Estimating and accounting for genotyping errors in RAD-seq experiments

Luisa Bresadola, Vivian Link, C. Alex Buerkle, Christian Lexer, Daniel Wegmann

Abstract

In non-model organisms, evolutionary questions are frequently addressed using reduced representation sequencing techniques due to their relatively low cost, ease of use, and because they do not require genomic resources such as a reference genome. However, evidence is accumulating that many such techniques may be affected by specific biases, questioning the accuracy of obtained genotypes, and as a consequence, their usefulness in evolutionary studies. Here we introduce three strategies to assess genotyping error rates in such data: through the comparison with high quality genotypes obtained with a different technique, from independent replicates of some samples, or from a population sample when assuming Hardy-Weinberg equilibrium. Applying these strategies to data obtained with Restriction site Associated DNA sequencing (RAD-seq), arguably the most popular reduced representation sequencing technique, revealed per-allele genotyping error rates that were much higher than sequencing error rates, particularly at heterozygous sites that were wrongly inferred as homozygous. As we exemplify through the inference of genome-wide and local ancestry of well characterized hybrids of two widespread and intensively studied Eurasian poplar (*Populus*) species, such high error rates may easily lead to wrong biological conclusions. By properly accounting for these error rates in downstream analyses, either through the incorporation of genotyping errors directly, or by recalibrating genotype likelihoods, we were nevertheless able to use the RAD-seq data to support biologically meaningful and robust inferences of ancestry among *Populus* hybrids.

Introduction

Despite the impressive advancements in sequencing techniques and the decrease of related costs, whole genome sequencing (WGS) remains prohibitively expensive when working with a large number of samples or species with large genomes. Since many applications do not require information on the whole genome, reduced representation sequencing techniques are valuable alternatives and have become widely used for genome-wide SNP discovery and genotyping, especially in species with poor genomic resources (Narum *et al.* 2013; Andrews *et al.* 2016).

A commonly used reduced representation sequencing technique is Restriction site Associated DNA sequencing (Miller *et al.* 2007; Baird *et al.* 2008), which allows the sequencing of massively multiplexed samples at minimal costs by focusing on the sequences adjacent to restriction sites. Since restriction sites are often shared between individuals within a species and often also between closely related species (Cariou *et al.* 2013), focusing on adjacent sequences guarantees that sequenced loci are mostly overlapping across samples.

Briefly, the first step in the original RAD-seq protocol is the digestion of genomic DNA with a restriction enzyme. The resulting fragments are ligated to an adaptor and a unique barcode for each sample, and multiple individuals are pooled. The fragments are then sheared using a sonicator and those showing the proper size are selected and amplified through polymerase chain reaction (PCR). At this point the library is suitable for sequencing. By focusing the sequencing effort on tagged restriction sites, rather than on all randomly sheared genomic fragments (Rowe *et al.* 2011; Arnold *et al.* 2013), the number of markers can be customized through the choice of restriction enzymes. This choice will also influence which features of the genome are sampled, since certain enzymes preferentially cut in exonic regions, while others target intergenic and intronic regions (Arnold *et al.* 2013; Pootakham *et al.* 2016).

Several alternative RAD-seq protocols allow for ample customization of this methodology. These include the elimination of the sonication step (Andrews *et al.* 2016), ddRAD (Peterson *et al.* 2012), which uses two restriction enzymes rather than one, 2bRAD (Wang *et al.* 2012), which uses IIb-type restriction enzymes and produces fragments of 36 bp, and ezRAD (Toonen *et al.* 2013), in which DNA is digested with isoschizomers (a pair of restriction enzymes recognizing the same sequence). However, each method has different advantages and pitfalls, and a specific protocol may be more suitable for some applications than for others (Puritz *et al.* 2014). Due to this versatility, RAD-seq has been used in diverse applications, including the study of the genomics of adaptation (Andrews *et al.* 2016), hybridization and speciation (Marques *et al.* 2016), inbreeding depression (Hoffman *et al.* 2014), genetic associations (Nadeau *et al.* 2014), genetic mapping (Chutimanitsakun *et al.* 2011) and phylogeographic and phylogenomic analyses (Emerson *et al.* 2010; Leaché *et al.* 2015).

Despite this widespread use, genotypes called from RAD-seq data have been associated with several biases, many of which are specific to RAD-seq. Several major biases potentially affect alleles differently, which may lead to their unequal representation in sequencing data, and hence to genotyping errors at heterozygous sites (Davey *et al.* 2013). For instance, polymorphisms occurring in the restriction site may result in one allele not being cut and therefore not sequenced, potentially causing linked sites to be erroneously called homozygous (“allele dropout”). Polymorphisms at neighbouring restriction sites may also result in genotyping biases, for example, if the length of the fragment of one allele falls short of the selected size range (Andrews *et al.* 2016). Yet size differences among longer fragments were also found to result in unequal sequencing depth at linked sites because sonicators shear shorter fragments less efficiently than longer fragments (Sambrook & Russell 2006). Finally, the PCR step present in most RAD-seq protocols may contribute to genotyping errors through unequal amplification of the two alleles (Casbon *et al.* 2011) or through so-called PCR duplicates, the sequencing of misleading clonal copies of the same initial molecule (Andrews & Luikart 2014). Since many protocols produce single-end libraries or libraries where both ends are defined by restriction sites (e.g. ddRAD), PCR duplicates cannot be reliably identified bioinformatically unless very many different adapter sequences are used (Schweyen *et al.* 2014). This is particularly problematic in the case of PCR errors that might be sequenced in many copies, resulting in wrongly called heterozygous genotypes (Andrews *et al.* 2016).

Some consequences of these biases in downstream analyses are well documented. Gautier *et al.* (2013), for instance, found that under certain circumstances allele dropout leads to incorrect estimates of genetic diversity. Arnold *et al.* (2013) demonstrated through both simulations and real data that estimates of summary statistics commonly used to infer diversity and past demography from RAD data are severely affected by missing haplotypes and may show strong deviations from true values. Cariou *et al.* (2016), finally, illustrated that allele dropout can lead to underestimation of diversity, especially in highly polymorphic species.

In light of these potentially common issues, our aims here were to develop strategies (1) to estimate genotyping error rates in RAD data, and (2) to properly incorporate the resulting genotyping uncertainty in downstream analyses to mitigate the consequences of errors. For this we present methods to estimate genotyping errors in RAD-seq data in three different ways: First, by taking advantage of available genotyping data based on a different, more reliable method (e.g. using a chip or high-depth sequencing). Second, by using independent RAD-seq replicates of individuals. And third, by assuming Hardy-Weinberg proportions among population samples. Using simulations we show that all these methods are powerful in inferring error rates even if limited samples are available. We then applied these methods to RAD-seq data of the two widespread, genetically and ecologically divergent tree species *Populus alba* (White poplar) and *P. tremula* (European aspen) and inferred high genotyping error rates of multiple percent. By properly accounting for genotyping uncertainty, however, we obtain biologically meaningful estimates of genome-wide and local ancestry.

Materials and Methods

Estimation of genotyping error rates

Let us denote by g_{il} the observed genotype of individual $i = 1, ..., I$ at locus $l = 1, ..., L$, where $g_{il} = 0, 1, 2$ reflects the number of copies of the alternative allele at a bi-allelic locus. Given per-allele genotyping error rates ε_0 and ε_1 at homozygous and heterozygous sites, respectively, the probabilities $P(g_{il}|\varepsilon_0, \varepsilon_1)$ of observing genotype g_{il} are given in Table 1. We next present three strategies to estimate the genotyping error rates ε_0 and ε_1 from called genotypic data.

