MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3165–3178 doi: 10.1111/mec.12089

# The effect of RAD allele dropout on the estimation of genetic variation within and between populations

MATHIEU GAUTIER,* KARIM GHARBI,† TIMOTHEE CEZARD,† JULIEN FOUCAUD,* CAROLE KERDELHUÉ,* PIERRE PUDLO,*‡ JEAN-MARIE CORNUET* and ARNAUD ESTOUP*

**Inra, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988, Montferrier-sur-Lez, France, †The GenePool, School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, UK, ‡I3M, UMR CNRS 5149, Université Montpellier 2, F-34095 Montpellier, France*

## Abstract

**Inexpensive short-read sequencing technologies applied to reduced representation genomes is revolutionizing genetic research, especially population genetics analysis, by allowing the genotyping of massive numbers of single-nucleotide polymorphisms (SNP) for large numbers of individuals and populations. Restriction site–associated DNA (RAD) sequencing is a recent technique based on the characterization of genomic regions flanking restriction sites. One of its potential drawbacks is the presence of polymorphism within the restriction site, which makes it impossible to observe the associated SNP allele (i.e. allele dropout, ADO). To investigate the effect of ADO on genetic variation estimated from RAD markers, we first mathematically derived measures of the effect of ADO on allele frequencies as a function of different parameters within a single population. We then used RAD data sets simulated using a coalescence model to investigate the magnitude of biases induced by ADO on the estimation of expected heterozygosity and $F_{ST}$ under a simple demographic model of divergence between two populations. We found that ADO tends to overestimate genetic variation both within and between populations. Assuming a mutation rate per nucleotide between $10^{-9}$ and $10^{-8}$, this bias remained low for most studied combinations of divergence time and effective population size, except for large effective population sizes. Averaging $F_{ST}$ values over multiple SNPs, for example, by sliding window analysis, did not correct ADO biases. We briefly discuss possible solutions to filter the most problematic cases of ADO using read coverage to detect markers with a large excess of null alleles.**

*Keywords*: allele dropout, allele frequency, $F_{ST}$, heterozygosity, next-generation sequencing, RAD markers, single-nucleotide polymorphisms

*Received 12 July 2012; revision received 4 September 2012; accepted 12 September 2012*

## Introduction

The advent of next-generation sequencing (NGS) technologies is dramatically changing our understanding of genetic variation in a growing number of species (e.g. Davey *et al.* 2012). Yet, whole genome individual (re) sequencing is still limited to a relatively small number of species for which a complete genome assembly is available and remains prohibitive for the large numbers of individuals needed in most population genetics studies. A potential answer to this challenge is in alternative approaches relying on whole genome sequencing of pools of individual DNAs (Futschik & Schlötterer 2010) or sequencing of reduced representation genomic libraries (RRL, Van Tassell *et al.* 2008; Luca *et al.* 2011). These approaches have recently proved powerful to survey a large number of markers on a genome-wide scale, with considerably lower library construction and sequencing efforts. Deep sequencing of restriction site associated DNA (RAD-seq) borrows from the RRL strategy (Baird *et al.* 2008). The technique is increasingly popular, especially in nonmodel species (Davey & Blaxter 2011; Davey *et al.* 2012), with successful applications

Correspondence: Mathieu Gautier, Fax: +33 (0)4 99 62 33 45; E-mail: mathieu.gautier@supagro.inra.fr

across a wide range of organisms and questions ranging from genetic map construction and mapping of genes underlying traits of interest (Baird *et al.* 2008; Baxter *et al.* 2011) to population genomics (Hohenlohe *et al.* 2010) and phylogeography (Emerson *et al.* 2010). More recently, the application of RAD-seq beyond the context of short evolutionary timescale typical of population genomics studies (i.e. phylogenetic studies) has also been investigated (McCormack *et al.* 2012; Rubin *et al.* 2012).

Any molecular approach relying on whole DNA digestion with restriction enzyme(s) to reduce genomic representation might lead to potential biases in the estimation of genetic variability because some copies of DNA fragments might be associated with the absence of a restriction site (i.e. a null allele due to one or several mutations in the recognition sequence of the restriction enzyme). For example, in an RAD-seq analysis based on barcoded individuals (i.e. individual-based approach), any individual heterozygous at a single nucleotide polymorphism (SNP) associated with a null allele will appear as a homozygote for the SNP allele associated with the successfully digested fragment. Similarly, when developing RAD-seq using barcoded pools of individuals (i.e. pool-based approach), the SNP copies associated with the null allele will simply not be sequenced. Thus, both strategies are expected to lead to potential biases in the estimation of allele frequencies at RAD markers affected by null alleles. Following Luca *et al.* (2011), we will hereafter refer to the experimental silencing of marker alleles within unsequenced DNA fragments associated with a null allele at the restriction site as ADO. Luca *et al.* (2011) provided the first empirical evidence of the magnitude and effect of ADO in a human population genomics study based on sequences obtained from RRL. When comparing SNP calling from the resulting raw sequencing data of each individual with the ones obtained for the same individuals using an SNP genotyping chip assay, they observed up to 9.4% of heterozygous miscalled as homozygous. Similarly, nucleotide diversity computed from the raw data was underestimated by about 3%.

Allele dropout at RAD loci is expected to bias the estimation of allele frequencies and hence statistics traditionally used to summarize variation within population, such as expected heterozygosity $H_e$ (Nei 1977), and between populations, such as $F_{ST}$ (Weir & Hill 2002). However, the magnitude of this bias remains unknown. Because NGS approaches relying on the sequencing of pools of individual DNA have recently been proved powerful to survey a large number of markers on a genome-wide scale at a low cost, we focused the present study on ADO at RAD loci on an experimental design based on pools of individuals. We also addressed the more traditional individual-based approach, but to a lesser extent. First, we analytically derived key measures of the effect of ADO on allele frequencies as a function of different parameters within a single population. We then used RAD data sets simulated using a coalescence model to investigate the magnitude of biases induced by ADO on the estimation of expected heterozygosity and $F_{ST}$ under a simple demographic model of divergence between two populations.

## Methods

### Definitions and notations

We considered an RAD locus as generated by the standard protocol of Baird *et al.* (2008): such a RAD locus includes a sequence of $L$ bp long associated with a single restriction site of length $S$ bp. The case of RAD loci generated by the (very) recent protocol of Peterson *et al.* (2012), based on the DNA digestion by two restriction enzymes, was not tackled in the present study.

The restriction site has two allelic states: $R$ for the functional restriction site and $r$ for the mutated version (null allele). Note that, excluding indel mutations in the restriction site, there are $4^S$-1 possible different variants of the null allele. In typical RAD-seq analyses, $S = 6$ (e.g. using the restriction enzyme *Pst*I) or 8 (e.g. using *Sbf*I) and $L$ might vary from 2 × 36 bp (upstream and downstream RAD sequences derived from single-end 36 bp Illumina sequencing) up to around 2 × 500 bp (using 100 bp paired-end reads with a mean paired-end contig length of ~400 bp) (Davey *et al.* 2012). In an individual-based study, four patterns are possible depending on the sequence of the two DNA fragments carried by a diploid individual (Fig. 1B). If the individual is homozygous for the functional restriction site allele (patterns 1 and 2), there is no ADO. If the individual carries the functional and mutated copies of the restriction site (patterns 3 and 4), ADO only arises if it is also heterozygous for the SNP allele (pattern 4). Similarly, in a pool-based experiment (i.e. using a pool of DNA from multiple individuals as experimental unit), ADO arises when the two alleles at the restriction site (functional and mutated) and the two SNP alleles are segregating in the sample (Fig. 1C).

