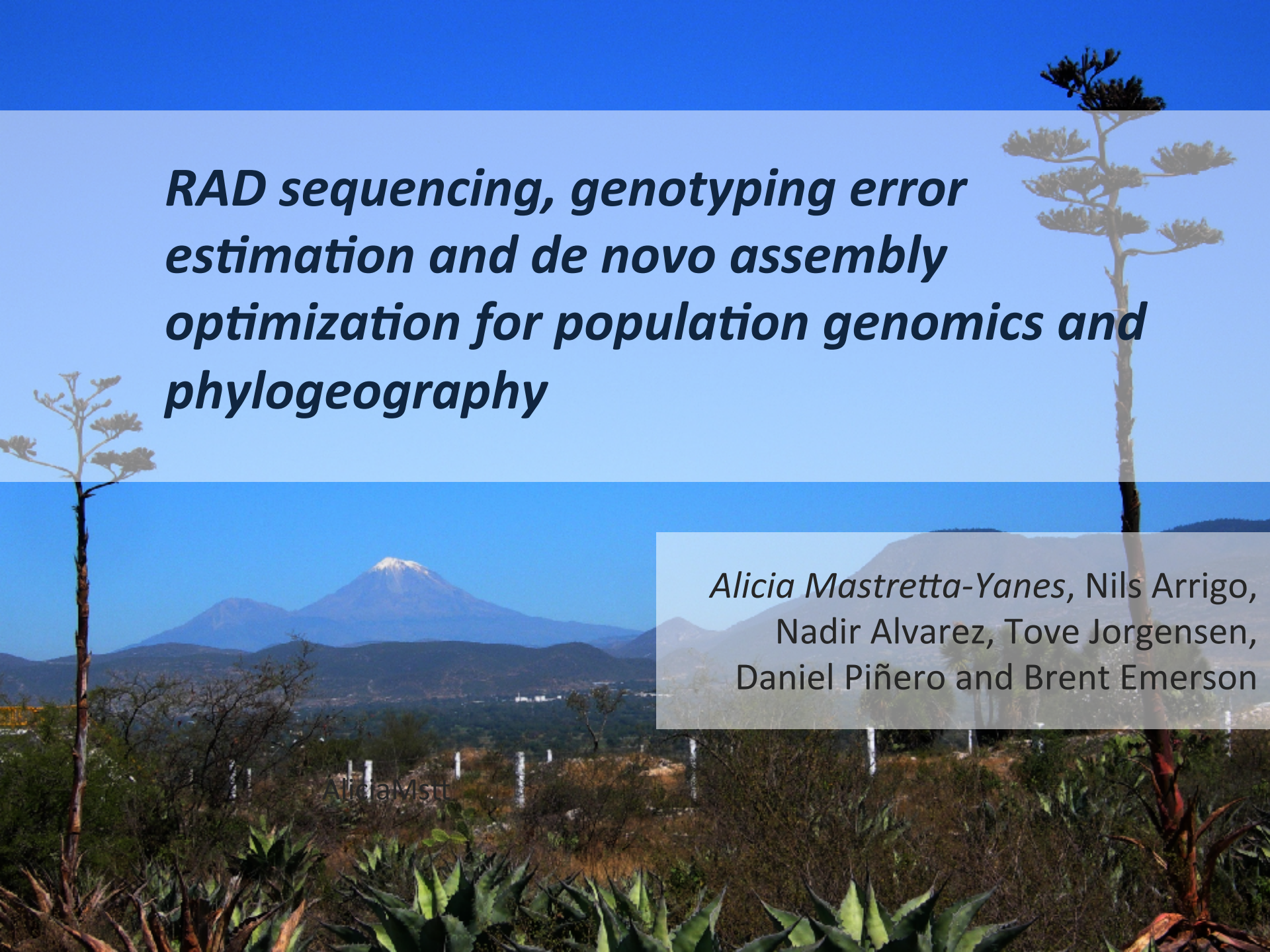
RAD sequencing, genotyping error estimation and de novo assembly optimization for population genomics and phylogeography

Alicia Mastretta-Yanes, Nils Arrigo, Nadir Alvarez, Tove Jorgensen, Daniel Piñero and Brent Emerson



Restriction site-associated DNA sequencing

Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers

Nathan A. Baird, Paul D. Etter, Tressa S. Atwood, Mark C. Currey, Anthony L. Shiver, Zachary A. Lewis, Eric U. Selker, William A. Cresko, Eric A. Johnson

