

Short INDELS: genetic markers
for adaptive divergence

Original aspects of the short INDELs project

- Divergent natural selection vs neutral processes
- Species with high diversity
- Systems with imperfect genomes can still contain useful functional information

INDEL-SNP comparisons

1. Outlier sharing
2. Clustering of (different types) markers
3. Derived allele frequencies (in progress and for now simply minor)
4. Displacement of cline centres

1. Outlier sharing

Total number of Anja's SNP: 55106

Proportion of SNP with significant clines.

CZA left: 0.5122128

CZA right: 0.4238377

CZB left: 0.3201829

CZB right: 0.4114071

CZD left: 0.4393351

CZD right: 0.4732697

Proportions of SNP outliers that are shared.

CZA left and right: 0.707804

CZB left and right: 0.5680581

CZD left and right: 0.6569873

CZA and CZB: 0.4650635

CZA and CZD: 0.508167

CZB and CZD: 0.5426497

Number of SNP outliers found in 1 hybrid zone(s): 602

Number of SNP outliers found in 2 hybrid zone(s): 258

Number of SNP outliers found in 3 hybrid zone(s): 137

Number of SNP outliers found in 4 hybrid zone(s): 117

Number of SNP outliers found in 5 hybrid zone(s): 95

Number of SNP outliers found in 6 hybrid zone(s): 139

Prop. of SNP outliers in inversions found in 1 zone(s): 0.648

Prop. of SNP outliers in inversions found in 2 zone(s): 0.698

Prop. of SNP outliers in inversions found in 3 zone(s): 0.912

Prop. of SNP outliers in inversions found in 4 zone(s): 0.923

Prop. of SNP outliers in inversions found in 5 zone(s): 0.979

Prop. of SNP outliers in inversions found in 6 zone(s): 1

Total number of SNP: 11225

Proportion of SNP with significant clines.

0.5317595

0.4457016

0.3277506

0.4244989

0.4473942

0.4823163

Proportions of SNP outliers that are shared.

0.6160714

0.5178571

0.6339286

0.359375

0.4107143

0.484375

142

66

29

27

25

13

0.556

0.636

0.862

0.889

0.92

1

Total number of INDEL: 1752

Proportion of INDEL with significant clines.

0.5296804

0.4549087

0.3413242

0.4092466

0.4737443

0.4834475

Proportions of INDEL outliers that are shared.

0.7058824

0.4705882

0.6470588

0.3529412

0.4117647

0.4411765

24

7

7

5

1

3

0.625

0.57

0.86

1

1

1

Except this difference in the total number of SNPs, the proportions look quite similar but I have not run any statistical tests

Given the difference between SAMtools and GATK in the total number of SNPs, I have run some diagnostics

- Contigs after coverage filter (in progress)

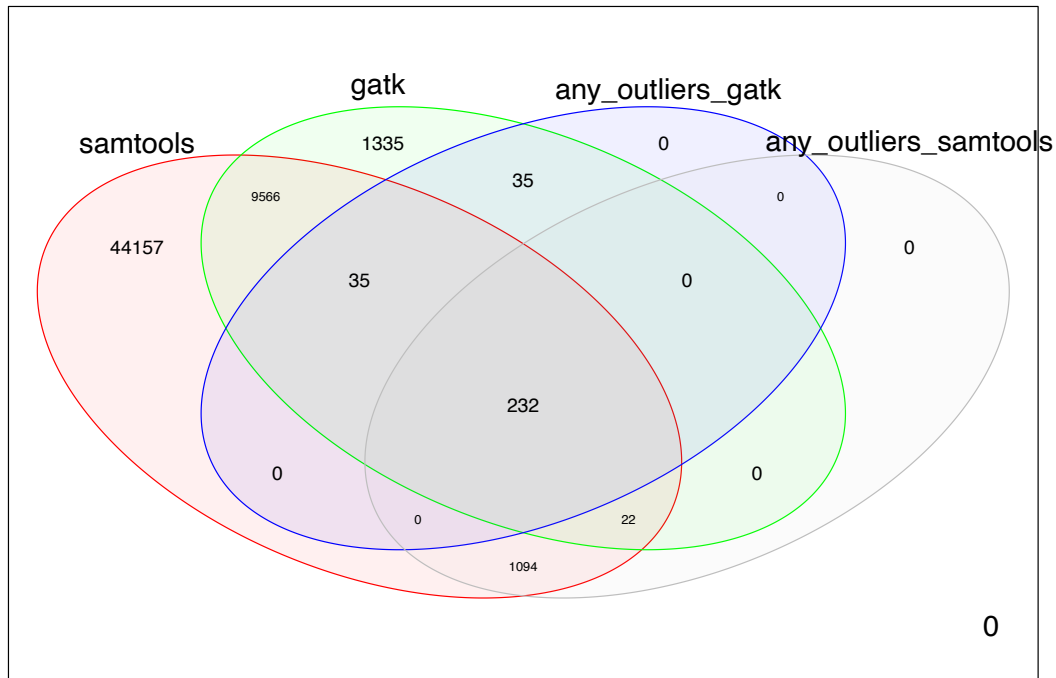
Given the difference between SAMtools and GATK in the total number of SNPs, I have run some diagnostics