From a Truth Set. Consider a set of accurate genotypes γ_{il} obtained independently for a common set of individuals and loci. Assuming all γ_{il} to be correct and genotyping errors to be independent between sites and individuals, the likelihood of the observed genotypes $g = \{g_{11}, ..., g_{I1}, ..., g_I\}$ is then given by

$$P(g|\gamma, \varepsilon_0, \varepsilon_1) = \prod_{i=1}^{I} \prod_{l=1}^{L} P(g_{il}|\gamma_{il}, \varepsilon_0, \varepsilon_1),$$

where $P(g_{il}|\gamma_{il}, \varepsilon_0, \varepsilon_1)$ is given in Table 1 and and $\gamma = \{\gamma_{11}, ..., \gamma_{I1}, ..., \gamma_{IL}\}$. We obtain maximum likelihood estimates of ε_0 and ε_1 through numerical maximization (see Supplementary Methods).

From Individual Replicates. Consider a set of individuals for which multiple independent sequencing experiments were conducted. Let us denote by $g_{il}^{(j)}$ the inferred genotype of individual $i = 1, \ldots, I$ at locus $l = 1, \ldots, L$ in replicate $j = 1, \ldots, r_i$. The likelihood of the full data g is then given by

$$P(g|\varepsilon_0, \varepsilon_1) = \prod_{i=1}^{I} \prod_{l=1}^{L} \sum_{\gamma=0}^{2} \left[P(\gamma|f_{i\gamma}) \prod_{j=1}^{r_i} P(g_{il}|\gamma, \varepsilon_0, \varepsilon_1) \right],$$

where γ denotes the unobserved true genotype, $P(\gamma|f_{i\gamma}) = f_{i\gamma}$ denotes the frequency of genotype γ among all loci of individual i and $P(g_{il}|\gamma, \varepsilon_0, \varepsilon_1)$ is given in Table 1. We obtain maximum likelihood estimates of the parameters ε_0, ε_1 and $f = \{f_{10}, \ldots, f_{12}, \ldots, f_{I2}\}$ with an EM algorithm as detailed in the Supplementary methods.

From Population Samples. Consider a set of individuals $i = 1, \ldots, I$ sampled from a random mating population such that the distribution of the true genotypes at loci $l = 1, \ldots, L$ are well described by Hardy-Weinberg proportions. While the allele frequencies f_l are unknown, let us assume that they follow a Beta distribution with parameters α, β such that $f_l \sim Beta(\alpha, \beta)$, as is expected under neutrality (Wright 1931). The likelihood of the full data is then given by

$$P(g|\varepsilon_0, \varepsilon_1, \alpha, \beta) = \prod_{l=1}^{L} \int P(f_l|\alpha, \beta) \prod_{i=1}^{I} \sum_{\gamma=0}^{2} P(g_{il}|\gamma, \varepsilon_0, \varepsilon_1) P(\gamma|f) df_l,$$

where the sum runs over the unknown true genotype γ, $P(\gamma|f)$ are the Hardy-Weinberg proportions and $P(g_{il}|\gamma, \varepsilon_0, \varepsilon_1)$ is given in Table 1. To obtain estimates under this model we resort to an MCMC approach under a Bayesian scheme (see Supplementary methods) with exponential priors $\varepsilon_0, \varepsilon_1 \sim Exp(\lambda)$ truncated at 0.5 and normal priors $\log(\alpha), \log(\beta) \sim N(\mu, \sigma^2)$. We used $\lambda = 5$, $\mu = \log(0.5)$ and $\sigma^2 = 0.25$ throughout.

Error rate classes. All above methods are readily extended to jointly infer error rates for multiple classes, such as bins of sequencing depth or groups of samples if libraries were prepared in multiple experiments. Inferring the error rates of all classes jointly is beneficial in the case of individual replicates or population samples, as information about hierarchical parameters such as individual genotype frequencies $f_{i\gamma}$ or the parameters α, β of the Beta distribution are shared across classes. Here, this allows us to infer error rates of multiple bins of sequencing depth.

Recalibrating genotype likelihoods. We recalibrate genotype likelihoods by treating obtained genotype calls g_{il} as data and determining the likelihoods $P(g_{il}|\gamma_{il}, \varepsilon_0, \varepsilon_1)$ for all $\gamma_{il} = 0, 1, 2$ according to Table 1 and using parameter estimates ε_0 and ε_1 obtained for the relevant error rate class. If a truth set is available, we also calculate the empirical likelihoods $P(g_{il}|\gamma_{il} = g)$ across all loci with $\gamma_{il} = g$ of a particular error rate class.

Implementation: We implemented all algorithms developed here in the open-source C++ program *Tiger* (Tools to Incorporate Genotyping ERrors), available through the git repository at https://bitbucket.org/wegmannlab/tiger.

Simulations. We used simulations to assess the power of the methods introduced above to infer genotyping error rates. All simulations were generated directly under the assumed model using routines we implemented in *Tiger*. Under the truth set or replicate model, we quantified the power separately for homozygous and heterozygous sites. This was not possible under the Hardy-Weinberg model, for which we draw true allele frequencies from a Beta distribution with parameters $\alpha = \beta = 0.7$, implying about 29% of heterozygous genotypes.

Application to Populus *species*

Study system and plant material. We generated RAD-seq data of 139 individuals of the two widespread tree species *Populus alba* and *P. tremula*, and their hybrids (*P. x canescens*) in two sets. The first set consisted of 136 individuals (Supplementary Table S1) grown with minimal interference in a common garden established at the University of Fribourg (Switzerland) and previously genotyped by Lindtke et al. (2014). All these individuals grew from seeds collected from 15 mother trees in a natural hybrid zone in the Parco Lombardo della Valle del Ticino in Northern Italy where individuals of the two species and their hybrids grow side by side (Lindtke *et al.* 2012; Christe *et al.* 2016).

The second set consisted of four individuals for which we generated multiple replicates: a hybrid individual (F039_05) also included among the samples of the first set, a second hybrid individual (I373_A) also from the Ticino hybrid zone but grown in a common garden in Salerno, Italy, a pure *P. alba* individual (J1) from the Jalón river in the Ebro watershed (Northeast of the Iberian Peninsula), an assumed F1-hybrid tree (BET) from a population in the Tajo river headwaters (Central Iberian Peninsula). The two Iberian individuals were previously genotyped using microsatellites (Macaya-Sanz *et al.* 2011).

DNA extraction and RAD sequencing. For all samples, DNA was extracted from 15-20 mg of silica-dried leaf material with the Qiagen DNeasy Plant Mini Kit (Valencia, CA). The concentration of DNA was measured with a Qubit 2.0 Fluorometer using the dsDNA HS assay kit (Invitrogen), and its integrity verified with electrophoresis on 1.5% agarose gels (1X TBE). Concentrations were standardized to 20 ng/µl and individual samples were submitted for library preparation and Restriction site Associated DNA sequencing (RAD-seq) to Floragenex (Eugene, OR). There, all extractions of the individuals of the first set, as well as the replicate extractions of F039_05 and I373_A, were processed (together with additional samples prepared in the same way), in five libraries of 95 individuals each. These libraries were prepared according to Floragenex' standard commercial protocol: genomic DNA was digested with the restriction endonuclease *PstI* (chosen according to previous studies on these species - Stölting *et al.* 2013; Christe *et al.* 2016) and RAD libraries were prepared with a method similar to the one described in Baird *et al.* (2008). This protocol included 18 PCR cycles, after which DNA fragments ranging from 300 to 500 bp were retained. All five libraries were sequenced in a single run on an Illumina HiSeq2500 instrument, but on individual lanes.