Published: October 13, 2008 • DOI: 10.1371/journal.pone.0003376

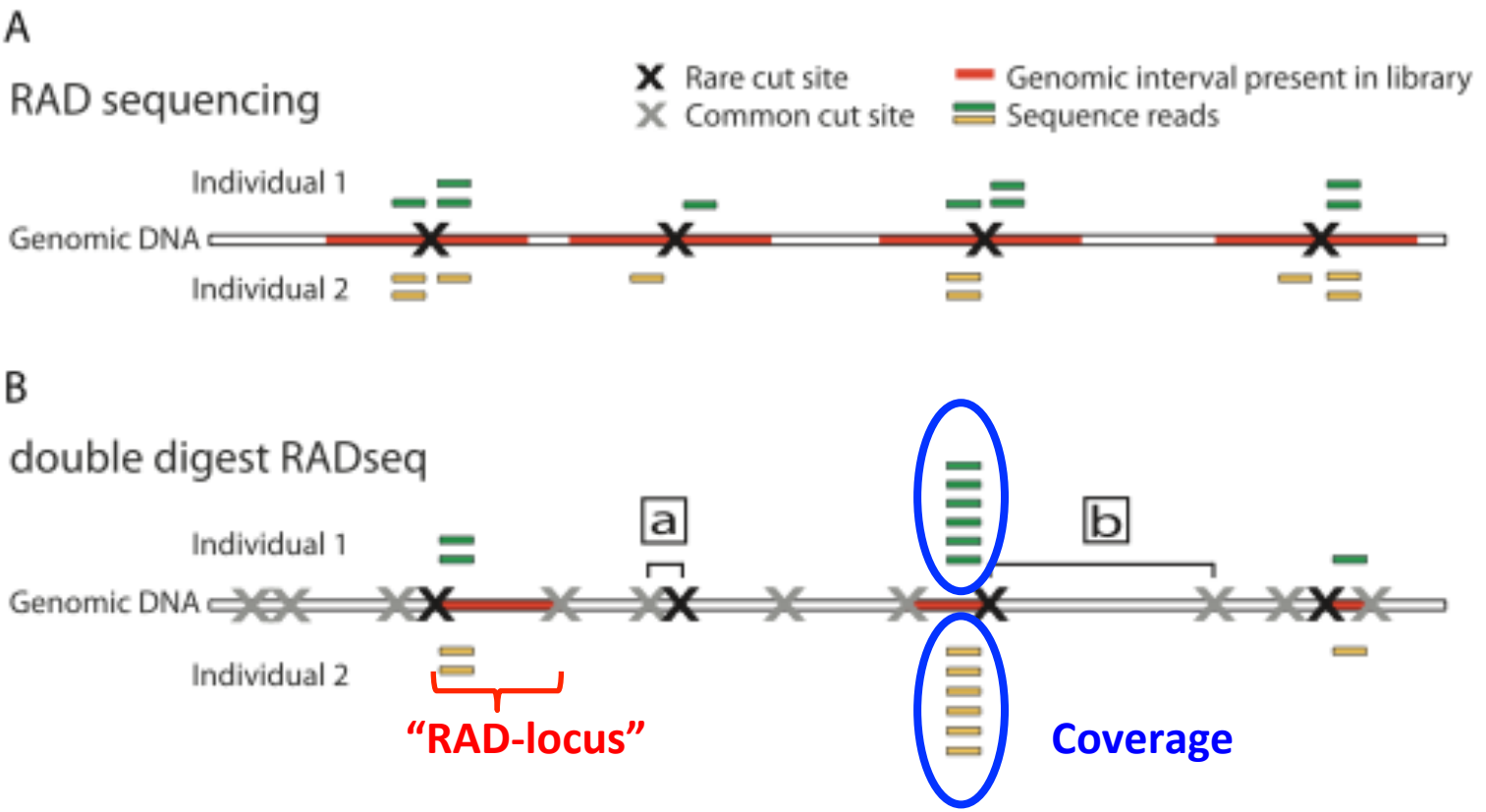
Proportion of the genome sequenced

Sanger sequencing

Whole genome sequencing

Reduced representation genome sequencing

non model species revolution



Peterson *et al.* 2012. PLoS ONE

Implications of genotyping errors for population genetics

- Artificial excess of homozygotes
- False departure from Hardy–Weinberg
- Overestimation of inbreeding
- Unreliable inferences about population structure
- Wrong estimations of nucleotide diversity
- Incorrect inferences at the genome-wide level

overlooked in RADseq

RADseq sources of error

Technical & human

Species genome

Wet lab

PCR stochastic sampling events

Bioinformatics

- Assembly algorithm
- Mismatches parameters

Next Generation Sequencing

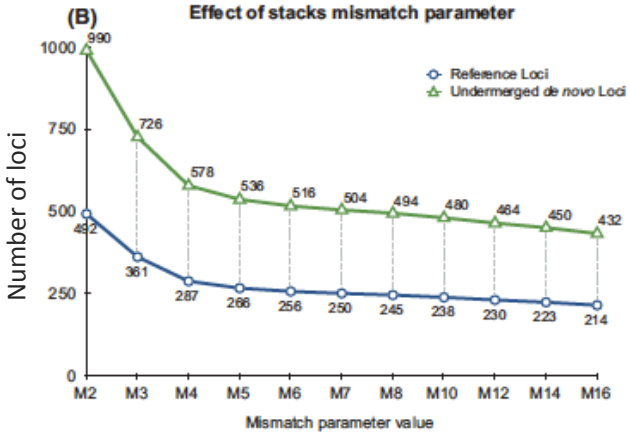
Stacks (Catchen *et al.* 2011, 2013)

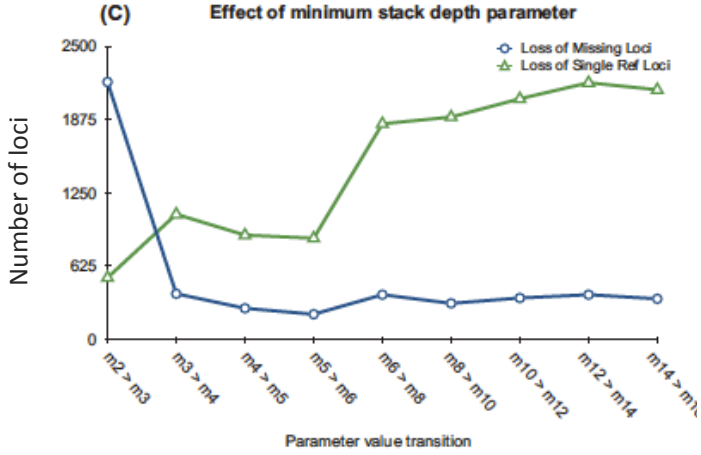
-m minimum number of identical raw reads required to form a stack

-M mismatches allowed between stacks when processing an individual

-max_locus_stacks maximum stacks allowed per locus

-n number of mismatches allowed between loci when building the catalog





Catchen et al 2013 Mol. Ecol.