### Probability of segregation patterns

Let us consider a panmictic population at equilibrium with a constant effective population size $N_e$ and an infinite-site neutral mutation model (Kimura & Ohta 1969). We denote the mutation rate per nucleotide scaled by the effective population size, $\theta = 4N_e\mu$ (with $\mu$ the nucleotide mutation rate). For simplicity, we hereafter disregard recombination within the (short) DNA

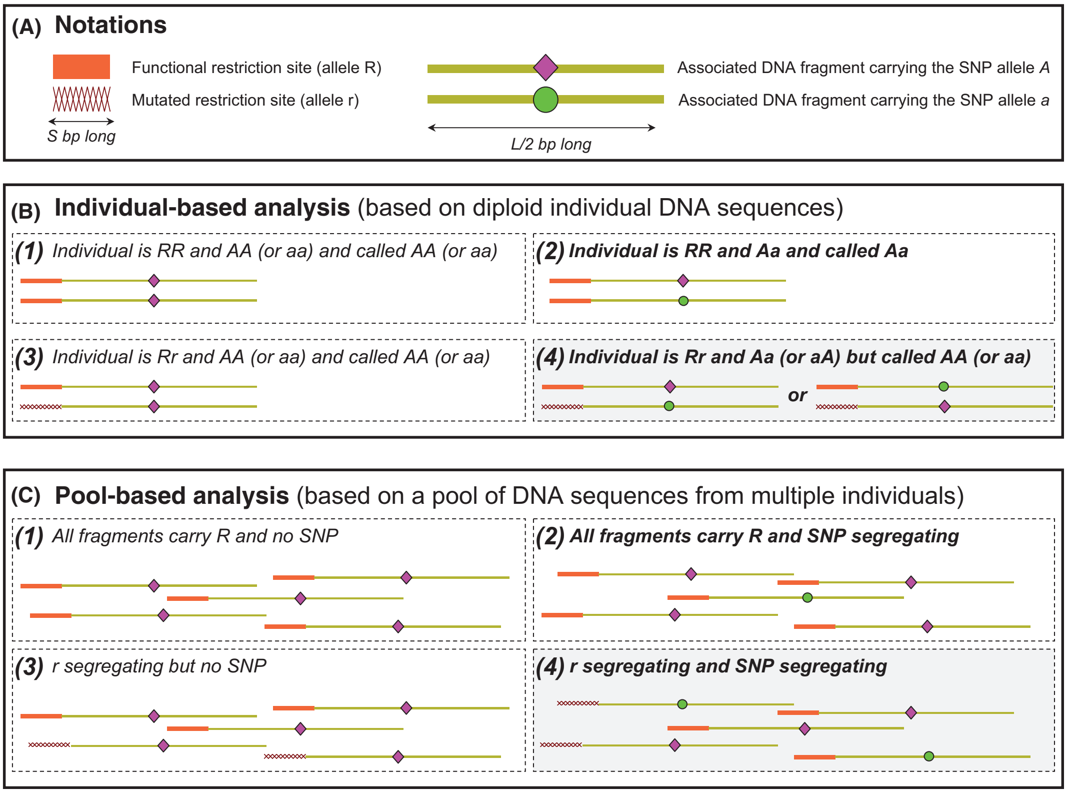
(A) Notations
Functional restriction site (allele R)
Mutated restriction site (allele r)
S bp long
Associated DNA fragment carrying the SNP allele A
Associated DNA fragment carrying the SNP allele a
L/2 bp long
(B) Individual-based analysis (based on diploid individual DNA sequences)
(1) Individual is RR and AA (or aa) and called AA (or aa)
(2) Individual is RR and Aa and called Aa
(3) Individual is Rr and AA (or aa) and called AA (or aa)
(4) Individual is Rr and Aa (or aA) but called AA (or aa)
or
(C) Pool-based analysis (based on a pool of DNA sequences from multiple individuals)
(1) All fragments carry R and no SNP
(2) All fragments carry R and SNP segregating
(3) r segregating but no SNP
(4) r segregating and SNP segregating

**Fig. 1** Allele dropout (ADO) in individual-based and pool-based restriction site associated DNA (RAD) analyses in relation to the segregation patterns between the restriction site and the associated DNA fragment. (A) The two alleles of the *S*-bp-long functional (respectively mutated) restriction site are represented by an orange (respectively dashed) rectangle. Only one of the two *L*/2 bp sequences associated with the restriction site are further represented focusing on the SNP allele (*A* or *a*) they carry. (B) The four possible segregation patterns in an individual-based analysis according to the two DNA copies carried by a diploid individual. ADO arises in the pattern 4 when the individual is heterozygous at both the restriction site (*Rr*) and the SNP (*Aa*). Note that we did not represent the case in which the individual carries two mutated copies of the restriction sites (*rr*; in this case the RAD locus is not observed whatever the SNP genotype). (C) The four possible segregation patterns in a pool-based analysis according to the DNA copies in the sample. ADO arises when the two alleles at the restriction site (functional and mutated), and both SNP alleles are segregating in the sample. In the same spirit than (B), we did not represent the case where all DNA fragments of a pool carried a mutated allelic form *r* of the restriction site (in this case, the RAD locus is not observed in the DNA pool).

sequence of size $S + L$ bp (<1000 bp). In addition, we implicitly assume that no additional restriction site evolves through mutation within the RAD locus. Assuming homogeneity of nucleotide composition across the genome and equal proportion of each nucleotide (50% GC), the distance $d_r$ between two consecutive restriction sites in the genome follows an exponential distribution with parameter $\lambda = \frac{1}{4^S}$ leading to $P(d_r < L) = 1{-}e^{-L\lambda}$. Under the range of parameters investigated here, the probability of having an additional restriction site (discarding the corresponding DNA fragment) within the associated DNA fragment varies from 0.001 ($S = 8$ and $L = 70$) to 0.071 ($S = 6$ and $L = 300$). Such low probability values are unlikely to challenge our conclusions.

Under the above assumptions, we could obtain analytical derivations of the probabilities for the four RAD loci segregation patterns described in Fig. 1B,C. The details of such derivations as well as the final formulae are given in the Appendix S1 (Supporting Information). These four different probabilities were numerically computed using an in-house *R* (R Development Core Team 2011) function (available upon request from MG).

### *Biases due to ADO on the estimation of SNP allele frequency within population*

Following the above notations and focusing on the segregation pattern 4 of Fig. 1, let $f_R > 0$ (respectively $f_r = 1{-}f_R$) represent the population allele frequency of a given functional (respectively null) restriction site. Similarly, let $f_A > 0$ (respectively $f_a = 1{-}f_A$) represent the frequency of allele *A* (respectively *a*) at an SNP in the associated DNA sequence. Four haplotypes denoted

*RA*, *Ra*, *rA* and *ra* are possibly segregating in the population. By definition, only DNA fragments associated with a functional restriction site (*R* allele) are read in RAD-sequencing analysis. Let $\alpha_A$ (respectively $\alpha_a$) represent the relative population frequency of allele *A* associated with an allele *R* background in the whole population:

$$\alpha_A = \frac{f_{RA}}{f_R} = \frac{f_A - f_{rA}}{1 - f_r}$$

(where $f_{RA}$ and $f_{rA}$ are the population frequencies of the *RA* and *rA* haplotypes, respectively). It follows that:

$$\max\left(0; \frac{f_A - f_r}{1 - f_r}\right) \leq \alpha_A \leq \max\left(1; \frac{f_A}{1 - f_r}\right)$$

and the $r^2$ measure (Hill & Robertson 1968) of linkage disequilibrium (LD) between the polymorphic restriction site and the SNP is equal to

$$r^2_{RA} = \frac{(\alpha_A f_R - f_A f_R)^2}{f_A f_R f_a f_r} = f_R \frac{(\alpha_A - f_A)^2}{f_A f_a f_r} .$$

As expected, when $r^2_{RA} = 0$ ($f_{RA} = f_R f_A$) then $\alpha_A = f_A$. When $r^2_{RA} > 0$, $\alpha_A > f_A$ if *A* and *R* are more frequently in coupling phase (i.e. $f_{RA} > f_R f_A$). Conversely, $\alpha_A < f_A$ if *A* and *R* are more frequently in opposite phase (i.e. $f_{RA} < f_R f_A$).

In an individual-based analysis involving the (independent) sequencing of *n* diploid individuals, the observed allele frequency is equal to

$$\widehat{f_A^{(d)}} = \frac{2n_{AA} + n_{Aa}}{2(n_{AA} + n_{Aa} + n_{aa})}$$

where $n_{AA}$, $n_{aa}$, and $n_{Aa}$ represent the observed number of individuals called as homozygous for allele A (i.e. genotypes $[RA/RA]$, $[RA/rA]$ or $[RA/ra]$), for allele a (i.e. genotypes $[Ra/Ra]$, $[Ra/rA]$ or $[Ra/ra]$) and heterozygous (genotype $[RA/Ra]$), respectively. Note that $n_{AA} + n_{Aa} + n_{aa} \leq n$ since individuals carrying two null copies of the restriction site (i.e. with a genotype $[r/r]$) will not be observed. In a pool-based analysis involving a DNA pool sample of haploid size 2*n* (or a pool of *n* diploid individuals), the observed allele frequency is equal to

$$\widehat{f_A^{(p)}} = \frac{n_{RA}}{n_{RA} + n_{Ra}}$$

where $n_{RA}$ and $n_{Ra}$ represent the observed count of *A* and *a* alleles in the pool. As noted previously, $n_{RA} + n_{Ra} \leq n$ due to ADO.