Following the same protocol, an additional library was generated and sequenced by Floragenex in a separate experiment, consisting of two and three replicate extractions of J1 and BET, respectively, as well as extractions from offspring of a controlled cross between them.

Bioinformatic data processing. We assigned reads to individuals or replicates with *fastq-multx* (ea-utils; Aronesty 2011), allowing one mismatch in the 15 bp including barcode and restriction site. Read quality was checked with FastQC 0.10.1 (Andrews 2010) and low quality bases and reads were removed with *condetri* v.2.3 (Smeds & Künstner 2011) using default parameters, except for the options -hq (high quality threshold) and -lfrac (maximum acceptable fraction of bases after quality trimming with quality scores lower than the threshold -lq), for which a value of 15 and 0.1 were chosen, respectively.

Good quality reads were aligned against the *P. tremula* mitochondrial reference sequence (Kersten *et al.* 2016) and against the nuclear reference genome of *P. trichocarpa* (Ptrichocarpa_210_v3.0; Tuskan *et al.* 2006) using Bowtie2 2.3.0 (Langmead & Salzberg 2012) with "end-to-end" and "very sensitive" settings. Reads with mapping quality lower than 20 were discarded using samtools 1.3 (Li *et al.* 2009) and read group information was added with picard tools 1.139 (http://broadinstitute.github.io/picard). We then used the tools TargetCreator and IndelRealigner of GATK 3.8 (DePristo *et al.* 2011) to realign around indels, and recalibrated base quality scores for each individual using the method by Kousathanas *et al.* (2017) implemented in ATLAS (Link *et al.* 2017) on mitochondrial sequences. This method does not require *a priori* information on genotyping information and instead learns base qualities from haploid regions while integrating over genotype uncertainty. Finally, we called genotypes with UnifiedGenotyper in GATK 3.8 (DePristo *et al.* 2011).

To then only retain reliable sites for comparison, we filtered resulting variants using vcftools (Danecek *et al.* 2011) and custom R scripts: first, we removed sites with an average depth across individuals ≥24 (the 98.7% quantile) to exclude potentially paralogous loci. Second, we only kept variants with at most two segregating alleles. Third, we removed indels and variant sites within 5 bp of an indel to avoid Single Nucleotide Variants (SNVs) originating from misalignments.

Truth Set. Genotypes of the 136 individuals grown in the common gardens were previously obtained (Lindtke *et al.* 2014) with a genotyping-by-sequencing (GBS) protocol very similar to the ddRAD protocol (Peterson *et al.* 2012) and using the restriction enzymes *EcoRI* and *MseI*. Importantly, Lindtke et al. (2014) generated sequencing data also for the 15 mother trees and used sibships in a Bayesian approach to infer genotypes while accounting for familial relationships.

To compare these high quality genotypes to those obtained from our own RAD-seq experiment, loci covered in both studies had to be identified first. Since Lindtke et al. (2014) used an older *P. tremula* reference, we extracted from this reference windows of 201 bp around each locus in the GBS data set (100 bp on either side). We then mapped these extracted sequences against the *P. trichocarpa* reference with Bowtie2 2.3.0 (Langmead & Salzberg 2012) with "end-to-end" and "very sensitive" settings and retained only those sequences that mapped uniquely with quality of 20 or more. We then kept all loci overlapping between the two data sets, but removed four loci for which different alternative alleles were called. To ensure high accuracy of the GBS data, we restricted all comparisons to genotypes called with a posterior probability ≥ 99% by Lindtke et al. (2014).

Estimation of genome-wide and interspecific ancestry. We estimated genome-wide (*q*) and interspecific (*Q_{12}*) ancestry for the 136 common garden samples using *entropy* (Gompert *et al.* 2014), a program that implements a model similar to the admixture model in *structure* (Pritchard *et al.* 2000). In contrast to *structure*, however, *entropy* can also make use of uncertain genotypes from low depth sequence data by working directly with genotype likelihoods, rather than genotype calls. Here we ran *entropy* on the raw genotype likelihoods, as well as on genotype likelihoods recalibrated using empirical likelihoods, excluding sites with >50% missing data in both cases. To stratify the estimates and have sufficient observations to estimate these probabilities reliably, we considered five RAD-seq depth classes: 1-3, 4-7, 8-15, 16-31 and ≥32.

Inference of locus-specific ancestry. To infer locus-specific ancestry, we ran RASPberry (Wegmann *et al.* 2011), which implements a Hidden Markov Model (HMM) to explain haplotypes of admixed individuals as a mosaic of provided reference haplotypes for each species. We obtained suitable reference haplotypes by phasing previously characterized pure *P. alba* and pure *P. tremula* individuals (51 each) from the Italian, Austrian and Hungarian hybrid zones (Christe *et al.* 2016) using FastPhase (Scheet & Stephens 2006), building input files with fcGENE (Roshyara & Scholz 2014). For use in RASPberry, individuals in the reference panels were not allowed to have missing data. We thus restricted the comparison to only the SNVs covered in all parental individuals.

To compute HMM transition probabilities in RASPberry, we used a default recombination rate of 5 cM/Mb as estimated by Tuskan *et al.* (2006) in *P. trichocarpa* and the estimates of the genome-wide ancestry *q* for each sample obtained with *entropy* from the error corrected data. For most other parameters we used previous estimates for *P. alba* and *P. tremula* hybrid zones (Christe et al. 2016), but scaled these as proposed by Wegmann et al. (2011) to reflect the size of the reference panel. These include the ancestral population recombination rates (315 and 900 for *P. alba* and *P. tremula*, respectively), mutation rates (0.00185 and 0.00349, respectively) and the miscopying rate (0.01). However, we set the time since admixture to five (rather than one) to reflect the different sampling strategy.

To account for genotyping errors, we estimated a per-allele genotyping error ε under the truth-set model with the constraint $\varepsilon_0 = \varepsilon_1 = \varepsilon$, and then added this estimate to the miscopying rate and the two mutation rates. Under the RASPberry copying model, these parameters control the rate at which the sample genotypes differ from the reference haplotype from which the sample is copying. That rate thus depends on the reference panel size, but also on genotyping errors.

We called ancestry segments as any stretch on a chromosome within which the posterior probabilities for a particular ancestry (homozygous *P. alba*, heterozygous ancestry or homozygous *P. tremula*) was > 0.5 at all SNVs, and measured its length from the first to the last SNV.

Results

Power to infer genotyping error rates

Simulations suggest that a few thousand loci are sufficient to accurately estimate genotyping error rates even from just a few samples (Figure 1). Due to the extra information provided, the smallest estimation errors were obtained when using a truth set: >90% of all estimates fell within a range from half to two-fold the true value (Q2, e.g., within [0.005, 0.02] for a true error rate of 0.01) if estimated genotypes were compared at 100 truly homozygous and 200 truly heterozygous genotypes for ε_0 and ε_1, respectively.