Optimal parameter values depend on:

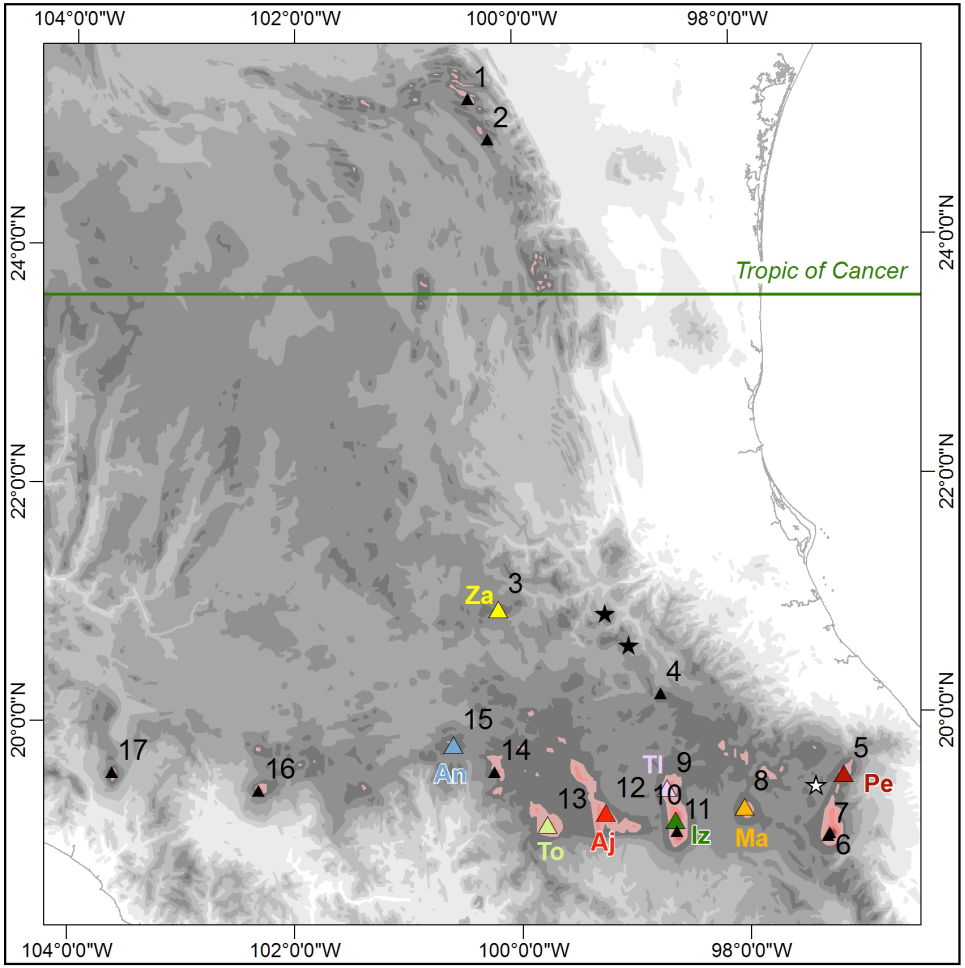
- Polymorphism of the genome
- Sequencing error
- Depth of sequencing

Explore values for each data set

No reference genome?

- DNA replicates
  - expectation of genome identity
- *De novo* assembly algorithms

Estimate error rates



Berberis alpina

Genome: ~1.5 Gb, diploid

ddRAD (EcoRI and MseI)
3 sequencing lanes
27 samples + 5 replicates
Populations and taxonomy represented

Exploratory analysis of *Stacks* key parameters using replicates

Assembly parameters

- *-m*: 2 to 15
- *-M*: 2 to 10
- *-n*: 0 to 5
- *-max_locus_stacks*: 2 to 6
- SNP calling model: free vs upper bound of 0.5, 0.25, 0.15, 0.1, 0.05 and 0.0056

- 11 replicate pairs

Evaluate

- Number of loci and SNPs
- Error rates
- Distribution of missing data

-m min. raw reads to form a stack ("minimal coverage")
-M mismatches between stacks when processing an individual

-max_locus maximum stacks allowed per locus
-n mismatches between loci when building the catalog

Error rates

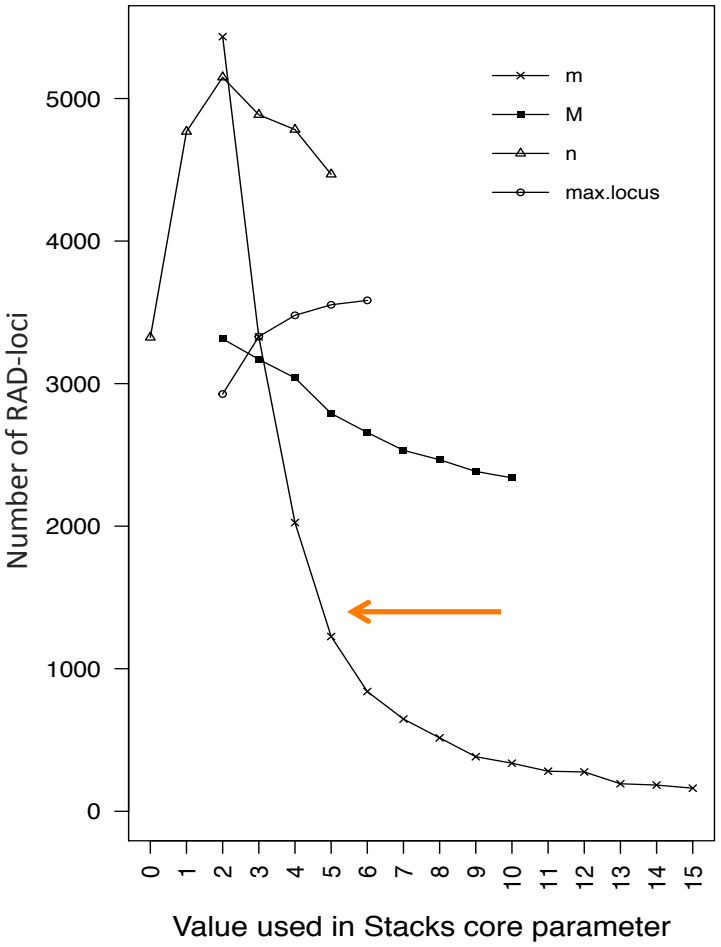
Locus error rate = $\dfrac{\text{Number of RAD-loci present in only one of the samples of a replicate pair}}{\text{Total number of RAD-loci}}$

