MOLECULAR ECOLOGY

Molecular Ecology (2013) doi: 10.1111/mec.12276

RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling

B. ARNOLD,[1] R. B. CORBETT-DETIG,[1] D. HARTL and K. BOMBLIES
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

Abstract

Reduced representation genome-sequencing approaches based on restriction digestion are enabling large-scale marker generation and facilitating genomic studies in a wide range of model and nonmodel systems. However, sampling chromosomes based on restriction digestion may introduce a bias in allele frequency estimation due to polymorphisms in restriction sites. To explore the effects of this nonrandom sampling and its sensitivity to different evolutionary parameters, we developed a coalescent-simulation framework to mimic the biased recovery of chromosomes in restriction-based short-read sequencing experiments (RADseq). We analysed simulated DNA sequence datasets and compared known values from simulations with those that would be estimated using a RADseq approach from the same samples. We compare these 'true' and 'estimated' values of commonly used summary statistics, π, θ_w, Tajima's D and F_{ST}. We show that loci with missing haplotypes have estimated summary statistic values that can deviate dramatically from true values and are also enriched for particular genealogical histories. These biases are sensitive to nonequilibrium demography, such as bottlenecks and population expansion. *In silico* digests with 102 completely sequenced *Drosophila melanogaster* genomes yielded results similar to our findings from coalescent simulations. Though the potential of RADseq for marker discovery and trait mapping in nonmodel systems remains undisputed, our results urge caution when applying this technique to make population genetic inferences.

Keywords: ascertainment bias, coalescent theory, population genomics, restriction-associated DNA sequencing

Received 21 September 2012; revision received 24 January 2013; accepted 25 January 2013

Introduction

High-throughput sequencing technology has revolutionized evolutionary genetics, enabling biologists to generate massive amounts of genomic data to address diverse questions in ecology and evolution. Importantly, new techniques allow high-throughput identification of variable sites [e.g. single nucleotide polymorphisms (SNPs)], even in species whose genomes are prohibitively large for sequencing or for which a reference genome is unavailable. In these situations, it is often preferable to eschew whole-genome sequencing in favour of a reduced representation approach that can be used to sample a fraction of the genome across many individuals at the same loci. A promising new technology, restriction-associated DNA (RADseq), is becoming popular for reducing genomic complexity in DNA libraries to sequence a small portion of the genome across many individuals (reviewed in Davey *et al.* 2011). Hundreds of indexed RAD libraries can be easily and inexpensively constructed and sequenced to characterize levels and patterns of genetic variation throughout the genome, even for nonmodel organisms. RADseq has already been employed in studies of population structure and biogeography (Emerson *et al.* 2010; Gompert *et al.* 2010; Hohenlohe *et al.* 2010), allele frequency estimation (Van Tassell *et al.* 2008), association studies (Parchman *et al.* 2012),

Correspondence: Kirsten Bomblies, Fax: (617) 495-9484;
E-mail: kbomblies@oeb.harvard.edu

[1]These authors contributed equally to this work.

genetic mapping (Baird *et al.* 2008; Andolfatto *et al.* 2011; Pfender *et al.* 2011), selection and introgression (Hohenlohe *et al.* 2011; Gompert *et al.* 2012), and linkage disequilibrium (Hohenlohe *et al.* 2012).

RADseq differs from other genome-sequencing approaches in that DNA fragments for construction of a library of sequences are generated by digesting genomes with a restriction enzyme, as opposed to random DNA shearing. Enzyme digestion results in nonrandom cleavage that ensures primarily the same regions are sampled across individuals. While powerful, the RADseq approach may be affected by numerous, largely uncharacterized biases. Potential problems arising from PCR bias in library construction, sequencing errors and inaccurate genotyping with lower sequencing depths have been recognized previously (Rokas & Abbot 2009), but these biases are expected to affect all re-sequencing projects. RADseq has an additional ascertainment bias whose effects have not been explored extensively: some recognition sequences will themselves be polymorphic, resulting in missing data for some chromosomes, and thus nonrandom sampling of lineages in a sample (Fig. 1).

How does nonrandomly missing data affect estimation of levels and patterns of genomic variation necessary for population genetic inference? Here, we address this question by developing a coalescent-simulation framework to mimic the biased recovery of haplotypes (hereafter genealogical bias) in RAD libraries. Our work is consistent with but extends beyond that of Gautier *et al.* (2012) who also studied how missing data bias estimates of expected heterozygosity and F_{ST}. We analyse our simulations with additional commonly used summary statistics (π, θ, Tajima's D, F_{ST} and the complete allele frequency spectrum) that are used to study demographic history and detect selection. We explore how RADseq affects genome-wide estimates of these statistics and how it impacts outlier analyses.

We show that RADseq nonrandomly subsamples the genome in two ways. First, within a locus, variants in a recognition sequence result in missing data and therefore truncate genealogies relative to the complete sample at these loci. This truncation results in underestimates of commonly used diversity statistics π and θ_w. Estimates of Tajima's D are also less accurate, but F_{ST} is relatively robust. Second, certain genealogies are more likely to result in missing haplotypes than others, such that RADseq samples a biased subset of all genealogies. For example, loci with intermediate amounts of missing data are more polymorphic than the simulation average and more likely have genealogies with deeper divergences. We show with *in silico* digests of 102 completely sequenced *Drosophila melanogaster* genomes that our coalescent simulations capture the major features of RADseq's genealogical bias. We discuss our findings and provide general guidelines for using RADseq for population genetic inference.