Similar accuracy was achieved under the Hardy-Weinberg model as soon as 5,000 sites were used. However, the accuracy is a function of the fraction of truly heterozygous genotypes in the data set, with the accuracy of ε_0 being much higher than for ε_1 if much fewer than 50% of all genotypes are heterozygous, and vice versa. Here, we simulated about 29% heterozygous genotypes and thus expect accuracy to be higher if a larger fraction of genotypes were heterozygous.

The lowest accuracy was observed under the replicates model, especially if error rates were low. Using 10^4 comparisons, for instance, all estimates were within Q2 for a true value of $\varepsilon_0 = \varepsilon_1 = 0.1$, but only slightly above 70% of all estimates for a true value of $\varepsilon_0 = \varepsilon_1 = 0.01$. This is readily explained by the fact that only very limited information about the true genotype is available: if the two replicates differ in their genotype, it is not clear which one is correct. Consequently, the accuracy of inference is much increased if more than two replicates are available per individual (Supplementary Figure S1). Nonetheless, even small error rates can be estimated relatively accurately, as >90% of all estimates for true value of $\varepsilon_0 = \varepsilon_1 = 0.01$ fell within Q2 as soon as $5{\cdot}10^4$ or more comparisons were used. Assuming 20% of all considered genotypes to be heterozygous, around 10^5 sites are required if two pairs of replicates are used.

High genotyping error rates in RAD-seq

We next used our inference methods to quantify genotyping errors from our RAD-seq experiment of 137 individuals of the two widespread tree species *Populus alba* and *P. tremula*, and their hybrids (*P. x canescens).* On average, our experiment resulted in 831,160.66 (sd 153,433.62) reads per sample that passed quality trimming and mapped against the reference genome of *P. trichocarpa* with mapping quality ≥20. From those, we called 529,305 Single Nucleotide Variants (SNVs), after removing multi-allelic sites, those with excess depth, indels and variant sites around indels. We estimated per-allele genotyping error rates from these SNVs through a comparison with previously published, high-quality genotypes (truth set), and from multiple replicate libraries sequenced for a subset of our samples (replicates).

Truth set. We estimated per-allele genotyping errors by comparing genotype calls from our RAD-seq experiment to those of a previously published GBS dataset (Lindtke *et al.* 2014) for 136 individuals present in both studies. In total, 7,426 SNVs overlapped between experiments, at which we could use a total of 16,610 genotype comparisons. Of those, only 69.9% matched, with matching rates increasing with RAD-seq depth (Figure 2A). Strikingly, RAD-seq genotypes were much less often heterozygous than GBS genotypes (Figure 2B), especially at low depth. In line with these observations, we inferred per-allele genotyping error rates > 10% whenever RAD-seq depth was ≤ 35 and when using a model assuming a single error rate ($\varepsilon_0 = \varepsilon_1$), driven by an exceptionally high error rate at truly heterozygous sites ε_1 (Figure 2C).

These are surprisingly high error rates, particularly when considering that sequencing error rates of Illumina machines are estimated at < 1% (Nielsen *et al.* 2011). Importantly, the bias towards homozygous genotype calls is not simply explained by low depth. Indeed, RAD-seq still resulted in less then half as many heterozygous calls at depths ≥ 40x, which are usually considered more than sufficient for accurate genotype calling (Nielsen *et al.* 2011). Instead, our results suggest an inherent bias in the RAD-seq data analyzed here.

However, our estimates rely on the assumption that the GBS data reflect true genotypes. This is based on good evidence: First, Lindtke *et al.* (2014) additionally sequenced the mother trees of all individuals considered here and estimated posterior genotypes using a hierarchical ancestry model that incorporated familial relationships with mothers and among siblings. These updated estimates correlated with the raw maximum likelihood genotype estimates ignoring familial data at 0.985 and differed from those in < 0.02% of all calls. Second, we restricted this comparison to GBS genotypes with a posterior probability ≥ 99%. Third, the fraction of concordant genotype calls between the GBS and RAD-seq data increased with RAD-seq depth. If the mismatches were driven by errors in the GBS data, no such dependence should be observed.

Replicates. We next benefitted from two sets of replicate libraries to estimate per-allele genotyping error rates. The first set consisted of two replicate libraries of each of two individuals (*F039_05 and I373_A*) sequenced along all other samples. Error rates estimated from that data corroborated the conclusion obtained from the comparison with GBS data (Figure 2D): error rates at truly homozygous sites (ε_0) were on the order of 1% or less, and those at truly heterozygous sites (ε_1) were equal or close to 50%, which is the largest value possible under our model.

In contrast to the estimates obtained in comparison to GBS genotypes, the error rates at truly heterozygous sites (ε_1) dropped to about 20% at high depth (≥ 20x). This difference might in part be driven by errors in the GBS data slightly inflating error rate estimates. However, given the high quality of the GBS data, it appears more likely that the error rates from replicates are underestimated. Polymorphisms in restriction cut sites or unequal PCR amplification rates of alleles, for instance, affect replicates systematically, while the statistical inference must assume independence of errors between replicates.

To verify that high error rates are not specific to the RAD-seq run performed on these hybrids, we carried out a second RAD-seq experiment including two and three replicates of a pure *P. alba* individual (J1) and a putative F1 *P. alba* x *P. tremula* hybrid (BET), respectively. This experiment resulted in 694,030.80 (sd 187,957.33) reads per sample that passed quality trimming and mapped against the reference genome of *P. trichocarpa* with quality ≥20.

Per-allele error rates estimated from this data were indeed lower, with error rates at truly homozygous sites (ε_0) on the order of 0.2% and those at truly heterozygous sites (ε_1) starting out at 50% and dropping to about 3% at a depth of 25x (Figure 2D). However, these lower errors still point to a particular issue in calling heterozygous genotypes: of all truly heterozygous sites with depth ≥ 25x in our data, > 5% are expected to be called homozygous. (The lower output of this sequencing experiment does not allow us to make reliable statements for higher depths).

Estimation of genome-wide and interspecific ancestry

We investigated the impact of genotyping errors in our RAD-seq data on the inference of genome-wide ancestry q as well as interspecific ancestry Q_{12}, which reflects the proportion of loci of heterospecific ancestry. We estimated these ancestry components with *entropy* from 230,805 SNVs, and compared them to estimates from GBS obtained by Lindtke *et al.* (2014).

The model implemented in *entropy* accounts for the genotyping uncertainty reflected in the genotype likelihoods. However, the raw genotype likelihoods obtained from our RAD-seq data are misleading: for sites with considerable depth, the RAD-seq genotype likelihoods often suggest almost certainty for wrong genotypes. Of all genotypes wrongly called as homozygous at depth ≥30x (judged by the comparison to GBS data), 90% had a variant quality of 77 or more. (A variant quality of 77 implies that it is more than 5 billion times less likely to observe the obtained data from a heterozygous than homozygous site).

As a result, ancestry estimates differed considerably between the GBS and RAD-seq data sets (Figure 3). Interestingly, the estimates of the genome-wide ancestry q were much less affected by genotyping errors than the estimates of the interspecific ancestry Q_{12}. This is readily explained, however, by the directionality of the most common error, which is to wrongly infer homozygous genotypes at heterozygous sites. We found these errors to result more frequently in a homozygous reference than homozygous alternative call (65.0%), particularly at low depth (84.5% at depth ≤5x, 49.9% at depth ≥20x). But this did not introduce a bias in q towards one of the species since we were using the reference sequence of the outgroup *P. trichocarpa*. The estimates of Q_{12}, however, are very sensitive to an underestimation of heterozygosity.