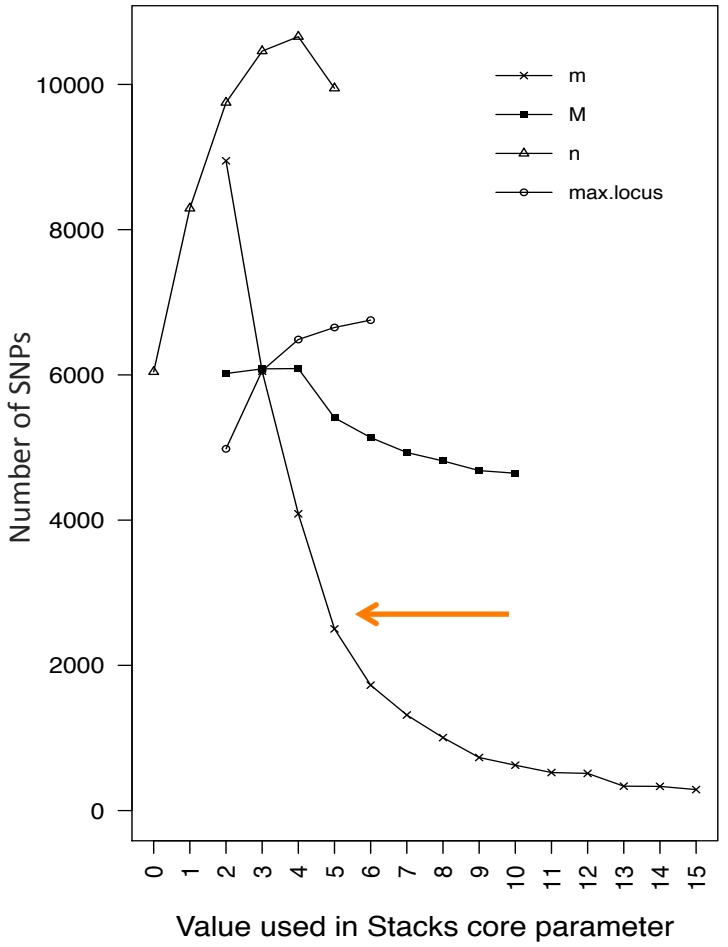
Allele error rate = $\dfrac{\text{Number of allele mismatches between a replicate pair}}{\text{Number of alleles for the replicate pair}}$

SNP error rate = $\dfrac{\text{SNP mismatches between a replicate pair}}{\text{Number of SNPs for the replicate pair}}$

Results

Output loci and SNPs





-m min. raw reads to form a stack (“minimal coverage”)

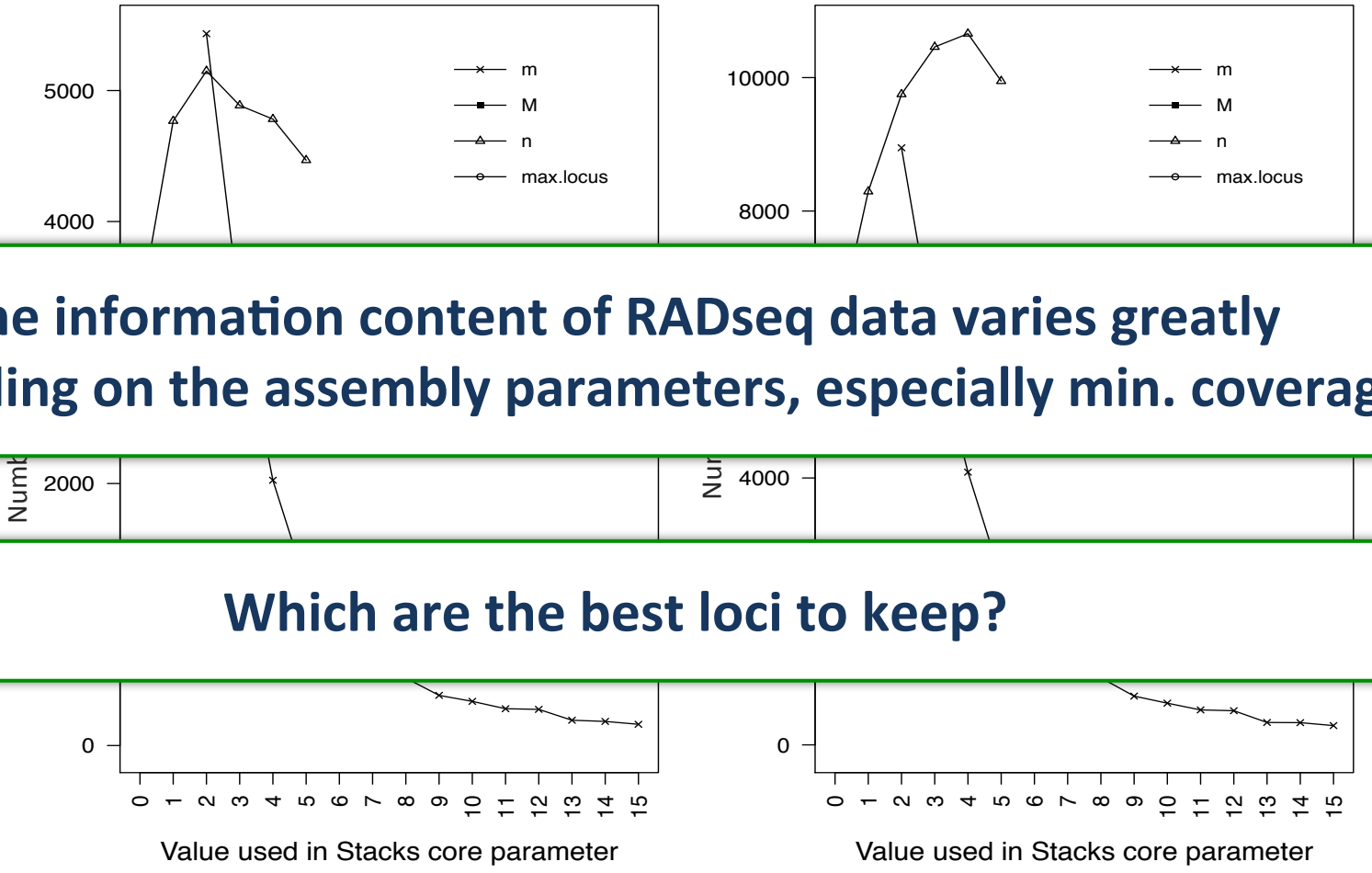
-M mismatches between stacks when processing an individual