Methods

Coalescent simulations

We used Hudson's ms (Hudson 2002) to simulate 10 kb DNA fragments for 100 haploid individuals with different population mutation and recombination rates (i.e. $\theta = 4N_e\mu$ and $\rho = 4N_er$, with μ and *r* being the

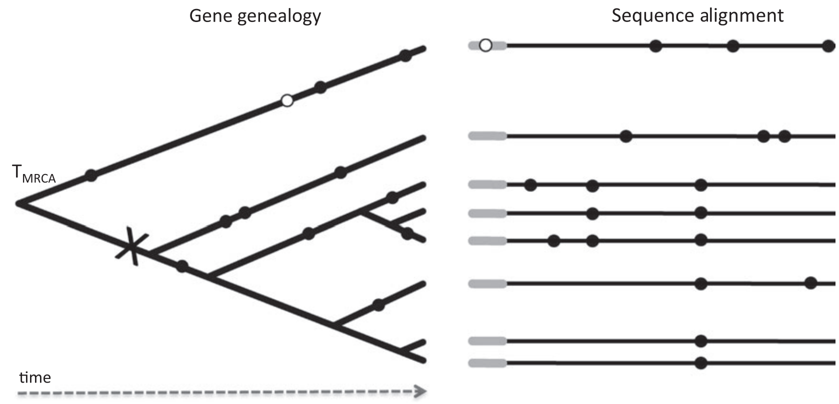

Fig. 1 An example of a DNA sequence alignment (horizontal lines at right) along with the underlying genealogy of the locus (left). Dots represent segregating mutations in the sequence and where in the genealogy they occurred. The wider grey portion of the sequence alignment represents the recognition sequence and a white dot indicates a mutation in the recognition sequence. Haplotypes are not observed in a recovery of chromosomes in restriction-based short-read sequencing experiment if mutations occur within this region. In this example, the true time to most recent common ancestor (T_{MRCA}) of the sample is lost since a mutation occurred within the recognition sequence in the most divergent haplotype; the genealogy is thus truncated to point 'X' and results in incomplete sampling that is biased against recovery of the most divergent haplotype(s).

mutation and recombination rates respectively). Three values of θ were used for simulations (0.0001, 0.001, or 0.01 per bp), with either $\rho = 0$ and $\theta = \rho$. We first simulated a single population at demographic equilibrium under each set of parameters above. To explore the effect of demographic history on RADseq, we modelled a bottleneck in which the population shrunk to 25% of the original size for 0.1 N_e generations, 0.1 N_e generations before present, after which it recovered to its original size. We also modelled an exponential growth scenario in which the population grows exponentially from 10% of its present-day size over 0.2 N_e generations. Simulations were repeated 100 000 times for each parameter set.

To explore the ability of RADseq to effectively detect population subdivision using a common metric of genetic differentiation (F_{ST}), we simulated two populations at demographic equilibrium that exchange migrants at a constant rate per generation. We simulated varying levels of population structure with 50 haploid individuals, or chromosomes, per population with migration rates (Nm) of 10, 1 or 0.1, and $\theta = \rho = 0.01$ per bp.

In silico RADseq experiment

Using custom Perl scripts, we performed an *in silico* digest by searching these simulated fragments for a specific recognition sequence. Since Hudson's ms (Hudson 2002) models DNA sequences with zeroes and ones, we used recognition sequences consisting of 12 zeroes and ones. Assuming equal nucleotide base composition, this motif occurs as frequently as a six-base DNA restriction enzyme site (about 2.8 times per 10 kb). Fragments that contained no recognition sequences were not analysed. After the *in silico* digest, we analysed the sequence 100 bp to the right of each recognition sequence to model the standard RADseq protocol (Baird *et al.* 2008). This length was chosen because it is currently a commonly used read length in Illumina sequencing. We compared 'true' summary statistic values (before digest) with 'estimated' ones (after digest, using only chromosomes that would have been recovered in a RADseq experiment). Here, we focus exclusively on biases induced by restriction site polymorphism, which ignores other potential sources of bias arising from sequencing and alignment, such as other sources of nonrandom sampling of haplotypes, sequencing errors and reference bias (reviewed in Rokas & Abbot 2009; Pool *et al.* 2010). These are expected to be general issues for most or all re-sequencing projects and are not addressed here.

Our simulation framework models the biased recovery of haplotypes in the RADseq protocol due to restriction site polymorphism. At a particular locus, a chromosome may not be sampled for two reasons: (i) a cut site, which is polymorphic in the population, is not present on that chromosome or (ii) a recognition sequence is present within 100 bp to the right of another recognition sequence, resulting in a fragment that is removed in the size-selection step and thus not sampled. As a result, the number of chromosomes sampled to the right of a particular recognition sequence, hereafter referred to as 'chromosome sampling depth,' varies among loci and may be less than the total 100 simulated DNA sequences. To demonstrate the effect of missing data due to the RADseq protocol, in the results below, we either binned loci by chromosome sampling depth or imposed cutoffs such that only loci with at least a minimum number of sampled chromosomes are analysed.

After the *in silico* digest of each fragment, we calculated the allele-frequency spectrum (AFS) for the 100 bp to the right of each recognition sequence using all simulated chromosomes (the 'true' AFS). We also calculated the AFS using only chromosomes that have the correct recognition sequence and would therefore be sampled by a RADseq protocol (the 'estimated' AFS). We then used these to calculate typical summaries of the data such as average number of pairwise differences (π, Tajima 1983), Watterson's θ (θ_w, Watterson 1975), Tajima's D (Tajima 1989) and F_{ST} (Weir & Cockerham 1984). As above, the true values for these summary statistics (π_t, $\theta_{wt,}$ D_t) were calculated using all chromosomes at the locus, and estimated values (π_e, $\theta_{we,}$ D_e) were calculated only for chromosomes that would be sampled in a RADseq experiment.

For the simulations with population subdivision, for any one locus, chromosomes are sampled according to criteria described above to mimic the RADseq protocol. F_{ST} can be inflated when one population has greater sampling depth, which may occur if a recognition-site mutation rises to a higher frequency in one population than the other, and this may confound inferences based on F_{ST}. Thus, for our analyses, we condition on sample sizes being the same for both populations to avoid these artefacts that inflate estimates of F_{ST}.

Double digest RADseq

We modified our framework to explore how summary statistics are affected by another RADseq protocol recently developed by Peterson *et al.* (2012), which relies on double digests. Briefly, this method requires first digesting the genome with two restriction enzymes and then selecting those fragments that fall within a defined size interval. We mimicked this process by sampling only fragments that were flanked by the same two complete recognition sequences of 6 zeroes and ones that were either within 150–250 or 350–450 bps of each other. The length of the restriction sequence was chosen to make the overall size of the mutational target associated with each chromosome at a locus the same as the standard RADseq protocol mentioned above. We further required that no additional cut sites be present in between that cause the fragment to be shorter than the selected size. We then sampled the 100 bp immediately adjacent to the left recognition sequence and analysed this as described above for the standard RADseq method. Although a double digest would normally involve sampling fragments flanked by two distinct recognition sequences, we only use a single recognition sequence for this *in silico* digest (repeated twice). However, since the sampling properties are the same for any arbitrary sequence of a specified length, we still refer to this modified framework as a 'double digest.' All analyses presented for the double digest protocol used the size selection with shorter fragments (150–250 bps) unless otherwise stated.