To improve these estimates, we propose to directly account for the elevated genotyping error rates in RAD-seq data by adjusting the genotype likelihoods according to the observed genotyping uncertainty. By treating the genotype calls as data, we can determine the probabilities $P(g|\gamma)$ of observing a RAD-seq genotype call g given the true genotype γ either by using estimates of the per-allele error rates $\varepsilon_0, \varepsilon_1$, or empirically from the comparison to a truth set. Using the latter approach on our data (individually for each depth) resulted in estimates of q and Q_{12} that were much closer to those obtained by Lindtke *et al.* (2014; Figure 3).

Some differences between the point estimates of Q_{12} remain, likely due to differences in the models and their information-sharing among individuals (in Lindtke *et al.* (2014), information about maternal plants and sibships was part of the model) and the extent to which information in uncertain genotypes was outweighed by hierarchical prior probabilities related to ancestry. Nonetheless, these results demonstrate the importance of accounting for uncertainty in genotyping data since the estimates of interspecific ancestry with and without correction lead to a very different biological interpretation: With correction, the hybrid individuals appear to be mostly early generation hybrids, suggesting meaningful reproductive isolation between the species. Without correction, the large number of individuals with low Q_{12} but intermediate values of q suggest considerable gene flow between the species, an interpretation at odds with recent work (Macaya-Sanz *et al.* 2011; Christe *et al.* 2016, 2017).

Inference of locus-specific ancestry

We next evaluated the impact of genotyping errors on local ancestry inference. Hybrid zones between *P. alba* and *P. tremula* are dominated by pure parental individuals and early hybrids (mostly F1), with only few adult recombinant hybrids (Lindtke *et al.* 2014; Christe *et al.* 2016). We chose ten individuals among our samples spanning that spectrum according to q and Q_{12} values from Lindtke *et al.* (2014): a putatively pure *P. alba* individual (F039_01), a putatively pure *P. tremula* individual (F030_01), two putative backcrosses to *P. alba* (F020_04 and F032_08), two putative backcrosses to *P. tremula* (I345_02 and I345_03), two putative F1 hybrids (I373_03 and F030_05) and two putative hybrids of later generations (Fn - F022_03 and F026_05). We then used *RASPberry* (Wegmann *et al.* 2011) to infer local ancestry along chromosomes 1 through 5, restricting our inference to the 6,445 SNVs that did not have missing data in the parental reference haplotypes we took from Christe *et al.* (2016).

In line with these expected simple ancestry make-ups, we inferred many large ancestry blocks often spanning almost entire chromosomes (Figure 4). Surprisingly, however, we inferred most of these blocks to be of homospecific ancestry, and also inferred many short segments, which is difficult to reconcile with the putative ancestries of our samples (Figure 4). As an example, consider the individual I373_03 in Figure 4 that was classified as an F1 hybrid by Lindtke et al. (2014), but for which we inferred homospecific ancestry blocks for both parental species. Such artifacts could arise from the reference panels being too small to properly reflect the haplotypes found in our hybrid individuals, large gaps between neighboring SNVs limiting the power of the HMM implemented in RASPberry, but most likely by genotyping errors towards homozygous genotypes.

While *RASPberry* does not account for genotyping uncertainty via genotype likelihoods, the implemented copying-model allows for “mutations”, or differences between the observed genotype of an admixed individual and the reference haplotypes from which it is copying. We thus repeated the inference by adding an average per-allele genotyping error rate of 13.7% (weighted average across depth ≥5x) to the mutation rate parameters to account for the high genotyping error in our RAD-seq data (Figure 4). Accounting for genotyping errors indeed improved our estimates. For the putative backcrosses, for instance, we called fewer segments homozygous for the “wrong” ancestry (11 versus 34) and these covered a smaller fraction of the first five chromosomes (6.1% versus 11.1% of the parts at which ancestry could be called). Similarly, we called a higher fraction of the first five chromosomes to be of heterozygous ancestry (45.0% versus 37.2% of the parts at which ancestry could be called). While these results corroborate the importance of accounting for the true uncertainty in genotypes in downstream analysis, they also illustrate that a method accounting for a uniform error fails to fully mitigate the bias against heterozygous genotypes present in our RAD-seq data sets.

Discussion

Here, we report high genotyping error rates in two independent RAD-seq data sets. We obtained these estimates by comparing RAD-seq calls of two independent experiments, either to published genotype calls for the same individuals (Lindtke *et al.* 2014) obtained with a different sequencing method (GBS), or to calls obtained from independent replicates of the same individuals. Both approaches provide evidence for high per-allele genotyping errors of several percent and show that RAD-seq has a strong bias towards calling homozygous genotypes at heterozygous sites that is not overcome with higher sequencing depth.

Only few studies have reported estimates of genotyping errors of reduced representation techniques to date, but all agree with the high estimates obtained here. Luca *et al.* (2011), for instance, compared genotypes of human samples obtained with a technique similar to RAD to those available in a public database, and estimated that between 6.3 and 9.7% of heterozygous sites were called as homozygous. Similarly, Mastretta-Yanes *et al.* (2015) found that, depending on the parameter settings chosen for the *de-novo* assembly, between 5.9 and 8.8% of alleles were not concordantly called between replicates.

Several factors could explain the high genotyping error and the lack of heterozygous genotypes in RAD-seq data. For example, one allele might not have been sequenced or sequenced only at very low depth because of differences in fragment length that can lead to amplification bias, less efficient shearing, or loss in size selection. This is a likely explanation, since Davey and colleagues (2013) showed that there is a high correlation between read depth and fragment length. Similarly, differential efficiency of PCR among alleles could have masked one allele (i.e., PCR duplicates), causing it to be represented in a very low number of reads (Schweyen *et al.* 2014). Finally, the well-known issue of allele dropout due to polymorphisms in the restriction site and the “loss” of one allele at heterozygous sites may have contributed to the inaccurate, low observed heterozygosity (Davey *et al.* 2013; Puritz *et al.* 2014; Andrews *et al.* 2016). However, this problem is a less likely explanation as it should not affect RAD-seq at a higher rate than the double-digest GBS protocol we used for the error estimation.

Several bioinformatic solutions have been suggested to mitigate the apparent biases in RAD-seq. Both Arnold *et al.* (2013) and Gautier *et al.* (2013) recommend the comparison of read depth across sites, to identify loci likely exhibiting allele dropout. In our case, however, depth varied substantially across sites, because of PCR duplicates or stochastic events, rendering such an approach difficult. Davey *et al.* (2013) also noted that alleles present in two copies at homozygous sites have higher depth compared to alleles present in single copy at heterozygous loci, but read depths for the two sets of alleles overlap, inhibiting the accurate detection of loci with allele dropout by using depth alone. To improve upon this, Cooke *et al.* (2016) developed a method to infer the likelihood of observing allele dropout at a site on the basis of the coverage of each sample, and suggested to ignore sites where this likelihood is high. Finally, it was also suggested to discard any locus with a missing genotype, since this might indicate a polymorphism in the restriction site. In many studies with moderate depth, including ours, the amount of missing data prevents the adoption of such drastic solutions. In summary, all these filtering suggestions result in a massive reduction in usable loci, and hence further accentuate the already limited genome-wide coverage of reduced library techniques such as RAD.