-max_locus maximum stacks allowed per locus

-n mismatches between loci when building the catalog

Results

Output loci and SNPs



The information content of RADseq data varies greatly depending on the assembly parameters, especially min. coverage

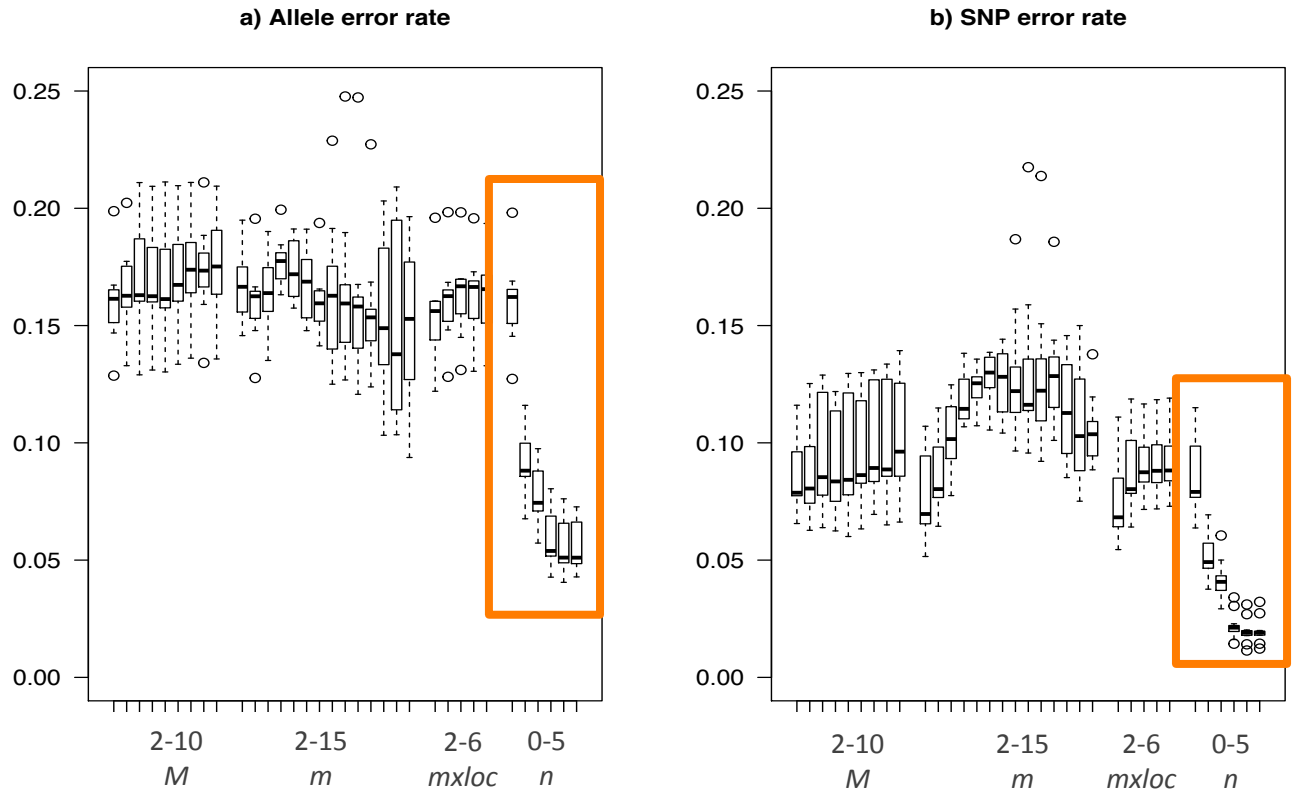
Which are the best loci to keep?

-m min. raw reads to form a stack ("minimal coverage")
-M mismatches between stacks when processing an individual

-max_locus maximum stacks allowed per locus
-n mismatches between loci when building the catalog

Results

Allele and SNP error rates



!!!

