however, that the effect on the AFS will be small, when only very high quality reads are used. For our data analysis, it is therefore reassuring that most of our sweep signals were confirmed when using only the high-quality reads. If we assume that local stretches of sequence where the mutation rate is reduced tend to be short, another argument for the limited influence of sequencing errors would be that the sweep windows we detected on chromosome X of *D. melanogaster* tended to be fairly large.

If there is uncertainty about the homogeneity of the mutation rate at a larger scale, sweep detection (but not AFS estimation) can also be based on sites with at least k observed minor alleles, for instance with $k = 2$ or $k = 3$. Our method can easily be adapted for this purpose by replacing $\mathbb{P}(\sum_{j=1}^{r_i} Z_{i,j}=0)$ by $\mathbb{P}(\sum_{j=1}^{r_i} Z_{i,j}<k)$ in equation (5). Note, however, that the computation time will increase. Indeed, while there is only one vector Z_i verifying $\sum_{j=1}^{r_i} Z_{i,j}=0$, there are $r_i!/l!(r_i-l)!$ vectors verifying $\sum_{j=1}^{r_i} Z_{i,j}=l$, and the likelihood of all these vectors needs to be computed for l from 0 to $k-1$.

Like several other methods for the detection of selection, our approach is designed for hard sweeps with the favorable allele being fixed recently. Partial selective sweeps, as well as soft sweeps, will therefore usually not be detected. On the other hand, it is well known that some demographic effects, in particular bottlenecks, can produce similar genomic patterns as selective sweeps, potentially leading to false positives. In a previous study focusing on standard sequencing data (Boitard et al. 2009), we simulated a wide range of bottleneck scenarios and showed that the HMM method proposed for individual sequencing generally led to fewer false positive signals than several competing methods. The reason is that HMMs do not only use the site frequency spectrum but take into account also the correlation of allele frequencies between sites. As bottlenecks tend to increase the correlation between sites, we expect also the Hidden Markov Model proposed here to be more robust against bottlenecks than for instance composite likelihood methods. To put us further on the safe side, the sweeps detected in *D. melanogaster* were identified using the HMM with very conservative tuning parameters (see Materials and Methods).

Overall, our study shows that sequencing large pools of individuals at low coverage is a promising strategy for population genetic analyzes. Indeed, the method we proposed permits for cost effective and powerful scans for selection using this type of data. Its practical applicability is demonstrated by the selective sweep signals we identified in *D. melanogaster*. Alternative cost effective sequencing strategies, such as Restriction site Associated DNA sequencing (Hohenlohe et al. 2010) and Genotyping-by-Sequencing (Andolfatto et al. 2011; Elshire et al. 2011), have been proposed for population genetic studies based on large samples. These molecular methods generate individual low-coverage sequence data for a subset of the genome, thus providing individual genotypes at a large number of SNPs (typically from tens to hundreds of thousands). This represents a clear advantage over Pool-Seq for applications requiring haplotype information. However, the estimation of individual genotypes requires a minimum per individual coverage, at the very least 3× for calling homozygotes and

5× for calling heterozygotes (Hohenlohe et al. 2010). In contrast, our study demonstrates that a per individual coverage around 1× is sufficient in a Pool-Seq analysis. Of course, sequencing individuals at 1× or 2× coverage is also a reasonable strategy for allele or haplotype frequency estimation, provided the uncertainty about individual genotype calls is taken into account in the analyzes (Gompert et al. 2012). Although this experimental design was shown to be less efficient than Pool-Seq for allele frequency estimation (Futschik and Schlötterer 2010), it provides partial information about individual genotypes. Another advantage of Pool-Seq over Genotyping-by-Sequencing is to explore the whole genome rather than a subset of positions. In populations with low levels of linkage disequilibrium, an (almost) exhaustive screening of the genome certainly increases the power of scans for selection or association studies. For inference based on allele frequencies only, such as our method of detecting hard sweeps, we therefore believe that Pool-Seq is an attractive design.

Supplementary Material

Supplementary material and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has been financially supported by the Austrian Science Funds (FWF) grants P19832 and P22725. Travels between France and Austria were funded by the grant 25154QH from the Partenariat Hubert Curien Amadeus program.

References

Achaz G. 2009. Testing for neutrality in samples with sequencing errors. *Genetics* 179:1409–1424.

Andolfatto P, Davison D, Erezyilmaz D, Hu T, Mast J, Sunayama-Morita T, Stern D. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21:610–617.

Bansal V. 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 26:i318–i324.

Beisswanger S, Stephan W, De Lorenzo D. 2006. Evidence for a selective sweep in the wapl region of Drosophila melanogaster. *Genetics* 172:265–274.

Boitard S, Schlötterer C, Futschik A. 2009. Detecting selective sweeps: a new approach based on hidden Markov models. *Genetics* 181:1567–1578.

Cirelli C, Bushey D, Hill S, Huber R, Kreber R, Ganetzky B, Tononi G. 2005. Reduced sleep in Drosophila shaker mutants. *Nature* 434:1087–1092.

DePristo M, Banks E, Poplin R, et al. (18 co-authors). 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498.

Dohm J, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36:e105.

Druley T, Vallania F, Wegner D, et al. (12 co-authors). 2009. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods.* 6:263–265.

Elshire R, Glaubitz J, Sun Q, Poland J, Kawamoto K, Buckler E, Mitchell S. 2011. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLoS One* 6(5): e19379.

Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics* 26(16):2064–2065.

Fiston-Lavier A, Singh N, Lipatov M, Petrov D. 2010. Drosophila melanogaster recombination rate calculator. *Gene* 463:18–20.

Futschik A, Schlötterer C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* 186:207–218.

Glinka S, De Lorenzo D, Stephan W. 2006. Evidence of gene conversion associated with a selective sweep in Drosophila melanogaster. *Mol Biol Evol.* 23(10):1869–1878.

Gompert Z, Buerkle C. 2011. A hierarchical Bayesian model for next-generation population genomics. *Genetics* 187:903–917.

Gompert Z, Lucas L, Nice C, Fordyce J, Forister M, Buerkle C. 2012. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution* doi:10.1111/j.1558-5646.2012.01587.x.

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

Hohenlohe P, Bassham S, Etter P, Stiffler N, Johnson E, Cresko N. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. *PLoS Genet.* 6(2):e1000862.

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

Jiang R, Tavaré S, Marjoram P. 2009. Population genetic inference from resequencing data. *Genetics* 181:187–197.

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

Lee J, Choi M, Yan X, Lifton R, Zhao H. 2011. On optimal pooling designs to identify rare variants through massive resequencing. *Genet Epidemiol.* 35:139–147.

Li H. 2011. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

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

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

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptative substitution in drosophilia. *PLoS Genet.* 2(10):e166.

Nielsen R, Williamson L, Kim Y, Hubisz M, Clark A, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–1575.

Rubin C-J, Zody M, Eriksson J, et al. (19 co-authors). 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464:587–591.

Wei Z, Wang W, Hu P, Lyon G, Hakonarson H. 2011. Snver: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nuclear Acids Res.* 39(19):e132.

Williamson S, Hubisz M, Clark A, Payseur B, Bustamante C, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet.* 3(6):e90.