As a model-based alternative, we propose here to properly account for the high genotyping errors in downstream analysis. A first such attempt was recently proposed by Cariou *et al.* (2016), who developed an Approximate Bayesian Computation (ABC) method to estimate genetic diversity while accounting for allele dropout, but found this method not to be accurate under elevated levels of diversity. A more general solution, we believe, is to make use of the large number of recently developed tools that do not require genotype calls but rather work directly from genotype likelihoods to account for uncertainty in the data (Fumagalli *et al.* 2014; Korneliussen *et al.* 2014; Kousathanas *et al.* 2017; Jørsboe *et al.* 2017). Using such tools minimizes the necessity to filter data stringently and is readily applied to low-depth data (Nielsen *et al.* 2011).

However, for such methods to work properly, the genotype likelihoods need to accurately reflect the uncertainty in genotypes. While all modern genotype callers also calculate genotype likelihoods, these do not reflect biases specific to individual sequencing protocols such as RAD-seq, as we illustrate here, and must thus be recalibrated. Here we propose two recalibration strategies: If accurate genotype calls are available for a subset of the individuals and markers, empirical genotype likelihoods can be obtained by comparing those to calls from a reduced representation sequencing experiment. Alternatively, sample replicates may be used to infer per-allele genotyping error rates, from which recalibrated genotype likelihoods are readily calculated. Tools for both of these strategies are available through the software *Tiger*, which also accounts for sequencing depth as an additional covariate. While we found sequencing depth to be a particularly important predictor, the model is also readily extended to additional covariates such as the raw genotype likelihood or genotype call, which might provide additional information about genotyping error rates.

However, both strategies might be biased. Genotyping errors in a set of genotype calls considered to be accurate (the truth set), for instance, will result in an overestimation of genotyping error rates. Consistent biases in genotype calls affecting replicates similarly, on the other hand, might be difficult to infer and result in an underestimation of genotyping error rates. But despite these caveats, recalibrated genotype likelihoods are likely reflecting genotype uncertainty much more accurately, particularly for protocols with error rates as high as those we report here for RAD-seq. Indeed, we found here that genotype recalibration was essential to avoid drawing inaccurate conclusions and instead recovered biologically meaningful results about the ancestry of *Populus* hybrids.

We also note that if no tools accepting genotype likelihoods as input information are available for specific applications, this should not discourage users from incorporating genotyping uncertainty in the analyses. We have shown here that local ancestries were more reliably estimated by RASPberry (Wegmann *et al.* 2011), a tool requiring genotype calls, when adding the estimated genotyping error rate to the parameters of the model. But we note that given the particular lack of heterozygous genotypes in the RAD-seq data analyzed here, a model using a single per-genotype error rate as implemented in RASPberry was not sufficient to overcome all biases.

In conclusion, and in line with others (Mastretta-Yanes *et al.* 2015; Cooke *et al.* 2016; Cariou *et al.* 2016), we strongly suggest to carefully assess genotyping error rates in reduced representation sequencing experiments, and to properly account for those in downstream analyses, for instance using the tools we present here through *Tiger*. For this purpose, we recommend to either sequence a subset of individuals and markers at much higher quality, or to include sufficient replicates, from which genotyping error rates can be inferred. Knowledge on these error rates then allows to properly account for genotyping errors in downstream analyses, rather than losing a large amount of information due to stringent filtering. However, with ever dropping costs for sequencing and library preparation, low-depth whole-genome sequencing may become a valuable alternative in many applications. Indeed, simulations have shown that low-depth data spanning a larger fraction of the genome yield accurate and precise estimates of population genetics parameters (e.g. Buerkle & Gompert 2013; Kousathanas *et al.* 2017; Rustagi *et al.* 2017). The problem of high error rates in RAD-seq or other reduced representation libraries is thus likely transient and we expect that the field will quickly adopt new sequencing technologies that circumvent it entirely.

Acknowledgements

We thank Kai N. Stölting, Camille Christe and Margot Paris for helpful insights and discussions, Thelma Barbará for help in the laboratory, David Frey for collecting seeds in the hybrid zone, and Santiago González Martínez for providing the Spanish samples. This work was supported by grant 31003A_149306 of the Swiss National Foundation to CL. All calculations were done on Vital-IT and the Bioinformatic Core Facilities of the Universities of Bern and Fribourg, and computers at the University of Wyoming.

Data Accessibility

The code of Tiger is available through a git repository at https://bitbucket.org/wegmannlab/tiger. The RAD-seq data are available on the Sequence Read Archive through bioprojects PRJNA528699 and PRJNA528706. The called genotypes used to estimate genotyping errors are available at Zenodo (DOI 10.5281/zenodo.2604109 and 10.5281/zenodo.2604124).

Author Contributions

LB and DW conceived the study; CL and DW provided funding; LB collected genetic data; LB, CAB, VL and DW performed the analyses; CAB, CL and DW supervised the study; LB and DW wrote the manuscript with input and revisions from all co-authors.

References

Andrews S (2010) *FastQC: a quality control tool for high throughput sequence data*.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature reviews. Genetics*, **17**, 81–92.

Andrews KR, Luikart G (2014) Recent novel approaches for population genomics data analysis. *Molecular ecology*, **23**, 1661–1667.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology*, **22**, 3179–3190.

Aronesty E (2011) *ea-utils: Command-line tools for processing biological sequencing data*.

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

Buerkle AC, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology*, **22**, 3028–3035.

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and evolution*, **3**, 846–852.

Cariou M, Duret L, Charlat S (2016) How and how much does RAD-seq bias genetic diversity estimates? *BMC evolutionary biology*, **16**, 240.

Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic acids research*, **39**, e81.

Christe C, Stölting KN, Bresadola L *et al.* (2016) Selection against recombinant hybrids maintains reproductive isolation in hybridizing Populus species despite F1 fertility and recurrent gene flow. *Molecular ecology*, **25**, 2482–2498.

Christe C, Stölting KN, Paris M *et al.* (2017) Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Molecular ecology*, **26**, 59–76.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A *et al.* (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *BMC genomics*, **12**, 4.

Cooke TF, Yee M-C, Muzzio M *et al.* (2016) GBStools: A Statistical Method for Estimating Allelic Dropout in Reduced Representation Sequencing Data. *PLoS genetics*, **12**, e1005631.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* , **27**, 2156–2158.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular ecology*, **22**, 3151–3164.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491–498.

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

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* , **30**, 1486–1487.

Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular ecology*, **22**, 3165–3178.

Gompert Z, Lucas LK, Alex Buerkle C *et al.* (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular ecology*, **23**, 4555–4573.

Hoffman JI, Simpson F, David P *et al.* (2014) High-throughput sequencing reveals inbreeding depression in a natural population. *Proceedings of the National Academy of Sciences*, **111**, 3775–3780.

Jørsboe E, Hanghøj K, Albrechtsen A (2017) fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* , **33**, 3148–3150.

Kersten B, Faivre Rampant P, Mader M *et al.* (2016) Genome Sequences of Populus tremula Chloroplast and Mitochondrion: Implications for Holistic Poplar Breeding. *PloS one*, **11**, e0147209.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics*, **15**, 356.

Kousathanas A, Leuenberger C, Link V *et al.* (2017) Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, **205**, 317–332.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**, 357–359.

Leaché AD, Chavez AS, Jones LN *et al.* (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome biology and evolution*, **7**, 706–719.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , **25**, 2078–2079.

Lindtke D, Buerkle CA, Barbará T *et al.* (2012) Recombinant hybrids retain heterozygosity at many loci: new insights into the genomics of reproductive isolation in Populus. *Molecular ecology*, **21**, 5042–5058.