Assuming the studied population is in Hardy–Weinberg equilibrium, it is possible to show (see Appendix S2, Supporting Information) that the expectations and variances of the observed allele frequency estimators defined above are as follows:

$$E\left(\widehat{f_A^{(d)}}\right) = E\left(\widehat{f_A^{(p)}}\right) = \alpha_A;$$

$$V\left(\widehat{f_A^{(d)}}\right) \simeq \frac{1}{2n_d f_R} \alpha_A (1 - \alpha_A) \frac{4 - 3f_R}{(2 - f_R)^2}$$

$$V\left(\widehat{f_A^{(p)}}\right) \simeq \frac{1}{n_c f_R} \alpha_A (1 - \alpha_A)$$

### *ADO in two diverging populations*

We simulated DNA sequence data using the coalescent-based algorithms implemented in routines of the software DIYABC V1 (Cornuet *et al.* 2010). We considered a simple demographic scenario in which two populations diverged $t_S$ generations ago from an ancestral one, with no migration after divergence and all populations having the same effective population size $N_e$ (in number of diploid individuals). Different combinations of $t_s$ and $N_e$ values were considered in order to cover a large range of divergence times and effective population sizes typical of within-species (i.e. population genetics) surveys (Table 1). However, we also ran a few additional simulations with larger $t_s$ values to illustrate the effect of ADO in the context of larger evolutionary times typical of phylogenetic studies. Each population sample of simulated DNA sequence data consisted of 40 copies of a non-recombining sequence fragment of length $S + L$ bp (usually $S = 6$ bp and $L = 300$ bp) combined into 20 diploid individuals per population (assuming Hardy–Weinberg equilibrium). The mutation rate per nucleotide ($\mu$) was sampled from a uniform distribution $[10^{-9}$–$10^{-8}]$, and we assumed a Kimura two-parameters (K2P) model of DNA evolution (Kimura 1980) with a fraction of constant sites (those that do not mutate) fixed to 10% and the shape parameter of the Gamma distribution of mutations among sites equal to 2. Note that this mutation model implicitly assumes equal proportions of bases and homogeneity of nucleotide composition across the genome. We verified our simulation program by estimating from simulated data sets the proportions of the four segregation patterns displayed in Fig. 1 and by comparing these estimations with analytical expectations; a good agreement was found between the two measures (results not shown). The simulation program we used (executable file including a short tutorial text file as well as source files) is available under request from AE.

For each combination of $t_s$ and $N_e$ values, we simulated 10 or 20 millions of (independent) DNA samples

**Table 1** Summary of RAD loci simulations under a model of two diverging populations

| $t_S$ | $N_e$ | $S$ (bp) | (a) | (b) | (c) | (d) |
|---|---|---|---|---|---|---|
| $10^3$ | $10^3$ | 6 | 0.243* | 4.88 (1.03) | 0.00 (0.08) | 0.00 |
|  | $10^4$ | 6 | 0.241 | 29.3 (1.30) | 4.13 (0.41) | 2.00 |
|  | $10^5$ | 6 | 0.263 | 83.4 (4.02) | 15.4 (1.71) | 7.66 |
| $10^4$ | $10^4$ | 6 | 0.249* | 36.3 (1.37) | 2.41 (0.84) | 1.02 |
|  |  | 8 | 0.015 | 36.8 (1.34) | 2.80 (0.72) | 1.48 |
| $10^5$ | $10^4$ | 6 | 0.240 | 51.9 (1.62) | 3.63 (1.46) | 1.57 |
|  | $10^5$ | 6 | 0.272 | 87.8 (5.57) | 22.3 (8.54) | 10.4 |
|  |  | 8 | 0.017 | 84.6 (5.83) | 29.7 (11.4) | 14.4 |
| *$10^6$* | *$10^4$* | *6* | *0.261* | *81.6 (4.71)* | *19.0 (14.6)* | *1.52* |
| *$10^7$* | *$10^4$* | *6* | *0.303* | *68.9 (42.7)* | *48.6 (42.5)* | *0.47* |

We considered a scenario of two diverging populations (20 diploid individuals sampled in each population), assuming various combinations of divergence time ($t_s$) and effective population size ($N_e$) values. Restriction sites of 6- or 8-bp ($S$ (bp)) were also considered. The results corresponding to parameter values used to illustrate the case of large evolutionary times typical of phylogenetic studies are in italics.
(a) ‰ of simulations retained (i.e. with at least one functional restriction site in the population samples) computed over $10^7$ or $2 \times 10^7$ (indicated by*) simulations.
(b) Percentage of RAD loci displaying sequence polymorphism. The average number of SNPs per locus is given in parentheses.
(c) Percentage of SNPs lost due to ADO. The percentage of RAD loci for which a mutated copy of the restriction site is fixed and hence absent in one of the two population samples is indicated in parentheses. In this case, the RAD locus is not observed. Other case of SNP loss is due to the strict association of one SNP variant with a mutated copy of the restriction site.
(d) Percentage of observed (i.e. not lost) SNPs still displaying ADO.
RAD, restriction site associated DNA; ADO, allele dropout.

(200 millions for an 8-bp restriction enzyme such as *Sbf*I). To mimic RAD sequence data, only those population samples in which at least one functional copy of the 6-bp restriction site (e.g. *Pst*I at the first $S$ = 6 bp) was observed in at least one of the two populations were recorded for post-processing treatments.

### *Post-processing treatments of simulated data sets*

For each recorded simulated data set, we used *gawk* scripts and R in-house routines to search for one (or several) SNP in the sequence of length $L$. When an SNP was present, its genetic variation was summarized within population by computing the expected heterozygosity ($H_e$; Nei 1977) for each population and between populations by computing unbiased estimates of $F_{ST}$ as described by Weir & Cockerham (1984) (see also Weir 1996, eqn 5.2).

Because we knew which of the sampled copies of the restriction site were mutated in our recorded population data sets, we could estimate $H_e$ and $F_{ST}$ in two different ways: (i) by keeping the sequence copies associated with a mutated restriction site and thus without taking into account ADO (hereafter named 'true' $H_e$ and $F_{ST}$ values) and (ii) by taking into account ADO (hereafter named 'observed' $H_e$ and $F_{ST}$ values) by computing the statistics after discarding the sequences associated with mutated restriction sites.

To summarize the effect of ADO on the estimation of $F_{ST}$, we introduced a measure of false outlier proportion at a threshold of $x$% ($FOP_{x\%}$). Briefly, let $n_T$ represent the number of SNPs with both a true and an observed $F_{ST}$ above the threshold $t$ defined by the $x$% quantile of the true $F_{ST}$ distribution. Similarly, let $n_F$ represent the number of SNPs with an observed $F_{ST}$ above this same threshold $t$ but a true $F_{ST}$ below $t$. Then we defined the $FOP_{x\%}$ as $FOP_{x\%} = n_F/(n_F + n_T)$ by analogy to the false-positive rate. Note that $FOP_{x\%}$ varies from 0 (if $n_F = 0$) to 1 (if $n_T = 0$). It is worth stressing that $FOP$ was not computed for expected heterozygosity due the discrete nature of this statistics and its U-shape distribution, which make it inappropriate to define a quantile as threshold.

To account for variation in the $FOP_{x\%}$ estimator on $F_{ST}$, we performed each time 50 000 bootstrap samples. For each random sample, $n_{RADloci}$ (corresponding to the observed number of RAD loci) are sampled with replacement, and one SNP per RAD locus is further randomly sampled. In the case of statistics combined across several SNPs (e.g. $F_{ST}$ average over $n_{snp}$ SNPs), the procedure was slightly modified. In the case where RAD loci were independent, each random sample consisted of 1000 $n_{snp}$-uplets SNPs. Each SNP of the $n_{snp}$-uplet was randomly sampled within one RAD locus sampled with replacement among the $n_{RADloci}$ available ones. In the case where RAD loci were simulated as completely linked (see below), a similar procedure was used except that SNPs from the $n_{snp}$-uplet all belong to the same 30 000-bp-long sequence, the 1000 sequences being randomly sampled with replacement among the available ones.

The *gawk* scripts and *R* in-house routines we used for the different post-processing treatments are available under request from MG.