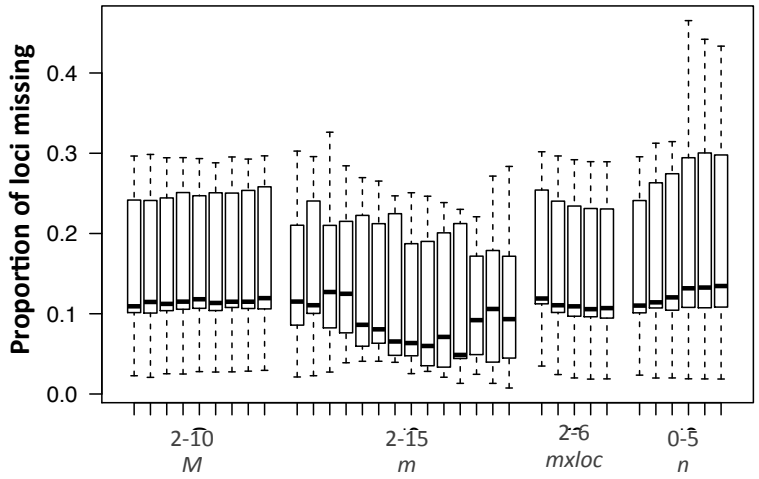
- **RADseq data has error, and it can be very high**
- **You won't know it without replicates**
- **Error can be decreased by tuning the assembly parameter values**

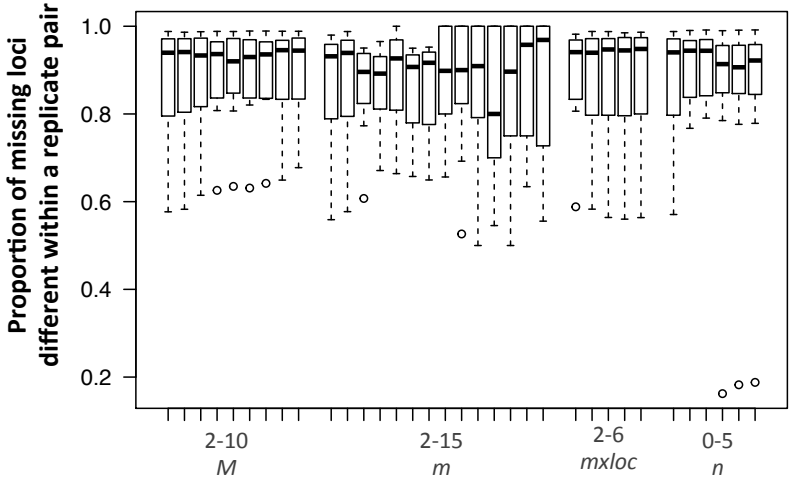
-m min. raw reads to form a stack ("minimal coverage")
-M mismatches between stacks when processing an individual

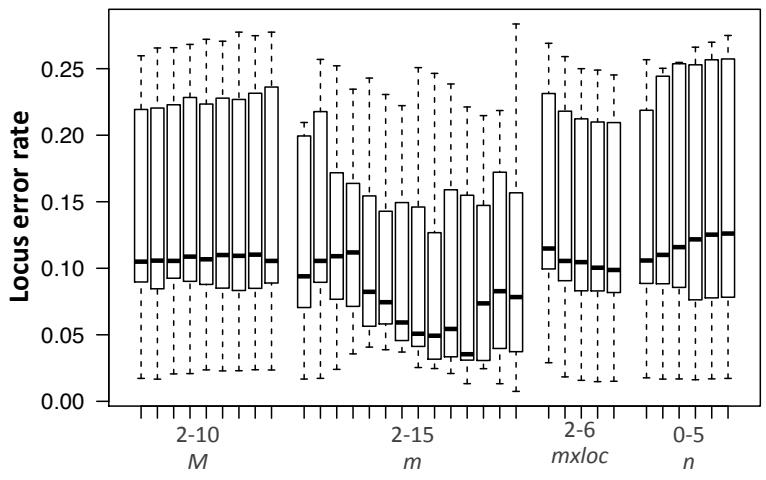
-mxloc maximum stacks allowed per locus
-n mismatches between loci when building the catalog

Results

Distribution of missing data





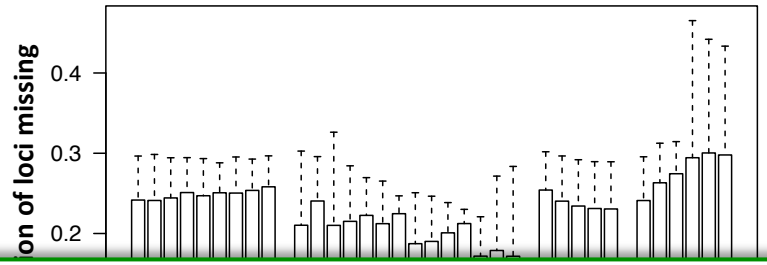


-m min. raw reads to form a stack ("minimal coverage")
-M mismatches between stacks when processing an individual

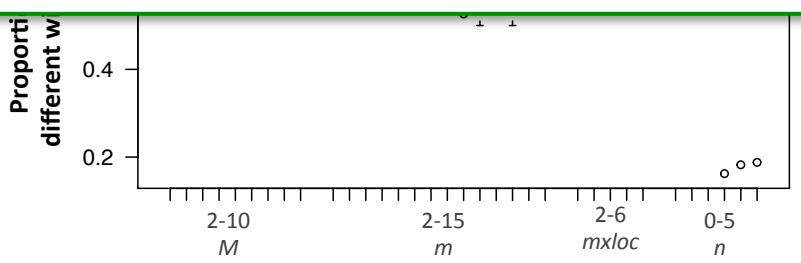
-mxloc maximum stacks allowed per locus
-n mismatches between loci when building the catalog

