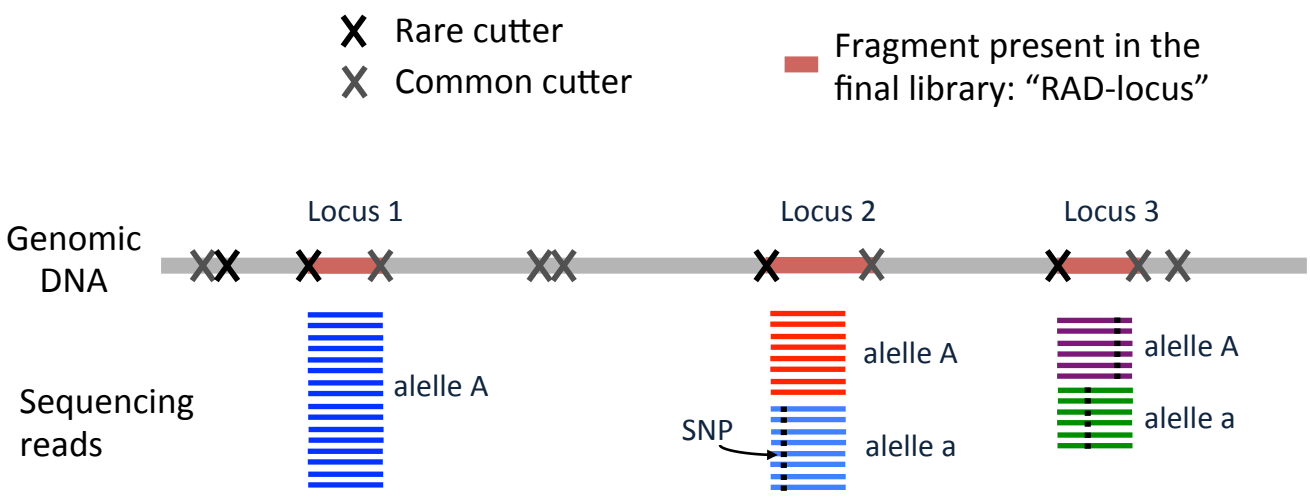
Supporting Information 1. Schematic diagram of RAD data genotyping and differences between replicates

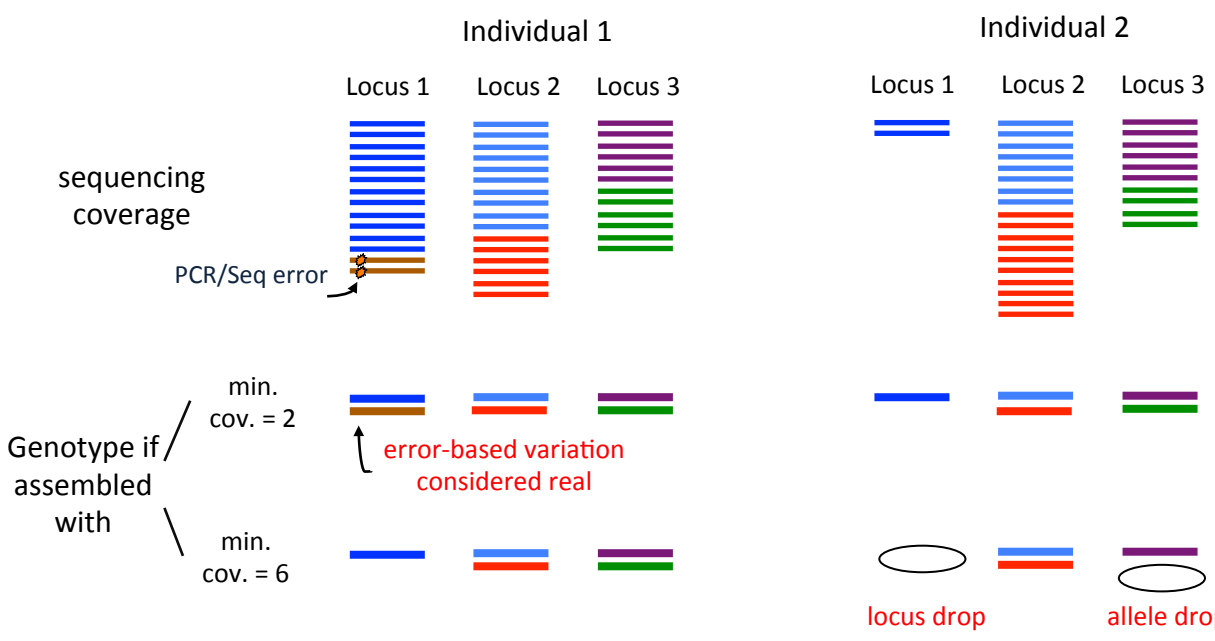
from Mastretta-Yanes *et al.* (2014). RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference

(I) RAD-loci, alleles, SNPs and coverage



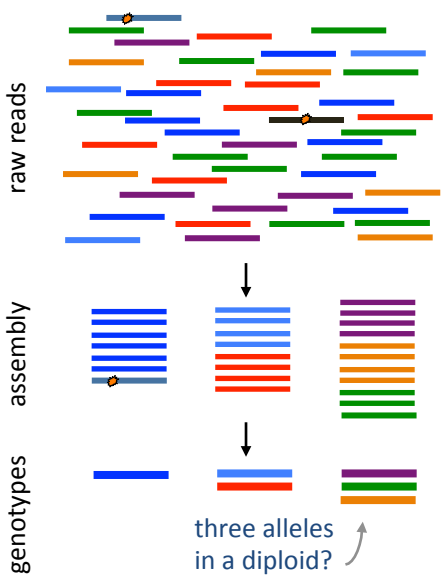
An example double digest RAD library: genomic DNA is digested with two restriction enzymes (a rare and a common cutter) and processed to create sequencing competent fragments. The RAD-loci present in the final library are the fragments kept after the size selection. A RAD-locus is thus a short DNA sequence. Each locus can have one or more alleles, which differ from each other by a small number of SNPs (black squares). Sequencing produces a number of reads per allele, which is referred to as coverage. The same principle applies to traditional RAD-seq libraries.

(II) The role of coverage



During assembly and genotyping, setting a threshold for minimal coverage (defined by researchers) allows to distinguish between PCR/sequencing error and real variation. If it is set too low it can lead to error-based variation being considered real. However, if it is set too high it can cause locus or allele dropout. Locus dropouts results in missing data, but allele dropout results in inferences of homozygosity, when the underlying state of the locus is heterozygous. See Table 1 for reasons that can lead to heterogeneous and low coverage, and for other sources of error.

(III) Loci, alleles, SNPs, error and *de novo* assembly



In the absence of a reference genome, loci are assembled by matching together similar sequences and considering them either different loci, or alleles from the same locus. This is done based on a given number of mismatches (which are defined by researchers).

The main challenges are that: 1) repetitive regions and paralogs can be erroneously assembled together, creating alleles that are not biologically meaningful; 2) PCR/sequencing can form error-based alleles; and 3) both 1 and 2 occur at the same time.

As a consequence, there can be genotypic differences between individuals that are not biologically true.

(IV) Differences between replicates and error rates

DNA replicates derived from the same sample should have the same genotype, and any differences can be considered error produced from any of several possible reasons (Table 1). The differences between replicates can be examined at the locus, allele and SNP levels. Consider 6 RAD-loci genotyped in 4 individuals, of which we have replicates for individual 1 and 2:

	Individual 1		Individual 2		Individual 3	Individual 4
	Replicate I	Replicate II	Replicate I	Replicate II		
Locus 1		**AA**		**aa**	**Aa**	**AA**
Locus 2	**Aa**	**Aa**	**aa**	**Aa**		**AA**
Locus 3	**AA**		**AA**	**AA**	**AA**	**AA**
Locus 4	**aa**	**aa**			**aa**	**aa**
Locus 5			**Ab**	**AA**	**aa**	
Locus 6		**Aa**	**Aa**	**Aa**	**Aa**	**AA**

If we look at the distribution of missing loci **per replicate pair**, we can see that for individual 1 loci 1 & 6 are missing in replicate I; locus 3 is missing for replicate II and locus 5 is missing in both replicates. This means that for the replicate pair of individual 1 the **number of missing loci** is 4 and that the **proportion of loci missing** relative to all loci in the population is 4/6. Note that of the four missing loci, only locus 5 was lost in both replicates (and therefore does not result in a genotypic difference between them), whilst loci 1, 3 and 6 were lost in one replicate or the other, but not in both. Therefore, the proportion of **missing loci where a locus was lost only in one of the replicates** is 3/4. If we estimate this same proportion but against the total number of loci found for all individuals we have 3/6, which is **the locus error rate**.

There is error at the allele level if the alleles of a locus are different for a replicate pair. This can be caused by allele dropout due to low coverage or assembly error. For example, for individual 2, locus 2 is homozygous for replicate I and heterozygous for replicate II, and in locus 5 there is a different allele not present in both replicates. If we count mismatches like these and divide by the number of loci present in both replicates we have that the **allele error rate** for individual 2 **is 2/4**. Since alleles of a RAD-locus can differ by more than one SNP (see diagram I), the same principle can be applied to estimate the **SNP error rate**.

(IV) Replicates can aid *de novo* assembly

Because DNA replicates derived from the same sample should have the same genotype, one can evaluate which parameter values of the assembly pipeline optimize for a high number of loci with less differences between replicate pairs.