## Results

### *ADO within population and effect on allele frequency estimation*

Figure 1 shows the four possible segregation patterns between an $S$–bp-long restriction site and an $L$-bp-long associated fragment that can be observed in a sample of $n$ DNA copies: (i) no variation at all; (ii) no variation within the restriction site and polymorphism(s) in the DNA fragment; (iii) no variation within the fragment but variation within the restriction site resulting in the segregation of both null and functional copies of the site and (iv) variation in both the restriction site and the associated fragment, which corresponds to the case where ADO is present. Under our assumptions (see Methods), it was possible to derive the proportion of each four segregation patterns within a single population as a function of the size of the restriction site (*S*; in number of bases), the length of the associated DNA sequences (*L*), the (haploid) sample size *n* and the scaled mutation rate $\theta = 4N_e\mu$. A successful RAD experiment generally requires a high proportion of RAD loci displaying pattern 2 (i.e. informative RAD markers), with a low proportion of RAD loci showing pattern 4 (i.e. null alleles).

Expected proportions of RAD loci showing patterns 2 ($P_2$) and 4 ($P_4$) are plotted in Fig. 2 as a function of $\theta$ for different values of *S* ($S = 6$ or $S = 8$ representing a 6- or an 8-bp cutter, respectively) and *L* ($L = 70$ or $L = 300$) assuming $n = 40$ (Fig. 2A) and for different values of *n* ($n = 2$, $n = 10$, $n = 40$, $n = 100$ or $n = 200$) assuming $S = 6$ and $L = 300$ (Fig. 2B). The values considered for $\theta$ ranged from $10^{-5}$ to $10^{-1}$, which includes most realistic mutation–drift situations (e.g. assuming $\mu = 10^{-8}$ such $\theta$ values correspond to $N_e$ ranging from 250 to 2 500 000). As shown in Fig. 2A, for low values

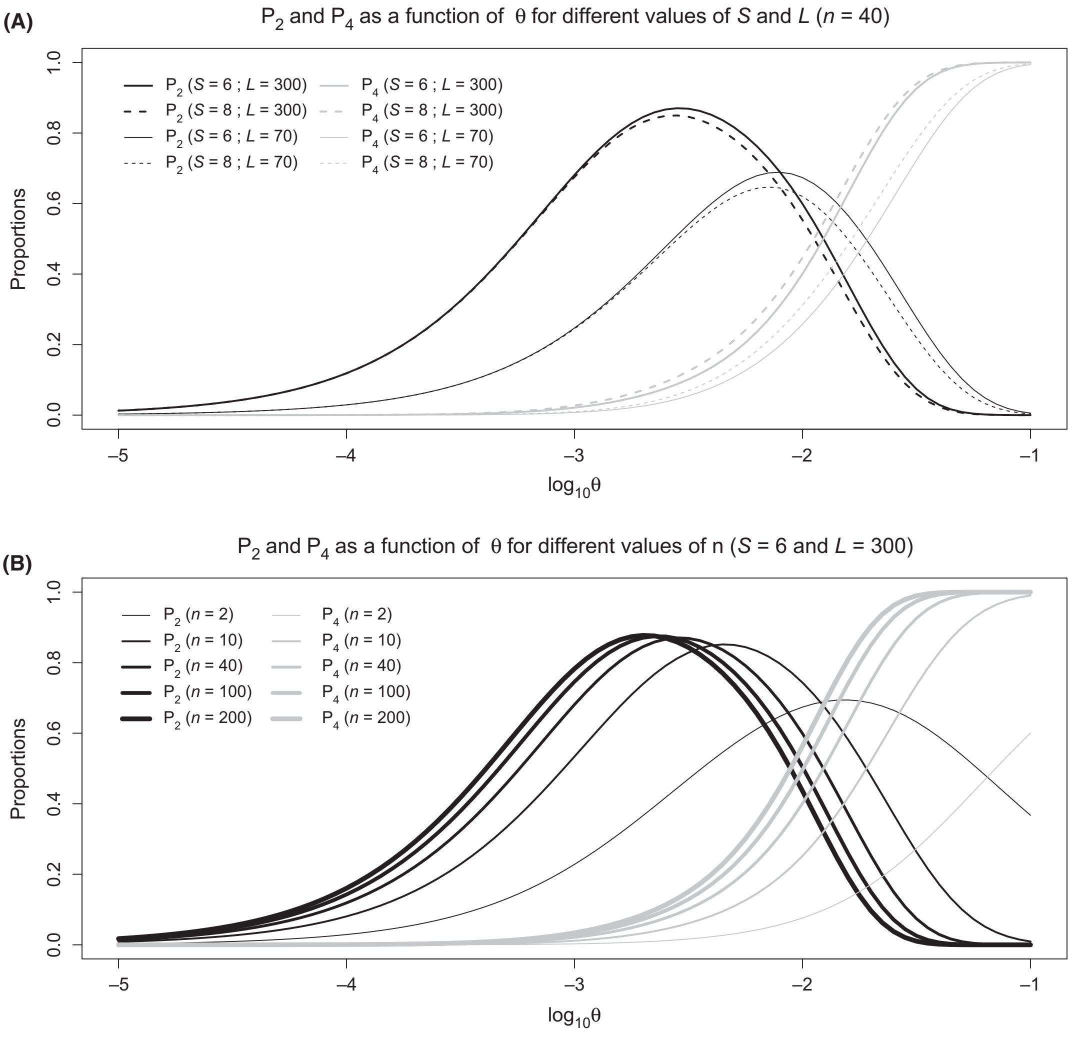

**Fig. 2** Expected proportions within a single close population of RAD loci displaying the segregation patterns 2 and 4 described in Fig. 1. Pattern 2: absence of variation within the restriction site and polymorphism(s) in the associated DNA sequence. Pattern 4: presence of variation within both the restriction site and the associated sequence, leading to ADO. The proportions of patterns 2 and 4 are given under simple demographic assumptions (see the main text) as a function of the scale mutation restriction $\theta = 4N_e\mu$ for: (A) different values of *S* (length of the restriction site in number of bp) and *L* (length of the associated DNA sequence in bp), assuming $n = 40$ (haploid sample size), and for (B) different values of *n* assuming $S = 6$ and $L = 300$. ADO, allele dropout; RAD, restriction site associated DNA.

of $\theta$ ($\theta < 10^{-4}$ i.e. $N_e < 2500$ if $\mu = 10^{-8}$), $P_4$ is almost null, and hence, no ADO is observed. However, only a small fraction of RAD loci are useful ($P_2 < 0.03$ with $L = 70$ and $P_2 < 0.12$ with $L = 300$), suggesting a reduced efficiency of RAD sequencing for populations of small effective size (at least in terms of marker productions). Conversely, for $\theta > 10^{-3}$ (i.e. $N_e > 25\,000$ if $\mu = 10^{-8}$), the proportions of RAD loci showing ADO start to sharply increase towards high values ($P_4$ varied from 0.26 to 0.45 when $\theta = 10^{-2}$ and $P_4 > 0.99$ for $\theta = 10^{-1}$). In other words, $\theta$ must be large enough to generate variability (typically $\theta > 10^{-4}$) but not too large (typically $\theta < 10^{-2}$) so that variability mostly (if not only) occurs within the DNA fragment associated with the restriction site. Increasing $L$ relatively to $S$ appears to be useful for most $\theta$ values, for instance, when $\theta = 10^{-3}$ and $S = 6$, $P_2 = 0.25$ (respectively $P_2 = 0.68$) and $P_4 = 0.01$ (respectively $P_4 = 0.02$) for $L = 70$ (respectively $L = 300$). We found that, for a given $L$, decreasing the size of the restriction site $S$ (which results in practice in an exponential increase in the number of RAD loci in the data set) had only marginal effects on the proportions of $P_2$ and $P_4$ (Fig. 2A). Moreover, the magnitude of this effect decreased with $L$. Finally, we found that, above a haploid sample size of reasonable size ($n > 10$), the sample size has marginal effect on the proportions of $P_2$ and $P_4$ (Fig. 2B).