Empirical confirmation with Drosophila melanogaster

To confirm whether the predictions of our simulation framework reflect biases that could arise in an actual RADseq experiment, we performed *in silico* digests of 102 fully sequenced hemizygous (i.e. only one chromosome is sampled) *D. melanogaster* individuals (Pool *et al.* 2012). We acquired genome assemblies in fastq format from www.dpgp.org and subsequently translated these to fasta format requiring a minimum nominal base quality of 30. We masked regions of putative identity-by-descent, described in Pool *et al.* (2012), using the conversion/masking script provided by www.dpgp.org. We selected three different recognition sequences representing distinct base compositions, AseI (TAATTA), EcoRI (GAATTC) and EagI (GCCGGC), to digest the assemblies *in silico*. Digests were performed as described above for coalescent simulations mimicking the standard RADseq protocol (Baird *et al.* 2008). In brief, we digested each genome with a specific recognition sequence and considered that chromosome to be sampled if there was not an additional recognition sequence within 100 bp to the right. In cases where there was missing data in the recognition sequence (i.e. due to masked low-quality base calls, and not due to high-quality variants in the recognition site), we excluded those chromosomes from calculations of both true and estimated π. Each recognition site with at least one observed chromosome was considered for downstream analysis if at least 100 of the chromosomes in the original genome assemblies were covered by quality 30 or greater sequence through the entire region spanned by the recognition sequence.

Results

RADseq underestimates polymorphism

We generated simulated datasets for 100 haploid individuals and analysed them mimicking a RADseq protocol (see Methods for details). In comparing 'true' values of summary statistics (π_t, $\theta_{wt,}$ D_t) with 'estimated' values (π_e, $\theta_{we,}$ D_e, calculated from the data using only chromosomes that would be sampled by RADseq), it is apparent that the RADseq protocol results in systematic underestimation of polymorphism (Fig. 2A). Not surprisingly, increasing amounts of missing data exacerbates this bias, and there is a strong positive correlation between chromosome sampling depth and estimates of polymorphism (Fig. 2A). Fortunately, a majority of loci have all chromosomes sampled, especially for lower parameter values of θ in the simulations (Fig. 2B). We found that π_t is more sensitive to missing data than θ_{wt}. Recombination decreases this sensitivity and brings values of both π_t and θ_{wt} closer to the simulation parameter value of θ (Fig. S1, Supporting information). The difference between estimated and true values is greater for loci from simulated data sets with higher input values of θ (Fig. S2, Supporting information), though increasing the recombination rate tends to decrease this difference. This is because recombination decreases correlations between variants in the recognition sequence and those in the flanking sequence.

Simulations of the double-digest RADseq protocol (Peterson *et al.* 2012) produce similar results. However, relative to the standard RADseq protocol, loci that have higher chromosome sampling depths are less frequent in the double-digest protocol (Fig. S3, Supporting information) and have true and estimated values of π and θ_w that are even lower than simulation averages (Fig. S4, Supporting information). As in the standard RADseq protocol, a lower population mutation rate mitigates this effect (Fig. S5, Supporting information).

Although by definition, true and estimated summary statistics are identical when all chromosomes are sampled, loci with complete data still tend to have lower polymorphism than simulation averages (Table 1, Fig. 2, Fig. S2, Supporting information), particularly for the double-digest protocol (Table 1, Figs S4 and S5, Supporting information). This bias is exacerbated in the double-digest simulation when longer fragments were selected (350–450 instead of 150–250 bps). Thus, while completely sampled loci are not biased individually, they will not capture the true genome-wide distribution of values. For simulations with higher polymorphism and no recombination, estimates of means and variances are further reduced below true simulation averages.

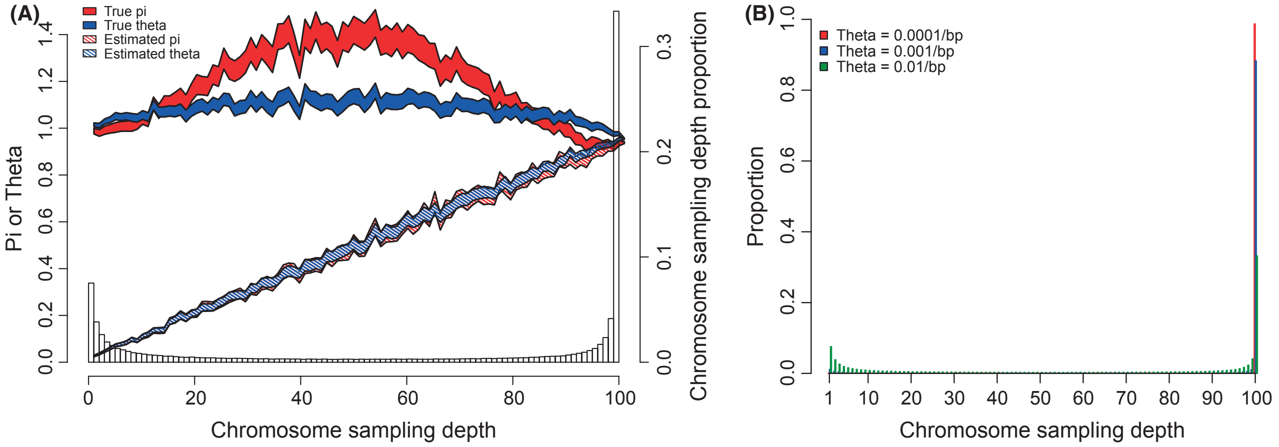

Fig. 2 (A) True and estimated values of π (red) and θ_w (blue) from *in silico* recovery of chromosomes in restriction-based short-read sequencing (RAD)seq as a function of chromosome sampling depth for θ = 0.01 per bp without recombination. Here, the simulation average of θ is 1 per 100 bp sequence read. Shaded regions show the 95% bootstrap percentile confidence intervals (1000 simulations) for the mean of true values of π (solid red) and θ_w (solid blue) and estimated values of π (shaded red) and θ_w (shaded blue) from *in silico* RADseq. 'Chromosome sampling depth' refers to the number of chromosomes that are actually sampled (have intact restriction sites) in the *in silico* experiment, and 'true' values are those calculated using the complete data for the same markers. The histograms in A (no recombination) and B (with recombination, ρ = θ) show the proportion of each chromosome sampling depth in the data and indicate that most markers are highly sampled with these simulation parameters, especially for lower values of θ (B).