Lindtke D, Gompert Z, Lexer C, Buerkle CA (2014) Unexpected ancestry of Populus seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. *Molecular ecology*, **23**, 4316–4330.

Link V, Kousathanas A, Veeramah K *et al.* (2017) ATLAS: Analysis Tools for Low-depth and Ancient Samples.

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.

Macaya-Sanz D, Suter L, Joseph J *et al.* (2011) Genetic analysis of post-mating reproductive barriers in hybridizing European Populus species. *Heredity*, **107**, 478–486.

Marques DA, Lucek K, Meier JI *et al.* (2016) Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLoS genetics*, **12**, e1005887.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular ecology resources*, **15**, 28–41.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, **17**, 240–248.

Nadeau NJ, Ruiz M, Salazar P *et al.* (2014) Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene and H. erato. *Genome research*, **24**, 1316–1333.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular ecology*, **22**, 2841–2847.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, **12**, 443–451.

Parchman TL, Gompert Z, Mudge J *et al.* (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 non-model species. *PloS one*, **7**, e37135.

Pootakham W, Sonthirod C, Naktang C *et al.* (2016) Effects of methylation-sensitive enzymes on the enrichment of genic SNPs and the degree of genome complexity reduction in a two-enzyme genotyping-by-sequencing (GBS) approach: a case study in oil palm (Elaeis guineensis). *Molecular breeding: new strategies in plant improvement*, **36**, 154.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular ecology*, **23**, 5937–5942.

Roshyara NR, Scholz M (2014) fcGENE: a versatile tool for processing and transforming SNP datasets. *PloS one*, **9**, e97589.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular ecology*, **20**, 3499–3502.

Rustagi N, Zhou A, Watkins WS *et al.* (2017) Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. *BMC genomics*, **18**, 396.

Sambrook J, Russell DW (2006) Fragmentation of DNA by Sonication. *Cold Spring Harbor protocols*, **2006**, db.prot4538–pdb.prot4538.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, **78**, 629–644.

Schweyen H, Rozenberg A, Leese F (2014) Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological bulletin*, **227**, 146–160.

Smeds L, Künstner A (2011) ConDeTri - A Content Dependent Read Trimmer for Illumina Data. *PloS one*, **6**, e26314.

Stölting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular ecology*, **22**, 842–855.

Toonen RJ, Puritz JB, Forsman ZH *et al.* (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Tuskan GA, Difazio S, Jansson S *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.

Wang S, Meyer E, McKay JK, Matz MV (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature methods*, **9**, 808–810.

Wegmann D, Kessner DE, Veeramah KR *et al.* (2011) Recombination rates in admixed individuals identified by ancestry-based inference. *Nature genetics*, **43**, 847–853.

Wright S (1931) Evolution in Mendelian Populations. *Genetics*, **16**, 97–159.

Tables

Table 1*. Per-allele genotyping error model. Shown are the probabilities of observing a RAD*-seq *genotype given the the true genotype and the per-allele genotyping error rate ε.*

	Observed genotype		
True genotype	0	1	2
0	$(1-\varepsilon_0)^2$	$2\varepsilon_0 * (1-\varepsilon_0)$	$\varepsilon_0^{\ 2}$
1	$\varepsilon_1 * (1-\varepsilon_1)$	$(1-\varepsilon_1)^2 + \varepsilon_1^{\ 2}$	$\varepsilon_1 * (1-\varepsilon_1)$
2	$\varepsilon_0^{\ 2}$	$2\varepsilon_0 * (1-\varepsilon_0)$	$(1-\varepsilon_0)^2$

Figures

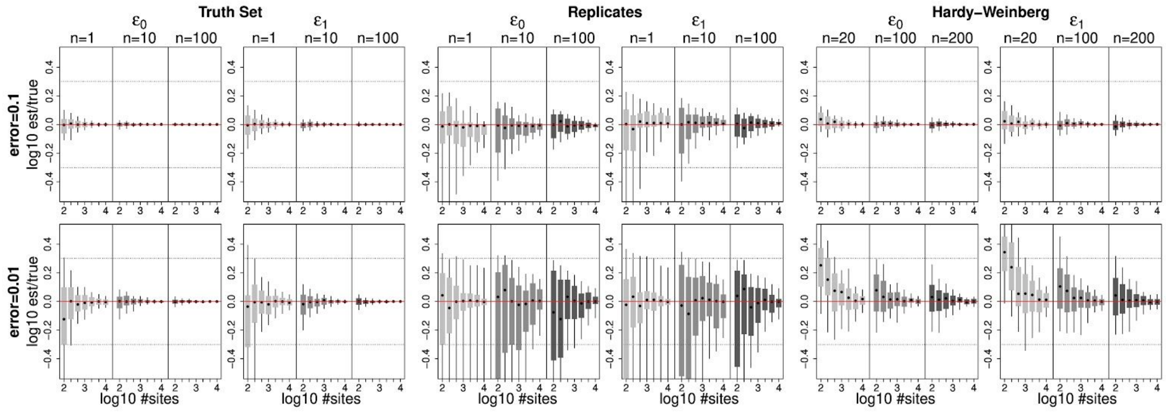

Fig. 1: *Accuracy of error rate estimates. Shown are the estimated error rates relative to the true error rates (red line) of 100 replicates for different samples sizes n (shown on top), the two error rates 0.1 and 0.01 (top and bottom row, respectively) and different numbers of unlinked loci. The horizontal dashed lines indicate Q2, the interval within which an estimate is less than a factor of two away from the true value (i.e. within half and two times the true value). Simulations were generated for the truth set (left), the individual replicate (middle) and Hardy-Weinberg models (right).*

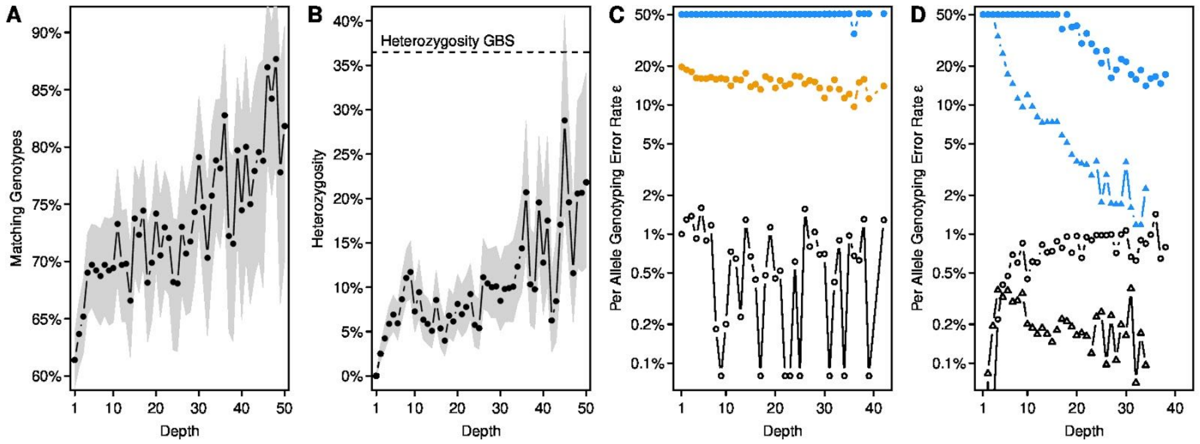

Fig. 2: *Estimates of genotyping errors in RAD-seq data.* **A**. *Maximum likelihood estimate (MLE) of the frequency of genotypes estimated from GBS and RAD-seq data that match as a function of RAD-seq sequencing depth. The gray region indicates the range of frequencies within two log-likelihood units of the MLE.* **B.** *MLE estimate of the frequency of heterozygous calls with RAD-seq among all calls of a particular depth. Gray region as in A. The heterozygosity observed among all GBS genotypes is given as dashed line.* **C**. *Per-allele RAD-seq genotyping error rates for homozygous (ε_0, black open symbols) and heterozygous (ε_1, blue filled symbols) genotypes estimated using a GBS truth set, limited to depth class with at least 100 comparisons. Estimates for a model assuming* $\varepsilon_0 = \varepsilon_1$ are shown with yellow closed circles. **D.** *Per-allele RAD-seq genotyping error rates obtained from two replicates of F039_05 and I373_A each (circles) and two and three replicates from J1 and Bet, respectively (triangles), limited to depth class with at least 1000 sites. Colors as in C.*