For allele frequency estimation, it is worth stressing that RAD loci matching pattern 4 (i.e. with ADO) show heterogeneous frequency of the functional copy of the restriction site (note that a minimum of one copy of the mutated version of the restriction site associated with a variable DNA sequence is sufficient to match pattern 4). We might expect that if the mutated restriction site is at a low frequency, the effect on allele frequency estimates would be negligible. As shown previously (see Methods), the expectations of the observed allele frequency at the SNP is equal to the relative frequency of the allele in the $R$ background (where $R$ denotes the functional copy of the restriction site), and these expectations are the same for the individual and pool-based approaches. Figure 3 illustrates how the resulting expected bias in the estimation of allele frequency due to ADO increases sharply with both the frequency of the mutated copies of the restriction site ($f_r = 1 - f_R$) and the magnitude of LD (measured here with $r^2$)

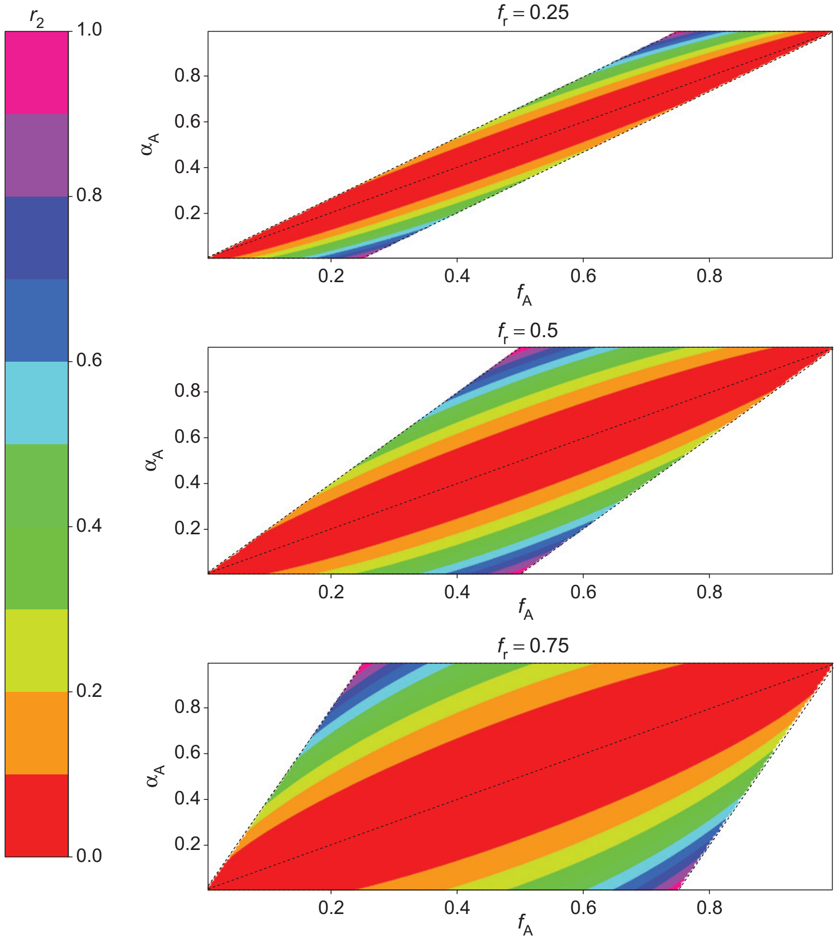

**Fig. 3** Relationship at a RAD locus between the expected observed allele frequency of one allele (termed $A$) at a SNP in the presence of allele dropout (ADO) and its true frequency ($f_A$). ADO occurs when mutated copies $r$ of the restriction site are segregating in the population at frequency $f_r > 0$. Envelopes of possible values of the expected observed allele frequency

$$E\left(\widehat{f_A^{(d)}}\right) = E\left(\widehat{f_A^{(p)}}\right) = \alpha_A$$

are displayed for $f_r = 0.25$; $f_r = 0.5$ and $f_r = 0.75$. Inside the envelope, regions are coloured according to the level of linkage disequilibrium (as measured by $r^2$) between the restriction and SNP sites. RAD, restriction site associated DNA.

between the restriction site polymorphism and the SNP within the associated DNA fragment. It is worth noting that relatively high values of $r^2$ are expected between the restriction and SNP sites since they are physically closely located. Indeed, we note that

$$E(r^2) \simeq \frac{10+\rho}{22+13\rho+\rho^2}$$

(Ohta & Kimura 1971; McVean 2002; see also Weir & Hill 1986, who included an additional sample size term in the equation), where $\rho = 4N_e c$ is the scaled recombination rate leading to $E(r^2) \simeq 0.45$ when $\rho \ll 1$ (assuming 1 cM is equivalent to 1 Mb, $\rho < N_e \times 10^{-5}$ if the SNP is <250 bp from the restriction site).

Although individual and pool-based analyses lead to the same expected bias in the estimation of allele frequency, the sampling variance of the estimator is increased in individual-based experiments by a factor

$$\lambda = \frac{n_h}{2n_d}\frac{3f_r+1}{(f_r+1)^2}$$

where $n_h$ is the haploid sample size in the pool-based analysis, and $n_d$ is the number of diploid individuals considered in the individual-based analysis (see Appendix S2, Supporting Information). Interestingly, given $n_h$ and $n_d$, λ only depends on the frequency of the null copies of the restriction site ($f_r$) and varies from $\lambda_{\min} = \frac{n_h}{2n_d}$ (when $f_r$ tends towards 0 or 1) to $\lambda_{\max} = \frac{9n_h}{16n_d}$ (when $f_r = 1/3$). The extra sampling variance in an individual-based analysis hence remains low (<12.5%) for equivalent sample size (i.e. if $n_d = 2n_h$).

### *Effect of ADO on the estimation of genetic variation within and between populations*

We used a coalescence model to simulate RAD sequences assuming a simple demographic scenario in which two populations diverged $t_S$ generations ago from an ancestral one, with no migration after divergence and all populations having the same diploid effective population size $N_e$. Different combinations of $t_s$ and $N_e$ values were considered in order to cover a large range of divergence times and effective population sizes typical of within-species surveys (i.e. population genetics). However, we also explored a few additional parameter values to illustrate the case of longer evolutionary times typical of phylogenetic studies. Table 1 provides the parameter values considered together with several descriptive statistics. Assuming equal proportion and distribution of bases across the genome, the probability that a single sequence in a population starts with a given 6- or 8-bp cutter recognition site is $2.44 \times 10^{-4}$ or $1.5 \times 10^{-5}$, respectively. Increasing the number of sampled copies of a given DNA sequence (here 40 copies in each population sample) only marginally affects this probability. The proportion of simulations retained (i.e. displaying at least one functional segregating site among the sampled sequence copies) was therefore close to the expected values (column (a) of Table 1) but tended to increase with long divergence times between populations (e.g. 0.303% of retained simulations for a 6-bp cutter for $t_S = 10^7$ and $N_e = 10^4$).

Table 1 shows that the effective population size had a stronger effect than the divergence time on most computed descriptive statistics. As expected, the percentage of RAD sequences showing polymorphism and the number of SNP per sequence strongly increased with $N_e$. The proportion of SNPs with more than two alleles remained negligible (at most 0.58% for $N_e = t_S = 10^5$). The percentage of observed SNP affected by ADO was low for most $N_e$ values but reached non-negligible values (around 10%) for $N_e = 10^5$ (column (d) of Table 1). The divergence time mostly impacted the proportion of retained sequence simulations for which the restriction site was absent from one of the two population samples due to fixation of the null variant of the restriction site in one of the two populations (in this case, the RAD locus cannot be observed). These proportions were found to be low to moderate, except for very long divergence times (e.g. 42.5% for $t_S = 10^7$; column (a) of Table 1). It is worth noting that for long divergence times, the percentage of the observed (i.e. not associated with a null allele) SNPs showing ADO was low (e.g. 0.47% and 1.52% for $t_S = 10^7$ and $10^6$, respectively; column (d) of Table 1). The size of the restriction site (6 or 8 bp) had only minor effect on the computed descriptive statistics (excluding the proportions of retained simulations).

For each retained simulated data set, we further estimated SNP-expected heterozygosities in one of the two populations ($H_e$) and $F_{\mathrm{ST}}$ between the two populations. Two estimations were derived for each data set: (i) by keeping the sequence copies associated with a null restriction site and thus without taking into account ADO (hereafter named 'true' $H_e$ and $F_{\mathrm{ST}}$ values) and (ii) by taking into account ADO and discarding the sequence copies associated with a null allele (hereafter named 'observed' $H_e$ and $F_{\mathrm{ST}}$ values). To evaluate to which extent the effect of ADO on these statistics depends on the frequency of mutated copies at the restriction sites ($f_r$), we considered different threshold on $f_r$ to remove RAD markers. More specifically, we considered $f_r^{\max}$ values as threshold, which corresponds to the maximum value of $f_r$ in the two populations ($f_r = f_r^{\max}$ values when considering SNP heterozygosities in one of the two populations). Figure 4 illustrates the relationship between true and observed values of $H_{\mathrm{e}}$ and $F_{\mathrm{ST}}$ in the case where the differences between the two measures were the strongest among the studied combinations of $t_s$

and $N_e$ values (i.e. for $N_e = t_S = 10^5$). The smaller differences observed for other combinations of $t_s$ and $N_e$ values are not shown in the figure, but quantitative estimations of such differences were summarized by computing the false outlier proportions ($FOP_{95\%}$) of $F_{\mathrm{ST}}$ due to ADO (Table 2). A first interesting result that can be visualized from Fig. 4A is that the observed heterozygosity tends to be overestimated relative to the true values. This result, which might seem surprising at first, can be explained by the fact that mutations altering the restriction site (and thus leading to ADO) tend to affect DNA sequences bearing ancestral SNP allelic states that themselves tend to be at high frequency within population. Therefore, ADO tends to inflate the value of the minimum allele frequency (MAF) at the SNP locus and thus $H_{\mathrm{e}}$ (Fig. S1, Supporting Information).