Table 1 Comparison of estimated values of summary statistics (θ_{we} or π_e) when all chromosomes are sampled to true simulation averages (θ_{wa} or π_a)

		Mean				Variance			
		Recombination		No recombination		Recombination		No recombination	
Protocol	θ per bp	θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a
Standard	0.0001	0.994	0.995	0.991	0.990	0.994	0.996	0.990	0.990
	0.001	0.987	0.982	0.988	0.984	0.988	0.980	0.988	0.979
	0.01	0.956	0.933	0.940	0.909	0.941	0.901	0.904	0.837
Double digest	0.0001	0.835	0.836	0.838	0.837	0.836	0.836	0.839	0.836
	0.001	0.858	0.851	0.829	0.823	0.857	0.841	0.830	0.815
	0.01	0.829	0.797	0.811	0.772	0.812	0.737	0.771	0.684

Results from two different simulation parameters of θ are shown. When recombination is present, ρ = θ. Results are given for both the standard and double digest RADseq protocols.

Chromosome sampling depth is correlated with particular genealogies

Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the AFS. The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the 'true' genealogies (Fig. 3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Fig. 3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (e.g. Miyashita & Langley 1988; Hohenlohe *et al.* 2012; Langley *et al.* 2012; Pool *et al.* 2012). Lastly, in agreement with their higher values of π_t, loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (T_{MRCA}'s, not shown) relative to the simulation average.

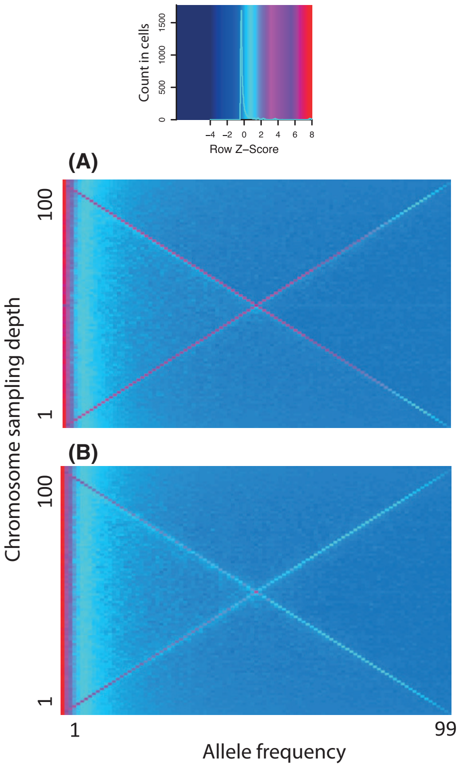

Fig. 3 Density plot of true allele frequency spectra (AFS) for loci with different chromosome sampling depths (A) without and (B) with recombination. Each row represents the AFS for a particular chromosome sampling depth with the density of a particular allele frequency indicated as a heat map. The *Z* score fits a normal distribution to the entries in each row, and each cell is coloured based on this fitting. This shows that loci with complete sampling (top of each graph) have an AFS characterized by abundant low-frequency polymorphisms, whereas loci with more missing data have greater proportions of intermediate frequency variants.

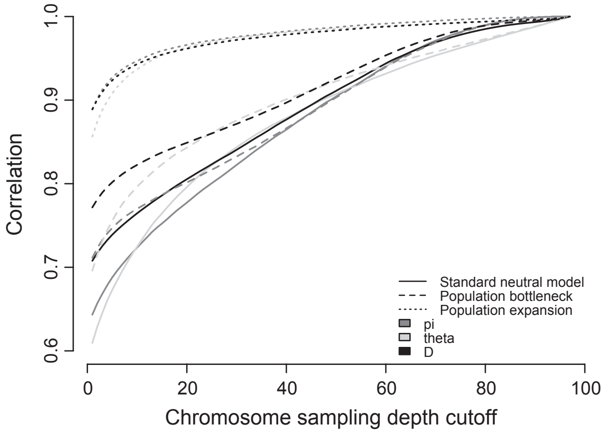

Fig. 4 Correlations between true and estimated values of summary statistics are sensitive to non-equilibrium demography and chromosome sampling depth cutoffs. Values for π (grey), θ_w (light grey) and Tajima's D (black) under different demographic models are plotted (solid lines = standard neutral model, dashed lines = bottleneck, dotted lines = population expansion). The *Y*-axis is the correlation between true and estimated values for loci that satisfy a given chromosome sampling depth cutoff (i.e. with at least a minimum number of chromosomes with intact recognition sequences).

Nonequilibrium demography and population subdivision affects true and estimated summary statistics

Nonequilibrium demographic processes can affect the AFS. Therefore, we asked what effect the introduction of a RADseq capture method can have on estimates of summary statistics for populations not at equilibrium. To this end, we simulated data under two commonly used demographic models: a population bottleneck and exponential growth. For the standard RADseq protocol, a population bottleneck followed by growth slightly decreases the effect missing data has on estimating true summary statistic values by slightly increasing the correlation between estimated and true values of θ_w and D (Fig. 4). However, bottlenecks have little effect on the estimates of π. Exponential population growth greatly reduces the sensitivity of π and θ_w to missing data and causes loci at all sampling depths to have estimated values of summary statistics that more closely resemble their true values (Fig. 4). Both of these scenarios mitigate the effect missing data has on estimation of summary statistics because effective population sizes are reduced (relative to an equilibrium population of equal present size), particularly for the exponential growth model.

A common goal of population genetic analyses is to detect and study population structure and differentiation. To explore the effects of RADseq on a common metric of genetic differentiation, F_{ST}, we simulated two populations at demographic equilibrium that exchange migrants at a constant rate per generation (described in Methods, performed only for the standard RADseq protocol). Unlike the results for metrics that summarize the AFS within a population, the distribution of estimated F_{ST} for loci with all chromosomes sampled is nearly identical to the true distribution (Fig. S6, Supporting information) for effective migration rates of $Nm = 10$ and $Nm = 1$. This strong concordance breaks down when populations exchange one migrant every 10 generations ($Nm = 0.1$; Fig. S6C, Supporting information). Importantly, including loci with missing data biases the estimated F_{ST} distribution, since missing data tends to inflate estimates of F_{ST} (Fig. 5). This is consistent with the results of Gautier *et al.* (2012) who considered biases of RADseq using a slightly different population subdivision demographic model.