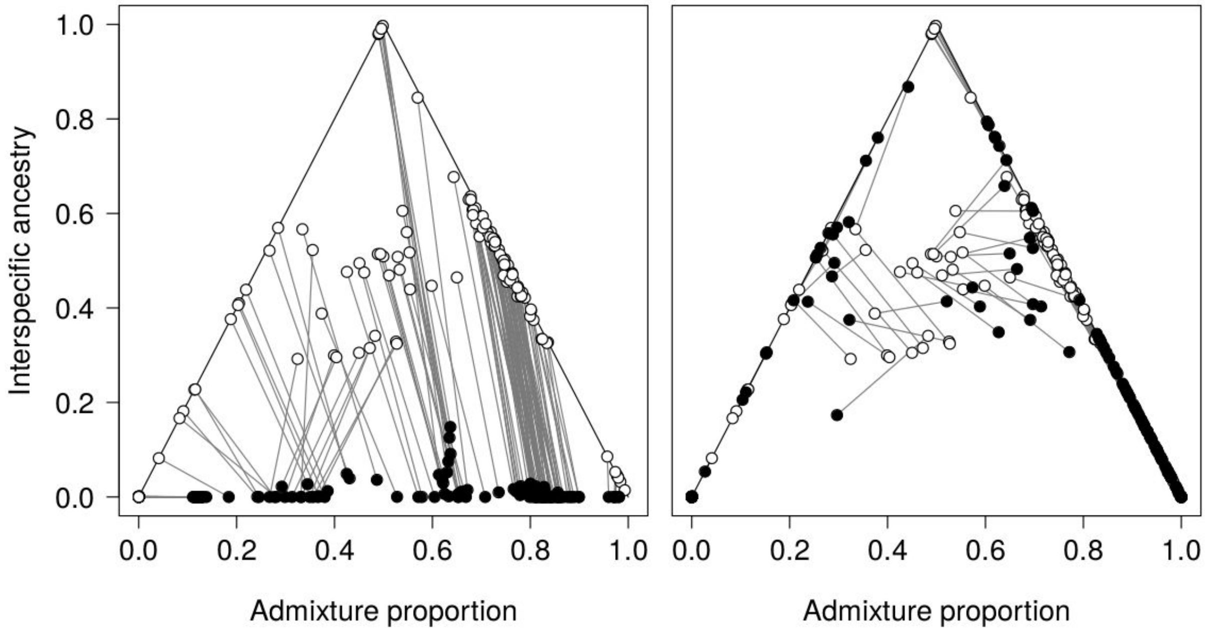

Fig. 3: *Comparison of genome-wide (*q*) and interspecific (*Q_{12}*) ancestry estimates of 136 individuals obtained from GBS data (open circles, Lindtke et al. 2014) and estimates from RAD-seq data (black circles) using either the raw (left panel) or corrected (right panel) genotype likelihoods. Grey lines connect values obtained for the same individual.*

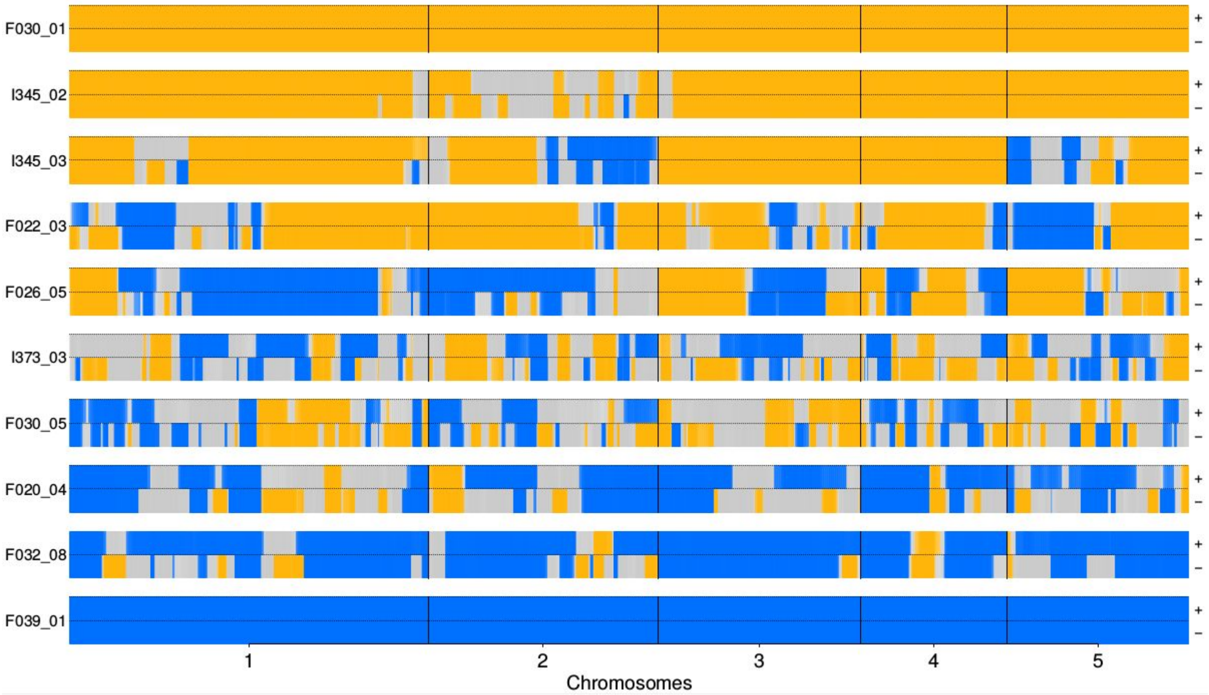

Fig. 4: *Comparison of local ancestry patterns on chromosomes 1-5 for (top to bottom) a putatively pure and P. tremula individual (F030_01), two putative backcrosses to P. tremula (I345_02 and I345_03), two putative hybrids of later generation (F022_03 and F026_05), two putative F1 hybrids (I373_03 and F030_05), two putative backcrosses to P. alba (F020_04 and F032_08) and a putatively pure P. alba individual (F039_01). For each individual, RASPberry results with (+) and without (-) correction are shown. Blue represents P. alba ancestry, orange P. tremula ancestry and grey heterospecific ancestry, with darker shades showing higher confidence in the ancestry estimates. To facilitate visualization, only sites with data are not shown.*