Figure 4B shows that observed $F_{\mathrm{ST}}$ values tend to be overestimated relative to true values (i.e. ADO tends to increase genetic differentiation between populations). A close examination of colour codes in Fig. 4B shows that strong $F_{\mathrm{ST}}$ biases preferentially affect SNPs with high frequency of mutated copies of the restriction site (i.e. those with $f_r^{\max} > 0.5$). However, Table 2 shows that $FOP_{95\%}$ remains low for most studied combinations of $t_s$ and $N_e$ values, except for large effective population sizes (i.e. $FOP_{95\%} > 10\%$ for $N_e = 10^5$ whatever the value of $t_s$). Interestingly, even for large $N_e$ values, $FOP_{95\%}$ substantially decreased when removing RAD loci with a high frequency of the null restriction site ($FOP_{95\%} < 5\%$ and 3% for $f_r^{\max}$ values $\leq 0.75$ and $\leq 0.5$, respectively). We found that *FOP* values increased when the $F_{\mathrm{ST}}$ threshold decreased (Table S1, Supporting Information). This reflects the fact that a majority of outlier SNPs are characterized by a strong $F_{\mathrm{ST}}$ bias due to a high frequency of the null restriction site (i.e. located in the upper part of the $F_{\mathrm{ST}}$ distribution tail, see Fig. 4B for an illustration). Finally, in agreement with analytical results on the possible segregation patterns within populations (detailed in Fig. 1), we found that *FOPs* were similar for all combinations of $t_s$ and $N_e$ values when considering diploid individual DNA sequences (i.e. individual-based analysis) instead of pools of sequences from multiple individuals (pool-based analysis; see Table S2, Supporting Information to be compared with Table 2).

We found that averaging $F_{\mathrm{ST}}$ values over multiple SNPs (e.g. sliding window analysis) did not reduce *FOP* values (Table 3). In fact, *FOP* values increased with the number of SNPs used to compute $F_{\mathrm{ST}}$. This result is true irrespectively of the level of genetic linkage among SNPs, although the *FOP* increase we observed was lower for completely linked than for independent SNPs. It is worth noting that, in agreement with previous analytical results, the distributions of $F_{\mathrm{ST}}$ averaged over multiple SNPs are characterized by a variance which decreases with the number of SNPs considered (Table S3, Supporting Information). Indeed, for $n_{snp}$ SNPs, the two-population $F_{\mathrm{ST}}$ estimates have a distribution close to a (scaled) chi-square distribution with $n_{snp}$ degree of freedom and so have a variance proportional to $1/n_{snp}$ (Weir & Hill 2002). Moreover, because a majority of outlier SNPs is characterized by a strong $F_{\mathrm{ST}}$ bias, these outliers have a strong effect on the observed average $F_{\mathrm{ST}}$ values. Altogether, these effects lead to an increase of *FOP* values with the number of SNPs used to compute $F_{\mathrm{ST}}$ values. The variance on average $F_{\mathrm{ST}}$ was found to decrease more

**Table 2** $F_{\mathrm{ST}}$ false outlier proportions ($FOP_{95\%}$ in%) in pool-based RAD analyses

| $t_S$ | $N_e$ | $S$ (bp) | $f_r^{\max} \leq 0.25$ | $f_r^{\max} \leq 0.5$ | $f_r^{\max} \leq 0.75$ | $f_r^{\max} \leq 1.00$ |
|---|---|---|---|---|---|---|
| $10^3$ | $10^3$ | 6 | 0.0 [0.0–0.0] (237) | 0.0 [0.0–0.0] (237) | 0.0 [0.0–0.0] (237) | 0.0 [0.0–0.0] (237) |
| | $10^4$ | 6 | 0.0 [0.0–0.0] (706) | 0.0 [0.0–7.7] (708) | 0.0 [0.0–7.7] (708) | 0.0 [0.0–7.70] (708) |
| | $10^5$ | 6 | 1.8 [0.0–4.5] (2119) | 2.7 [0.0–6.4] (2147) | 4.6 [0.9–9.4] (2174) | 10.7 [5.26–16.4] (2193) |
| $10^4$ | $10^4$ | 6 | 0.0 [0.0–1.0] (1797) | 0.0 [0.0–1.0] (1799) | 0 [0.0–2.3] (1804) | 1.10 [0.0–4.30] (1806) |
| | | 8 | 0.0 [0.0–0.0] (1109) | 0.0 [0.0–0.0] (1116) | 0.0 [0.0–0.0] (1116) | 0.0 [0.0–0.0] (1116) |
| $10^5$ | $10^4$ | 6 | 0.0 [0.0–0.0] (1231) | 0.0 [0.0–0.0] (1234) | 0.0 [0.0–0.0] (1236) | 0.0 [0.0–0.0] (1236) |
| | $10^5$ | 6 | 0.0 [0.0–1.7] (2247) | 0.9 [0.0–3.3] (2299) | 4.5 [1.5–8.5] (2324) | 10.1 [5.5–15.1] (2346) |
| | | 8 | 0.7 [0.0–2.4] (2738) | 1.4 [0.0–4.11] (2790) | 3.3 [0.7–6.4] (2832) | 13.0 [8.3–18.0] (2878) |
| $10^6$ | $10^4$ | 6 | 0.0 [0.0–0.0] (2024) | 0.0 [0.0–0.0] (2028) | 0.0 [0.0–0.0] (2032) | 0.0 [0.0–0.0] (2036) |
| $10^7$ | $10^4$ | 6 | 0.0 [0.0–0.0] (1737) | 0.0 [0.0–0.0] (1738) | 0.0 [0.0–0.0] (1739) | 0.0 [0.0–0.0] (1741) |

We considered a scenario of two diverging populations (20 diploid individuals sampled in each population), assuming various combinations of divergence time ($t_s$) and effective population size ($N_e$) values. Restriction sites of 6- or 8-bp ($S$ (bp)) were also considered. The table reports median [$q_{2.5\%}$–$q_{97.5\%}$] of the $FOP_{95\%}$ distribution obtained across 50 000 random samples for different cut-off values on $f_r^{\max}$ (maximal frequency of the mutated restriction site in both populations). For each random sample, $n_{RADloci}$ RAD loci (corresponding to the number given in parentheses) are sampled with replacement (bootstrap sample), and one SNP per locus is further randomly sampled.

RAD, restriction site associated DNA.