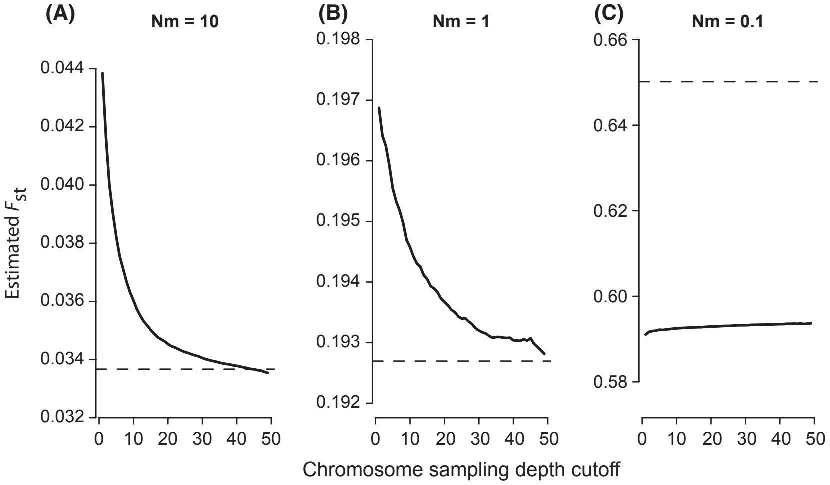

Fig. 5 Estimated F_{ST} (solid line) as a function of chromosome sampling depth cutoff per population (each consisting of 50 chromosome total) for three different migration rates: *Nm* = 10 (A), *Nm* = 1 (B) and *Nm* = 0.1 (C). The dashed line is the true simulation average. Here, we condition on sample sizes being the same in both populations to avoid inflated estimates of F_{ST}. Note that the *Y*-axes do not start at zero to more clearly illustrate differences between true and estimated values.

F_{ST}, θ_w, π and D outliers are sensitive to missing data

Although the levels and patterns of genetic variation in neutral loci that are linked to locally adapted alleles will depend on demographic and selective circumstances, it is interesting to consider outliers in the distributions of summary statistics as potential metrics for detecting positive selection and local adaptation. In particular, high F_{ST} may indicate that a locus is in linkage disequilibrium with locally adapted alleles. However, we show that missing data may inflate F_{ST} values, and rates of false positives quickly increase as the chromosome sampling depth cutoff decreases, especially when chromosome sample sizes amongst populations are allowed to vary as little as 20% (Fig. 6). Thus, it may be wise to constrain analyses to loci with complete chromosome sampling, but of loci in the upper 5% tail of true F_{ST} distribution, only 13%, 11% and 5% have complete chromosome sampling in both populations for *Nm* = 10, 1 and 0.1 respectively.

Within a population, genomic regions with low nucleotide diversity and left-skewed site frequency spectra may indicate the presence of a recent selective sweep via the hitchhiking effect (Maynard-Smith & Haigh 1974), or strong purifying selection (Charlesworth *et al.* 1993). We explored the effect of missing data on outlier analyses involving the commonly used diversity statistics θ_w and π. Specifically, we examined the lower 5% tail of the distributions of these statistics to assess how missing data affects false positive and false negative rates. Using different sampling depth cutoffs, rates of false positives and false negatives increase with the inclusion of loci with missing data for both the standard RADseq protocol (Fig. 7) and the double-digest protocol (Fig. S7, Supporting information). Similar analyses with lower values of θ (0.001 per bp and lower) were not possible since the 5% quantile of summary statistics contained the majority of loci due to low levels of polymorphism.

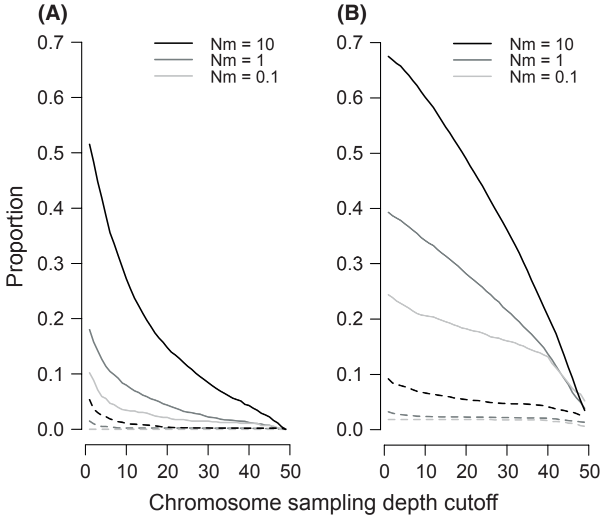

Fig. 6 Proportion of estimated F_{ST} 5% outlier loci that are false positives (solid lines) or false negatives (dashed lines) relative to the true distribution for different chromosome sampling depth cutoffs (50 chromosomes per population in complete sampling). Three different rates of migration are represented: *Nm* = 0.1 (light gray), *Nm* = 1 (gray), and *Nm* = 10 (black). If missing data are present, analyses were performed on loci for which chromosome sample sizes are exactly the same in both populations (A) or allowed to vary by 20% (B).

Since loci with missing data have more false positives and negatives, a possible solution is to limit outlier analyses to loci with complete chromosome sampling. If outliers were evenly distributed across loci irrespective of missing data, 5% of loci in each sampling depth category would be outliers. However, in agreement with the results presented in Table 1, loci with complete sampling have slightly decreased diversity and are more likely to fall within the lower 5% tail of the true distribution of π and θ_w and less likely to fall within the upper 5% tail (Table 2). Thus, limiting analyses to completely sampled sites may inadvertently enrich for loci that have experienced recent positive selection or are highly constrained by strong purifying selection.