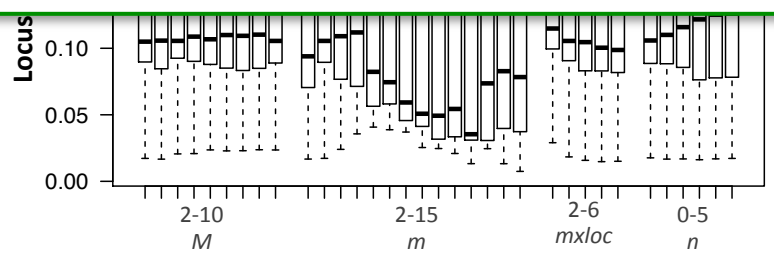
Results

Distribution of missing data



- **RADseq produces missing loci that are randomly distributed and little affected by assembly parameters**
- **Missing loci could be handled as missing data, but erroneous alleles and SNPs have negative consequences on biological inferences**





-m min. raw reads to form a stack ("minimal coverage")
-M mismatches between stacks when processing an individual

-mxloc maximum stacks allowed per locus
-n mismatches between loci when building the catalog

Results

"Best" parameter values

Highest number loci & SNP
Lowest allele & SNP error rate

m= 3, M=2, *N*=4, *n*=3, *max_locus_stacks*=3 and a SNP calling model with an upper bound of 0.05

Specific for this data set

Does it matter?

Effect on output information content and on detection of genetic structuring

Profiles

- **default**
- **optimal** (best from replicates analysis and *m*=3)
- **near optimal** (best from replicates analysis and *m*=4)
- **high coverage** (best from replicates analysis and *m*=10)

Samples

- *B. alpina* populations (75)
- Closest outgroup (*B. trifolia*) (3)
- Replicate pairs (10)

Results

Information content

	optimal	*near optimal*	*high coverage*	*default*
Number of **RAD-loci**	6,292	2,449	292	4,554
Number **SNPs**	11,057	4,353	502	7,736
Mean **coverage**	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)

Error rates

	optimal	*near optimal*	*high coverage*	*default*
Mean **locus error rate**	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean **allele error rate**	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean **SNP error rate**	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)

Results

Information content

	optimal	*near optimal*	*high coverage*	*default*
Number of **RAD-loci**	6,292	2,449	292	4,554

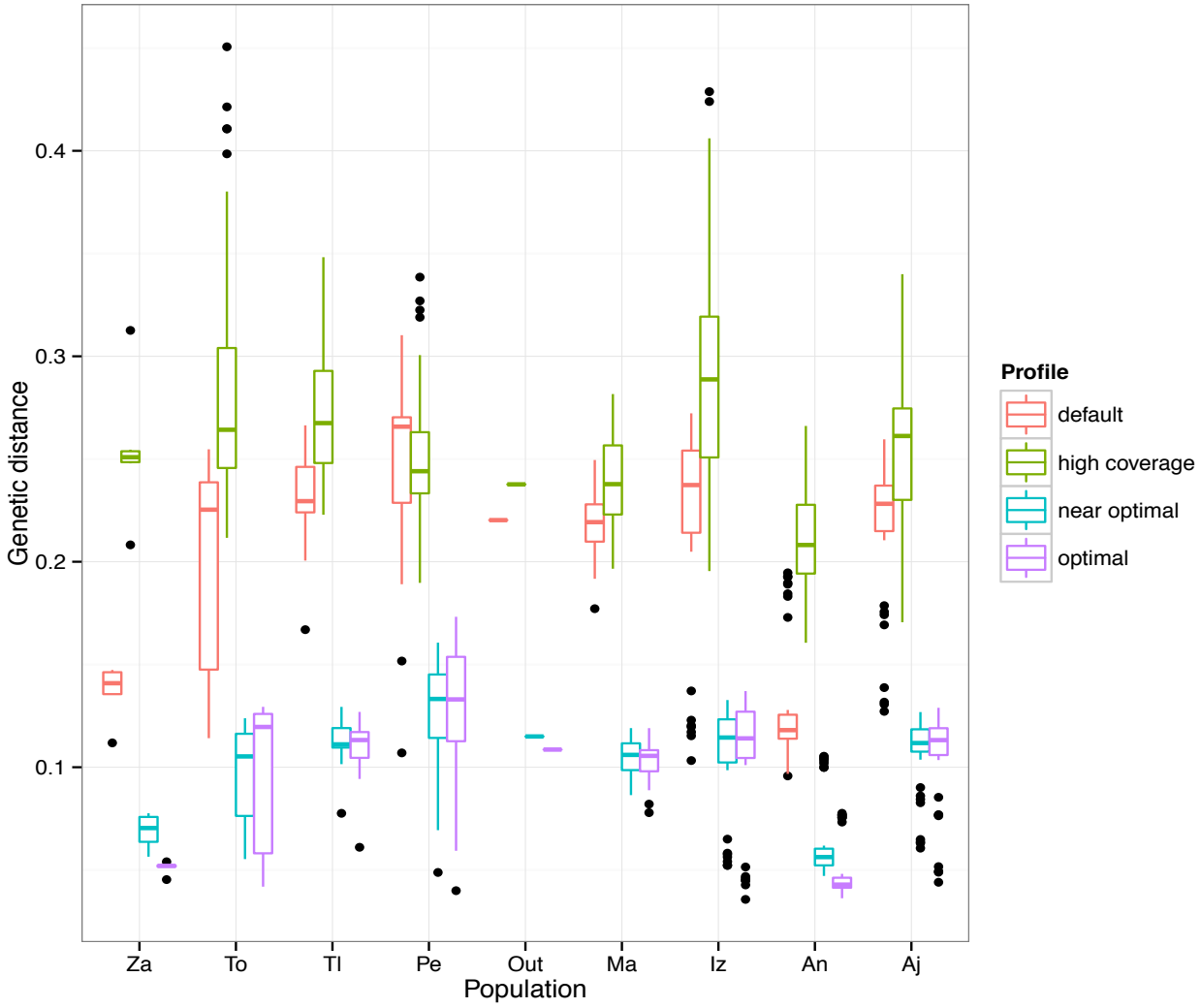
Using replicates to optimize *de novo* assembly provides the highest information content at the lowest allele and SNP error rates

Does optimizing for small error rates estimated from replicates improve the data quality of the rest of the samples?