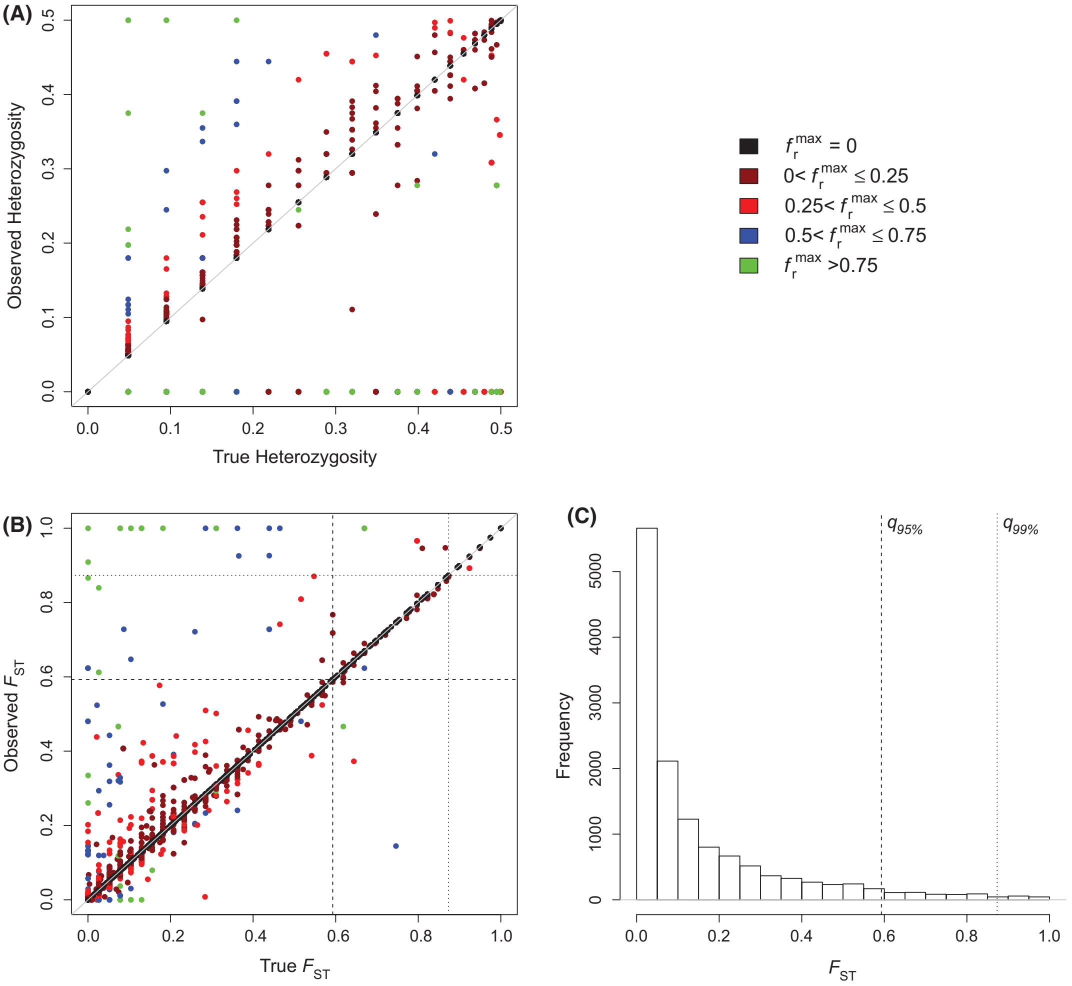

**Fig. 4** Relationships between observed (taking ADO into account) and true estimations of heterozygosity (panel A) and $F_{ST}$ (panel B) under a scenario of two diverging populations assuming $N_e = 10^5$ and $t_S = 10^5$. The panel C represents the distribution of true $F_{ST}$ with 95% and 99% quantiles (also drawn in panel B). The different colours of points correspond to different maximal frequencies of the mutated restriction site in both populations ($f_r^{max}$ values, see colour code legend). Because heterozygosity ($H_e$) was estimated within a single population, $f_r^{max} = f_r$ in panel A. $H_e$ and $F_{ST}$ were computed following Nei (1977) and Weir & Hill (2002), respectively. ADO, allele dropout.

quickly with the number of SNPs when the level of linkage decreases, hence the higher *FOP* increase observed for independent than completely linked SNPs (Table 3 and Table S3, Supporting Information).

## Discussion

### *Effect of ADO on genetic variation estimation*

Analytical derivations confirmed intuitive expectations that the frequency of ADO within population depends on the mutation rate per nucleotide scaled by the effective population size (i.e. $\theta = 4N_e\mu$). The parameter $\theta$ must be large enough so that there is polymorphism in the population sample but not too large, to prevent the occurrence of mutations within the restriction site. The optimal design depends on the ratio between the length of the restriction site and the length of DNA sequence associated with the latter. For RAD loci characterized by relatively long DNA sequences (e.g. 300 bp as in our simulation study), this ratio is very similar for 6- and 8-bp cutters, which explains that we obtained relatively similar results for both classes of restriction enzymes.

We found that ADO tends to overestimate genetic variation both within and between populations, as measured by expected heterozygosity and $F_{ST}$, respectively. The overestimation of expected heterozygosity relatively to the true values is due to the fact that we only consider (as in practice) observed SNPs (i.e. the sets of RAD loci and SNPs that were not lost due to ADO). When all the SNPs were considered (which would be impossible to achieve in practice but can be done from

**Table 3** False outlier proportions ($FOP_{95\%}$ in%) for $F_{\mathrm{ST}}$ averaged over $n_{snp}$ SNPs under a scenario of two diverging populations assuming $N_e = 10^5$ and $t_S = 10^5$

| $n_{snp}$ | $c$ | $f_r^{\max} \leq 0.25$ | $f_r^{\max} \leq 0.5$ | $f_r^{\max} \leq 0.75$ | $f_r^{\max} \leq 1.00$ |
|---|---|---|---|---|---|
| 1 | 0.5 | 0.0 [0.0–2.0] (2247) | 0.0 [0.0–3.9] (2299) | 3.8 [0.0–10.7] (2324) | 10.2 [3.6–18.0] (2346) |
| | 0.0 | 0.0 [0.0–3.8] (1626) | 2.0 [0.0–7.4] (1669) | 3.8 [0.0–10.7] (1686) | 8.8 [1.9–16.4] (1711) |
| 3 | 0.5 | 0.0 [0.0–5.7] (2247) | 2.1 [0.0–9.1] (2299) | 9.1 [2.0–16.7] (2324) | 19.4 [9.3–27.9] (2346) |
| | 0.0 | 0.0 [0.0–3.9] (1626) | 2.0 [0–7.5] (1669) | 5.7 [0.0–12.5] (1686) | 11.3 [3.8–20.6] (1711) |
| 6 | 0.5 | 2.0 [0.0–5.8] (2247) | 3.9 [0.0–10.7] (2299) | 10.7 [3.8–18.6] (2324) | 21.9 [12.3–31.0] (2346) |
| | 0.0 | 0.0 [0.0–5.7] (1626) | 2.0 [0.0–7.5] (1669) | 5.7 [0.0–12.5] (1686) | 12.5 [3.9–21.0] (1711) |

We considered a pool-based RAD analysis with pools of DNA sequences from 20 diploid individuals per population as experimental unit. Different cut-off values on $f_r^{\max}$ (maximal frequency of the mutated restriction site in both populations) are considered. The table reports median [$q_{2.5\%}$–$q_{97.5\%}$] of the $FOP_{95\%}$ distribution obtained across 50 000 random samples. In the independent cases (recombination rate, $c$ = 0.5), each random sample consists of 1000 $n_{snp}$-uplets SNPs. Each SNP of the $n_{snp}$-uplet was randomly sampled within one RAD locus sampled with replacement among the available ones (number given in parentheses). In the completely linked case ($c$ = 0), a similar procedure was considered except that SNPs from the $n_{snp}$-uplet all belong to the same 30 000-bp-long sequence (see the main text), the 1000 sequences being randomly sampled with replacement among the available ones (number given in parentheses).

RAD, restriction site associated DNA.

our computed simulated data), ADO tended to underestimate expected heterozygosity (results not shown), in agreement with the aforementioned results by Luca *et al.* (2011) who reported a decrease in observed nucleotide diversity due to ADO.

Assuming a mutation rate per nucleotide between $10^{-9}$ and $10^{-8}$ and a simple population model of two populations of effective size $N_e$ diverging $t_s$ generations ago, we found that the biases due to ADO remained low for most studied combinations of $t_s$ and $N_e$ values, except for large effective population sizes (i.e. $F_{\mathrm{ST}}$ *FOP* > 10% for $N_e = 10^5$, whatever the value of $t_s$). It is worth noting here the antagonistic effects of effective population size on ADO. On the one hand, large $N_e$ values favour the occurrence of mutations within the restriction site and therefore increase the frequency of RAD loci with ADO. On the other hand, large $N_e$ values increase the recombination rate within any DNA fragment of a given size so that the SNPs bearing an allele strongly associated with the mutated restriction site copies tend to be only those physically very close to the restriction site (McVean 2002; Fig. 3). Interestingly, we also found that, for all studied combinations of $t_s$ and $N_e$ values, $F_{\mathrm{ST}}$ *FOP* were similar when considering diploid individual DNA sequences as experimental unit (i.e. individual-based analysis) instead of pools of sequences from multiple individuals (i.e. pool-based analysis).