In silico digestion of Drosophila melanogaster genomes

To test whether our framework captures the major biases associated with RADseq, we performed *in silico* digests of 102 recently released *D. melanogaster* genome assemblies (Pool *et al.* 2012) using the standard RADseq protocol (Baird *et al.* 2008). The choice of restriction enzyme greatly affects which features of the genome are sampled (Fig. 8A). The GC-rich recognition sequence of EagI samples exons more frequently than loci sampled at random and much more frequently than the AT-rich AseI, which disproportionately samples intronic and intergenic regions. EcoRI, which has an intermediate base composition, samples genomic regions at frequencies similar to their abundance in the genome. It is likely that owing to different levels of polymorphism in different parts of the genome (e.g. due to stronger purifying selection in exonic vs. intergenic sequences), choice of restriction enzyme results in different estimates of nucleotide diversity (Fig. 8B).

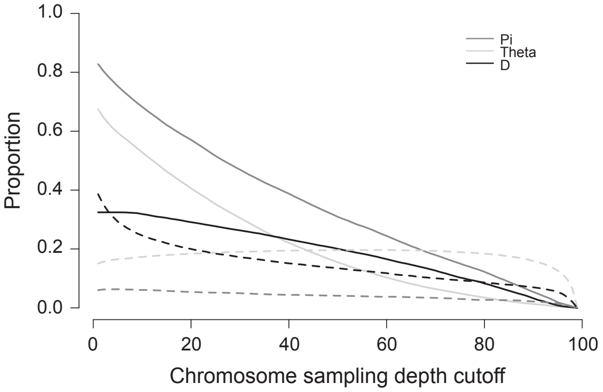

Fig. 7 Proportion of estimated π (grey), θ_w (light grey) or D (black) outliers that are false positives (solid lines) or false negatives (dashed lines) for inclusion in the lower 5% tail for different chromosome sampling depth cutoffs.

Table 2 Loci with complete sampling are more likely to fall within the lower 5% tail of the true distribution of π and θ_w

Protocol	Tail	Ratio	
Standard	Lower tail	1.19	1.17
	Upper tail	0.76	0.74
Double digest	Lower tail	1.76	1.95
	Upper tail	0.45	0.37

Shown are the ratios of the proportion of loci with complete chromosome sampling depth that are true outliers to the proportion of true outliers in the entire simulated data set.

Similar to the simulation results, in regions of the genome where π_t is higher, it is more common for an intermediate number of chromosomes to be sampled (Fig. 8C), which is consistent with the results of our simulations (above). This difference in π_t between loci with different chromosome sampling depths changes depending on the recognition sequence of the restriction enzyme used and increases as more polymorphic regions of the *D. melanogaster* genome are sampled (i.e. with an enzyme with an AT-rich recognition sequence). Again, we observe similar patterns in our simulation framework (above), suggesting that our simulations accurately reflect much of the bias associated with RADseq.

To compare our framework to the *D. melanogaster* data, we ran simulations with an increased recombination rate ($\rho = 0.1$ per bp, $\theta = 0.01$ per bp); $\rho = 10*\theta$ has been used previously in demographic inference of this species (e.g. Thornton 2009). We then recorded the median of the ratio expected to true π (π_e/π_t) for each locus with a particular number of sampled chromosomes (Fig. 8D). While our simulation appears to accurately model the majority of genealogical bias, we did not perfectly capture the dynamics of loci that have <10 sampled chromosomes, perhaps as a result of violations of the infinite sites mutation-model (see Discussion).

Discussion

RADseq provides a simple and inexpensive means of collecting genome-wide sequence data from diverse nonmodel organisms (e.g. Emerson *et al.* 2010; Hohenlohe *et al.* 2011; Gompert *et al.* 2012; Parchman *et al.* 2012). This approach is increasing in popularity as a means of population genomic inference, but the effects of the ascertainment bias associated with polymorphism in recognition sites have not been extensively explored (but see Gautier *et al.* 2012). Biases can arise from mutations segregating in the recognition sequence such that haplotypes are nonrandomly sampled for loci linked to these polymorphisms. Though it may seem comparatively rare for a mutation to occur within a recognition sequence, these variants are frequent enough to enable detailed population genetic analyses (e.g. Restriction Fragment Length Polymorphisms, Botstein *et al.* 1980). Consequently, a thorough examination of RADseq bias is essential for enabling detailed and accurate population genetic analyses based on this methodology.

Our coalescent simulations model two separate RADseq protocols (Baird *et al.* 2008; Peterson *et al.* 2012) and