Mean **locus error rate**	0.1738 (SD 0.103)	0.1657 (SD 0.100)	0.0882 (SD 0.088)	0.1590 (SD 0.094)
Mean **allele error rate**	0.0592 (SD 0.013)	0.0599 (SD 0.010)	0.0879 (SD 0.023)	0.0841 (SD 0.017)
Mean **SNP error rate**	0.0243 (SD 0.006)	0.0321 (SD 0.006)	0.0578 (SD 0.019)	0.0423 (SD 0.010)

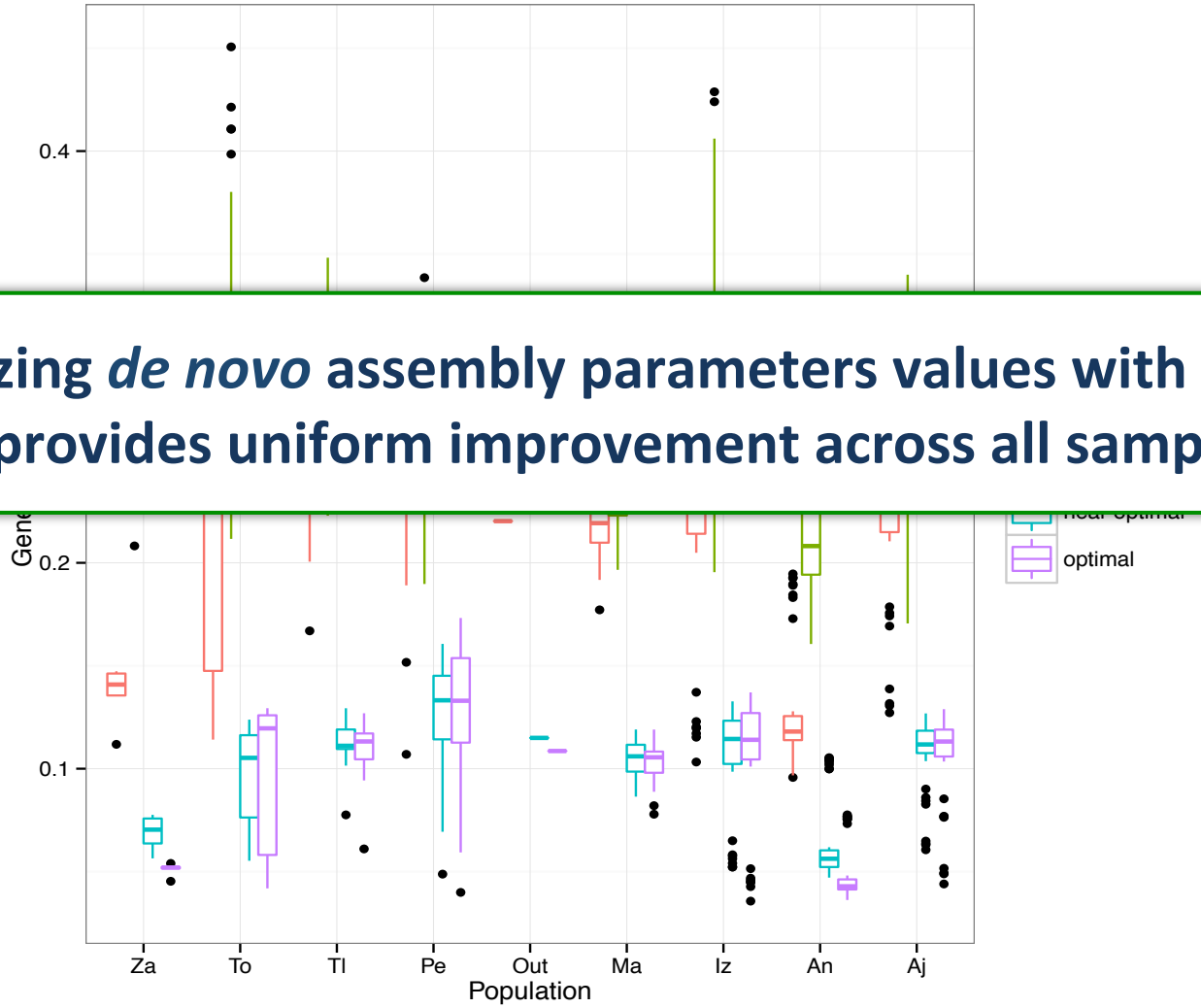
Results

Genetic distance between individuals of the same locality



Results

Genetic distance between individuals of the same locality



Optimizing *de novo* assembly parameters values with replicates provides uniform improvement across all samples

Results

Detection of structuring of genetic variation

	optimal	*near optimal*	*high coverage*	*default*
Variation explained by first two axes of PCoA*	80%	82%	47%	57%
Mean of F_{ST} pairwise matrix*	0.19	0.15	0.03	0.07

	optimal	*near optimal*	*high coverage*	*default*
Number of **RAD-loci**	6,292	2,449	292	4,554
Number **SNPs**	11,057	4,353	502	7,736
Mean **coverage**	10.32 (SD 4.16)	15.30 (SD 5.9)	58.92 (SD 21.9)	11.50 (SD 4.65)

Conclusions and recommendations

- Locus, allele and SNP error rates can be quantified with DNA replicates
- Information content and error rates can be simultaneously optimized
- Biological inferences should be discussed in the context of error rates
- Error rates should be reported
- Full power of RADseq