### *Mitigating the effect of ADO*

Our results indicate that for species characterized by large effective population sizes, it might be necessary to find means to reduce the effect of ADO at RAD loci. Increasing sample size is unlikely to be a useful approach as both analytical derivations and additional simulations based on a larger sample size (100 instead of 20 diploid individuals per population) indicated that the number of sampled individuals only marginally affected ADO occurrence and effect on genetic variation (results not shown). Using a lower threshold for detecting SNPs with outlier $F_{\mathrm{ST}}$ values does not represent an operational solution either since we found that *FOP* values increase when the detection threshold decreases. Finally, we found that averaging $F_{\mathrm{ST}}$ values over multiple SNPs, as is generally implemented in sliding window analysis, did not reduce *FOP* and in fact leads to an increase in *FOP*. Table 2 however shows that, even for large $N_e$ values, *FOP* values substantially decreased when removing RAD loci with high frequency of mutated restriction site ($FOP_{95\%}$ < 5% and 3% for $f_r^{\max}$ $\leq$0.75 and $\leq$0.5, respectively). This result strongly suggests that a practical solution might consist in detecting and removing the RAD loci characterized by high ADO frequency (say with $f_r \geq 0.5$).

Restriction site associated DNA markers affected by ADO are expected to show reduced read coverage due to missing RAD fragments associated with the null allele at the restriction site. Thus, the analysis of read coverage data might provide a practical solution for detecting and ruling out problematic cases of ADO (i.e. those with $f_r \geq 0.5$). Estimation of ADO allele frequencies within population in an individual-based analysis could be based on the Dempster *et al.*'s (1977) algorithm which has been widely used for null allele frequency estimation at molecular markers such as microsatellites or allozymes (e.g. Chapuis & Estoup 2007), provided that one can add a statistical layer including features typical of NGS markers (e.g. reduced coverage). RAD experimental designs involving high-throughput sequencing of pools of individuals might be more difficult to deal with. However, in this case too, the detection of problematic cases of ADO analysis using read coverage data might provide a solution. For example, the reduction of read coverage for RAD loci showing ADO might be assumed simply related to the underlying frequency of the mutated restriction site. This was empirically confirmed by an observed reduction in coverage around 50% (from 38% to 73%) in sequences associated with heterozygous restriction sites in the study by Luca *et al.* (2011). A mixture of overdispersed Poisson distributions seems a promising approach to model observed read count and to ultimately detect RAD loci displaying unexpected low coverage.

### *Limits of our approach*

A first limitation of our approach is that we did not take into account any potential error typical of NGS data production. Different sources of errors in NGS data have been previously identified and commented on (e.g. limited coverage, PCR bias, sequencing error, misalignment of reads; for a review, see Rokas & Abbot 2009; Pool *et al.* 2010). A full understanding of the errors structure has not been reached yet, and many of these errors are a function of the specific study design, including sequencing depth, platform and chemistry, library preparation and post-sequence processing. Error and bias profiles specific to the RAD technique are also emerging (Davey *et al.*, this issue). It was therefore beyond the scope of this study to include these errors in our mathematical and simulation settings through additional statistical layers. However, we believe that these errors are unlikely to change the general conclusions of our study, providing that read coverage is large enough so that raw sequencing data allow to give a comprehensive picture of the underlying RAD fragments.

A second limitation of our study is that only simple population models have been considered (i.e. a single close population of stable effective population size or a model of two populations of effective size $N_e$ diverging $t_s$ generations ago). However, these basic population models represent a first key step towards an understanding of the effect of ADO on genetic variation within and between populations at RAD loci. One could imagine more complex population demographic scenarios under which ADO is likely to be stronger than what we have observed in the present study. In particular, scenarios involving recent genetic admixtures between two source populations that diverged some substantial time ago are expected to favour the presence of RAD loci with high frequency of ADO in the admixed populations. To confirm this expectation, we ran additional simulations in which RAD loci were sampled from two populations originating from two recent (i.e. 50 generations ago) independent admixture events involving two parental populations characterized by large populations sizes ($N_e = 10^5$) and divergence time ($t_s = 10^6$). We found that in this case, the $F_{\mathrm{ST}}$ $FOP_{95\%}$ statistics computed between the two admixed populations at single SNP reached values as high as 27.8% (results not shown). A (more) stringent filtering of RAD loci with ADO was necessary in this case to reduce $FOP_{95\%}$ values below 5% (i.e. removal of RAD loci with ADO frequency $f_r \geq 0.3$). Additional studies are needed to further investigate the occurrence and effect of ADO at RAD loci under complex and potentially problematic population demographic scenarios.

Another limitation of our simulation-based approach is that we did not consider the effect of recombination within the RAD locus. As shown by analytical results, LD between SNPs and restriction site polymorphisms might be viewed as a key parameter to understand the effect of ADO on the SNP variability. Indeed, when LD tends towards 0, the allele frequencies in the functional restriction site background tend to be similar to the actual ones. Therefore, as the extent of LD is an inverse function of the genetic distance and the effective population size, for very large population sizes (where ADO occurrence is particularly problematic), our conclusions might be somewhat conservative in the sense that we might have overestimated ADO effect for large effective population sizes.

Finally, only a few parameter values were explored to tackle the case of large evolutionary times typical of phylogenetic studies. However, this limitation is justified by the fact that we deliberately focused our study on RAD loci data sets produced in the context of population genetics surveys and thus mostly considered combinations of divergence times and effective population sizes typical of such surveys. The few simulations for which large divergence times were considered ($t_s = 10^6$ and $10^7$) indicated a substantial loss of RAD markers due to the fixation of a mutated variant of the restriction site in one of the two populations (e.g. for 42.5% of the RAD loci when $t_S = 10^7$). Although genetic differentiation at RAD loci should be summarized by other statistics than $F_{\mathrm{ST}}$ in the context of phylogeographical or phylogenetic study (e.g. Emerson *et al.* 2010), our simulation studies indicated that most RAD loci for which a functional copy of the restriction site remained present in the two taxonomical units under study were characterized by limited ADO effect, at least as measured by *FOP* estimation. However, we found that the indirect selection of RAD markers conserved in two diverging taxonomical units tends to induce a bias towards DNA sequences characterized by low mutation rates, a feature that might lead to underestimating genetic divergence between species (results not shown).

Additional studies here are also needed to better understand the occurrence and effect of ADO at RAD loci in the context of phylogenetic studies.

## Acknowledgements

This research was financially supported by the Institut National de la Recherche Agronomique (grant dedicated to the Jeune équipe IGGIPOP) to M.G., J.-M.C. and A.E. and by the French 'Agence Nationale de la Recherche' through the ANR project GenoPheno (2010-JCJC-1705-01). The GenePool is funded in part by grants from the UK Natural Environment Research Council (R8/H10/56) and Medical Research Council (G0900740).

## References

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS ONE*, **6**, e19315.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621–631.

Cornuet JM, Ravigné V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics*, **11**, 401.

Davey JW, Blaxter ML (2011) RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, **9**, 416–423.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Ctachen JM, Blaxter ML, (2012) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

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

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

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

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.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–771.

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and applications to human evolution. *Genome Research*, **21**, 1087–1098.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2012) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, in press. doi:10.1016/j.ympev.2011.12.007.

McVean GA (2002) A genealogical interpretation of linkage disequilibrium. *Genetics*, **162**, 987–991.

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics*, **41**, 225–233.

Ohta T, Kimura M (1971) Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, **68**, 571–580.

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 non-model species. *PLoS ONE*, **7**, e37135.

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

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available from http://www.R-project.org/.

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

Rubin BE, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

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

Weir BS (1996) Genetic Data Analysis II. Sinauer Associates, Sunderland, Massachusetts. pp. 173.

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

Weir BS, Hill WG (1986) Letters to the Editor: nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, **38**, 776–778.

Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721–750.

Conceived and designed the study: M.G., K.G., A.E. Performed analytical derivations: M.G., P.P. Developed the simulation software: J.M.C., A.E. Performed simulation studies: M.G., A.E. Wrote the paper and approved the final version: M.G., K.G., T.C., J.F., C.K., P.P., J.M.C., A.E.

## Supporting information

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

**Table S1** $F_{ST}$ false outlier proportions (*FOP*) at different threshold under a scenario of two diverging populations assuming $N_e = 105$ and $t_S = 105$.

**Table S2** $F_{\mathrm{ST}}$ false outlier proportions ($FOP_{95\%}$ in %) in individual-based RAD analyses.

**Table S3** Standard deviation (SD) of the $F_{\mathrm{ST}}$–based scores (average over *nsnp* SNPs) under a scenario of two diverging populations assuming $N_e$ = 105 and $t_S$ = 105.

**Fig. S1** Relationship between observed and true minor allele frequencies (*MAF*) under a scenario of two diverging populations assuming $N_e = 10^5$ and $t_S = 10^5$.

**Appendix S1** Probability of the four RAD loci segregation patterns described in Fig. 1 of the main text.

**Appendix S2** Biases in allele frequency estimation within population due to allele drop-out (ADO).