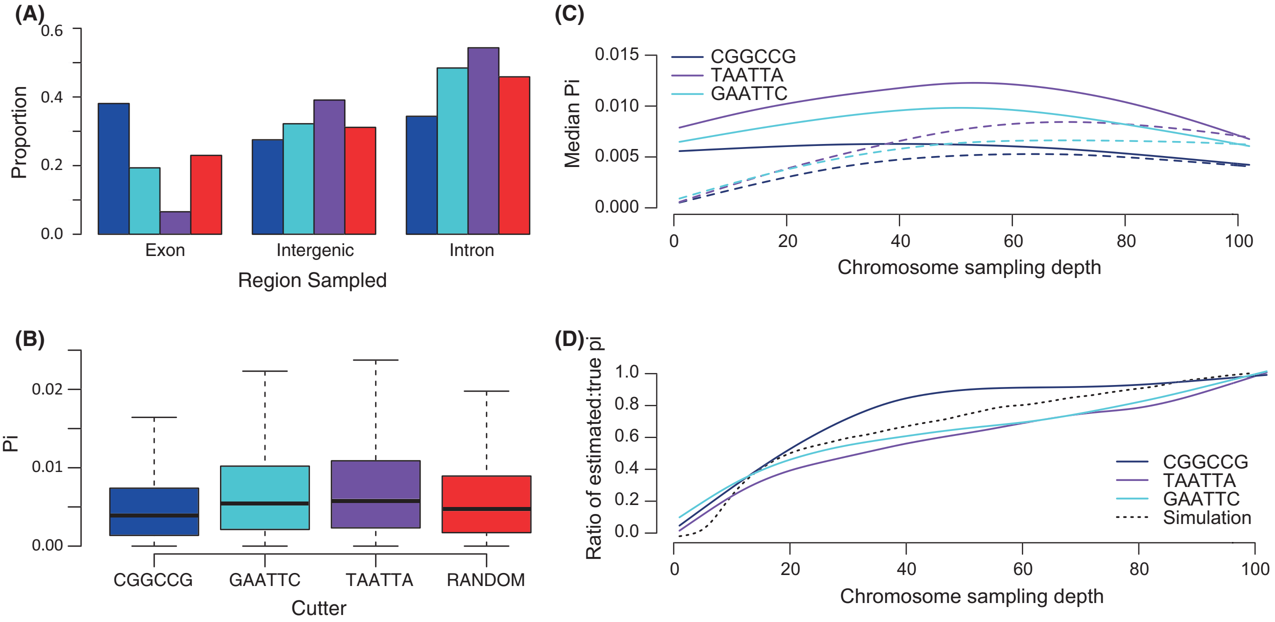

Fig. 8 Results for the *in silico* digests 102 *Drosophila melanogaster* genomes. (A) Proportion of sites located in distinct regions of the genome when *in silico* digests are performed with different enzyme recognition sequences. GC-rich recognition sequences sample more exons, whereas AT-rich recognition sequences sample comparatively more introns and intergenic regions. 'Random' values are calculated from fragments selected at random throughout the genome. (B) Box plots of true π for regions sampled by enzymes with different recognition sequences. (C) The median true π (solid line) and estimated π (dashed line) as a function of chromosome sampling depth for three different recognition sequences. (D) Median of the ratio of estimated π to true π as a function of the number of sampled chromosomes. Dark blue, purple and cyan lines represent the three different restriction enzymes used in the *in silico* digest of the *D. melanogaster* genomes, and the dotted black line is from simulations with $\rho = 0.1$ per bp, $\theta = 0.01$ per bp

show that in both cases, true and estimated values of π and θ_w vary with the amount of missing data that would occur in a RADseq experiment. Loci with higher π_t and θ_{wt} generally have fewer sampled chromosomes. These loci also have distinct frequency spectra and deeper divergence times. These patterns indicate that certain genealogies are particularly prone to missing data in RADseq experiments. Both π_e and θ_{we} and their correlations with true values decrease systematically as a function of the chromosome sampling depth, making loci with higher diversity the most strongly underestimated. Tajima's D is also sensitive to missing data. One potential solution might be to limit analysis to loci for which one can be certain of complete sampling. However, while this will reduce bias from sampling particular branches of the genealogy, it is important to remember that loci where RADseq samples all chromosomes are also a nonrandom subset of genome-wide π and θ_w distributions. Underestimated polymorphism has been previously observed in RADseq but was attributed to conservative SNP calling (Hohenlohe *et al.* 2010).

Loci with complete sampling for the double-digest protocol have further decreased estimates of diversity (compared to the true genome-wide estimate) than the standard protocol because missing data may arise not only from mutations within recognition sequences but also from novel restriction sites that cause some haplotypes to be outside of the size-selection range. Indeed, this problem is exacerbated for the simulation in which longer fragments were selected since there is a larger region within which novel restriction sites may occur. In reality, segregating insertions or deletions may also contribute to missing data by changing the length of sequences between cut sites to outside the range of size selection, but this additional source of bias was not modelled in this study.

Importantly, inclusion of loci with incomplete sampling may actually invert relative estimates of π and θ, such that loci that are in reality more diverse will have lower estimates for these parameters than loci with complete sampling that are taken from less diverse regions. In practice, for a particular locus with incomplete chromosome sapling depth, it may not be feasible to determine if chromosomes were not sequenced from polymorphism in the restriction site or from low sequencing depth.

The correlations between estimated and true values of summary statistics are also sensitive to nonequilibrium demography. Both population bottlenecks and expansions increase correlations between true and estimated values. The greater correlations presumably occur because both demographic scenarios decrease the effective population size and therefore reduce genetic diversity, so fewer loci have missing data and thus inaccurate estimates of summary statistics. Since natural populations are likely to have complex evolutionary histories, summary statistics may be affected by a combination of multiple demographic events in addition to the population mutation and recombination rates. Having estimates of these parameters a priori for a given study system help predict how frequently loci will contain missing data and how sensitive estimated values of summary statistics are to missing data.

We also explored the ability of RADseq datasets to detect population structure and differentiation by calculating F_{ST} between two populations at demographic equilibrium that exchange migrants at a constant rate per generation.

The distribution of estimated F_{ST} values for loci with all chromosomes sampled is very similar to the true distribution. The relative robustness of F_{ST} to the RADseq protocol suggests that this methodology is perhaps well suited to estimating rates of migration between populations.

Since outliers in the distributions of summary statistics are frequently used as metrics for detecting selection, we explored the sensitivity of F_{ST}, θ_W, and π outliers to missing data. We find that rates of false positives and false negatives increase for F_{ST}, θ_W and π as chromosome sampling depth decreases, since missing data biases estimates. This has important implications for outlier analyses as tests for selection or local differentiation and indicates that empirical outliers obtained from RADseq experiments where complete chromosome sampling cannot be established with certainty should be interpreted with caution. Again, a potential solution is to restrict analyses to loci with complete chromosome sampling depth, but with this correction, a vast majority of true F_{ST} outliers would be missed since many true outliers have incomplete sampling. Moreover, since many investigators sequence diploid organisms, it may be difficult to quantify the amount of missing data and the sample size variation amongst populations, both of which would inflate estimated F_{ST} values.

Our *in silico* RADseq analyses of 102 *Drosophila melanogaster* genomes were largely consistent with the results of our simulations, in that polymorphism is underestimated, especially for more diverse genomic regions. Although undoubtedly the populations from which these samples are derived are experiencing nonequilibrium selective and demographic processes that we did not model (Corbett-Detig & Hartl 2012; Pool *et al.* 2012), the overall congruence of our simulations with the *Drosophila* data suggests that our basic simulation framework captures the major biases that affect RADseq. One possible explanation of the poor fit of our model at low chromosome sampling depths is that the real data includes violations of the infinite-sites mutation model, such that mutations recur within nascent recognition sequences on different haplotypes. This would effectively inflate diversity relative to infinite-sites assumptions of the coalescent simulations. Nonetheless, it is clear that even though our simulations are relatively simplistic, we have identified a major potential bias inherent to the RADseq methodology.

The nucleotide composition of the recognition sequence affects which features of the genome are sampled and this suggests an appealing means of tuning RADseq for the specific goals of each respective study. For example, for the purpose of SNP discovery, one may prefer to select an enzyme with an AT-rich recognition sequence; conversely, if the goal is to study genetic differentiation between divergent populations, GC-rich recognition sequences will generally access a higher proportion of conserved regions of the genome and may increase the overlap in sampled loci between populations. However, such choices must still be considered with appropriate caveats, for instance, in species with DNA-methylation, CG sites are known to mutate at significantly higher rates than the genomic average (Cooper *et al.* 1995). In this case, using an enzyme which cuts sequences that contain these motifs may increase the amount of missing data, and violate a tacit assumption of our model that the per-site mutation rate in the recognition sequence is identical to that in the sequenced read. We thus emphasize that because each restriction enzyme will access different genomic regions, which may not have identical allele frequency spectra, the choice of restriction enzyme will also affect population genetic inferences.

Our results are also consistent with those of Gautier *et al.* (2012), but our interpretation of how RADseq affects estimates of diversity is different. In their study, Gautier *et al.* state that RADseq results in overestimates of heterozygosity because they only consider segregating sites that are observed after the *in silico* digest of simulated fragments. This effect occurs because mutations in linked recognition sequences are more to likely arise on the major allele haplotype, thus inflating minor allele frequencies and estimates of heterozygosity. Here, we examine the effect that RADseq has on commonly used diversity statistics per site and thus account for both observed and unobserved segregating sites. Because variants in a recognition sequence truncate genealogies relative to the complete sample at these loci, some true variants are not observed, overall resulting in underestimates of π and θ_W.

RADseq is an important emerging methodology, and is likely to see increased use; it is therefore important to identify biases and where possible to develop a means of accounting for them. In general, the assumption of identical per-site mutation rates in the cut-site and sequenced read is likely to be reasonable. Given this assumption, it may be possible to account for the genealogical bias of RADseq using our (or a similar) coalescent simulation modification framework. That is, standard coalescent simulations can be performed, and the resulting sequence digested and analysed as we describe. If the resulting biased summary statistics are then compared with empirically obtained RADseq summary statistics (e.g. using approximate Bayesian computation software such as ABCreg; Thornton 2009), it may be possible both to directly account for this source of bias in population genetic analyses and to recover unbiased estimates of the true distributions of relevant summary statistics.

This study can serve as a useful guide for investigators using RADseq for population-genomic analyses. From our simulations and empirical *in silico* digests, loci with missing data give inaccurate estimates of summary statistics and may increase the rate of false positives in outlier analyses. Thus, identifying and pruning loci with incomplete sampling will be important in any RADseq experiment aimed at accurately estimating commonly used summary statistics. Since RADseq will generally produce thousands or tens of thousands of markers throughout the genome, pruned datasets that retain only loci with complete sampling will still be substantial (Fig. 2). However, if RADseq is to be used for demographic inference, it remains important to recognize that ignoring loci with missing data, which are enriched for particular genealogical structures, will also affect estimation of evolutionary parameters and may not accurately represent a 'genome average' value. If many loci have high sequencing depths such that sites with missing data can easily be detected by differences in coverage, RADseq provides a powerful way to estimate genome-wide divergence amongst populations to describe biogeographic patterns. Thus, though our findings urge caution, with careful consideration of experimental design, data use and interpretation, RADseq will likely continue to develop as a powerful technique for addressing questions in evolutionary biology.

Acknowledgements

We thank John Wakeley for insightful suggestions about designing the simulation framework and analyzing the genealogical bias. We also thank sources of funding (NSF DDIG to BJA and Harvard Prize Fellowship to RBCD).

References

Andolfatto P, Davison D, Erezyilmaz D *et al.* (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, E3376.

Botstein DR, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314–331.

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

Cooper DN, Antonarakis SE, Krawczak M (1995) The nature and mechanisms of human gene mutation. In: *The Metabolic and Molecular Bases of Inherited Disease*, 7th edn (eds Scriver CR, Beaudet AL, Sly WS, Valle D), pp. 259–291. McGraw-Hill, New York City, New York.

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

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, **107**, 16196–16200.

Gautier M, Gharbi K, Cezard T *et al.* (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, AOP.

Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle CA (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, **19**, 2455–2473.

Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly. *Evolution*, **66–7**, 2167–2181.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology*, **11**, 117–122.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **367**, 395–408.

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

Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.

Maynard-Smith J, Haigh J (1974) The hitchhiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

Miyashita N, Langley CH (1988) Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. *Genetics*, **120**, 199–212.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and nonmodel species. *PLoS ONE*, **7**, e37135.

Pfender W, Saha M, Johnson E, Slabaugh M (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *TAG. Theoretical and Applied Genetics.*, **122**, 1467–1480.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Pool JE, Corbett-Detig RB, Sugino RP, *et al.* (2012) Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genetics*, in press.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution*, **24**, 192–200.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

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

Thornton KR (2009) Automating approximate Bayesian computation by local linear regression. *BMC Genetics*, **10**, 35.

Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

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

B.A. and R.C.D. designed simulations, performed analyses and wrote the manuscript. D.H. and K.B. wrote the manuscript, helped with the study design and offered suggestions on the analyses.

Data accessibility

The scripts used for our simulations are available on the Bomblies Lab website (http://www.oeb.harvard.edu/faculty/bomblies/resources.html) and Dryad doi: 10.5061/dryad.c9s0 g.

Supporting information

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

Fig. S1 True and estimated values of π and θ_w as a function of the chromosome sampling depth for $\theta = \rho = 0.01$ per bp.

Fig. S2 Mean values of π_e (solid red) and θ_e (solid blue) for loci with different sampling depth cutoffs (i.e. loci that have at least the specified number of sampled chromosomes with intact recognition sequences).

Fig. S3 Chromosome sampling depth proportions for the standard (red) and double digest (blue) RADseq protocols.

Fig. S4 True π (solid lines) and estimated π (dashed lines) vary as a function of chromosome sampling depth for the standard (blue) and double-digest (red) RADseq protocols.

Fig. S5 Mean values of π_e (solid red) and θ_e (solid blue) for loci with different sampling depth cutoffs (CSDC, i.e. loci that have at least the specified number of sampled chromosomes with intact recognition sequences).

Fig. S6 Distribution of estimated F_{ST} when all haplotypes are sampled (blue) vs. the true distribution (black), for $Nm = 10$ (A), $Nm = 1$ (B), and $Nm = 0.1$ (C).

Fig. S7 Proportion of estimated π (red), θ_w (blue), or D (purple) 5% outlier loci that are false positives (solid lines) or false negatives (dashed lines) relative to the true distribution for different chromosome sampling depth cutoffs.