- SNPs after all the filters (in progress)

Given the difference between SAMtools and GATK in the total number of SNPs, I have run some diagnostics

- SNPs after all the filters and cline/outlier analysis (i.e., clinal and non-clinal SNPs)

any_outliers = outlier in at least one CZ.



all_outliers = outlier shared by all six CZs (two per islands).

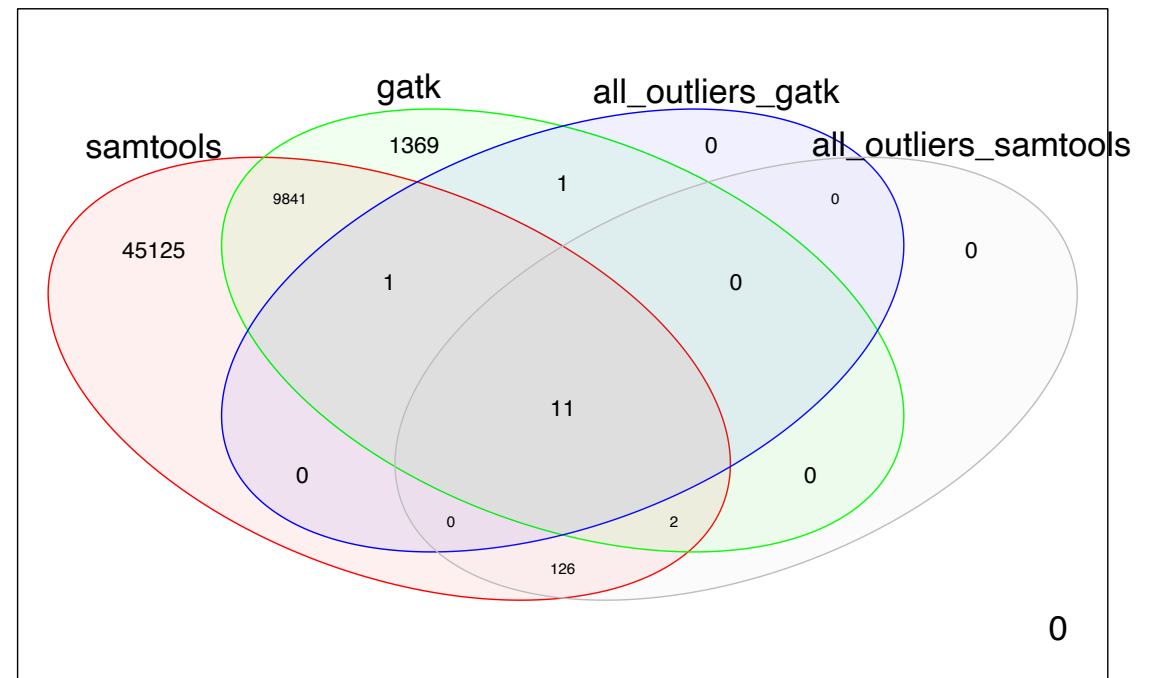
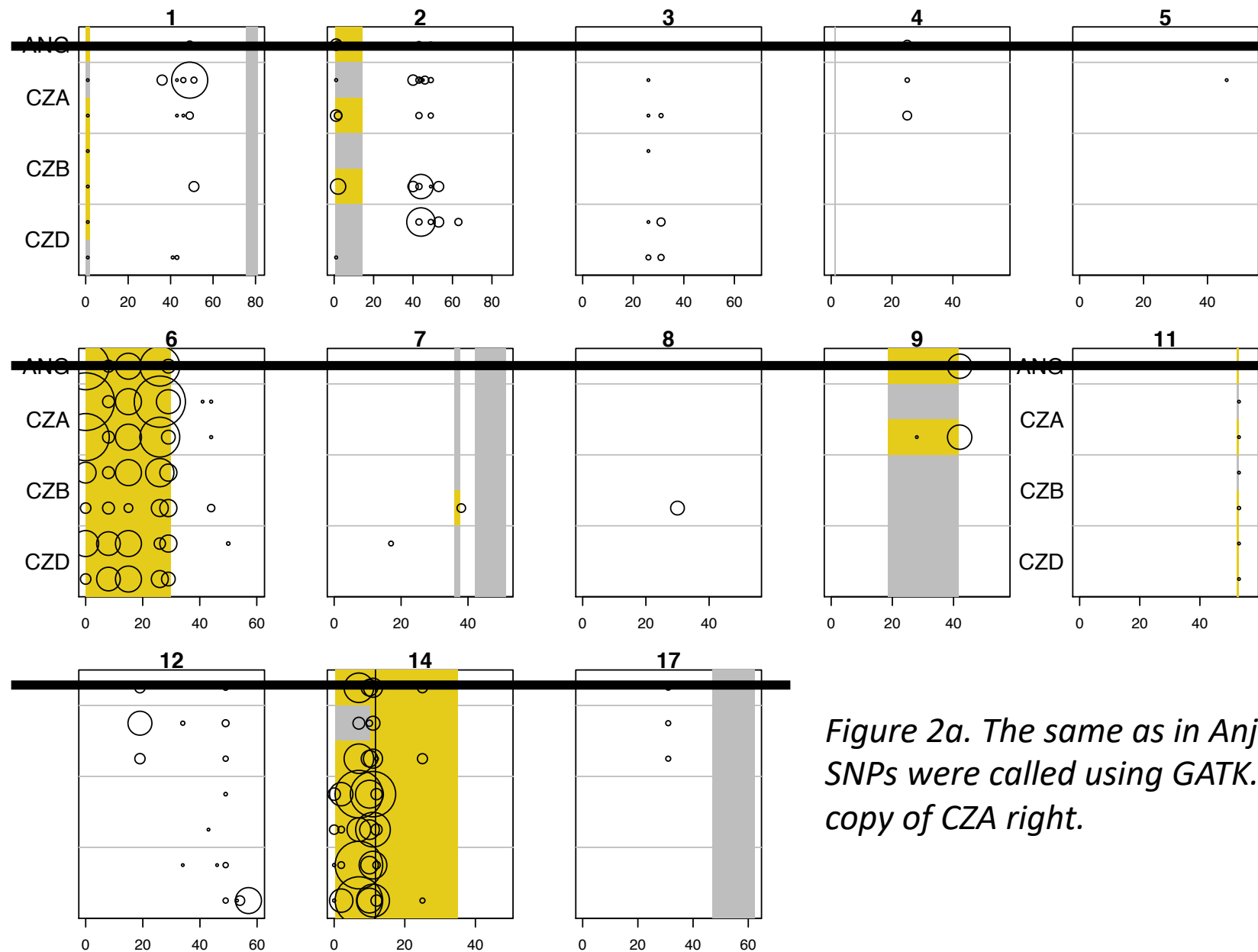


Figure 1. Venn diagrams of the number of SNPs after filtering and cline analysis. Left: SAMtools and GATK calls are intersected with the respective outliers that were at least present in one hybrid zone. Right: SAMtools and GATK calls are intersected with the respective outliers that were present in all six hybrid zones (two per islands).

GATK SNP call



Except this difference in the total number of SNPs, the map positions look quite similar but I have not run any statistics to test for significance.

Figure 2a. The same as in Anja's paper with the exception that the SNPs were called using GATK. ANG is crossed out because it was a copy of CZA right.

GATK INDEL call

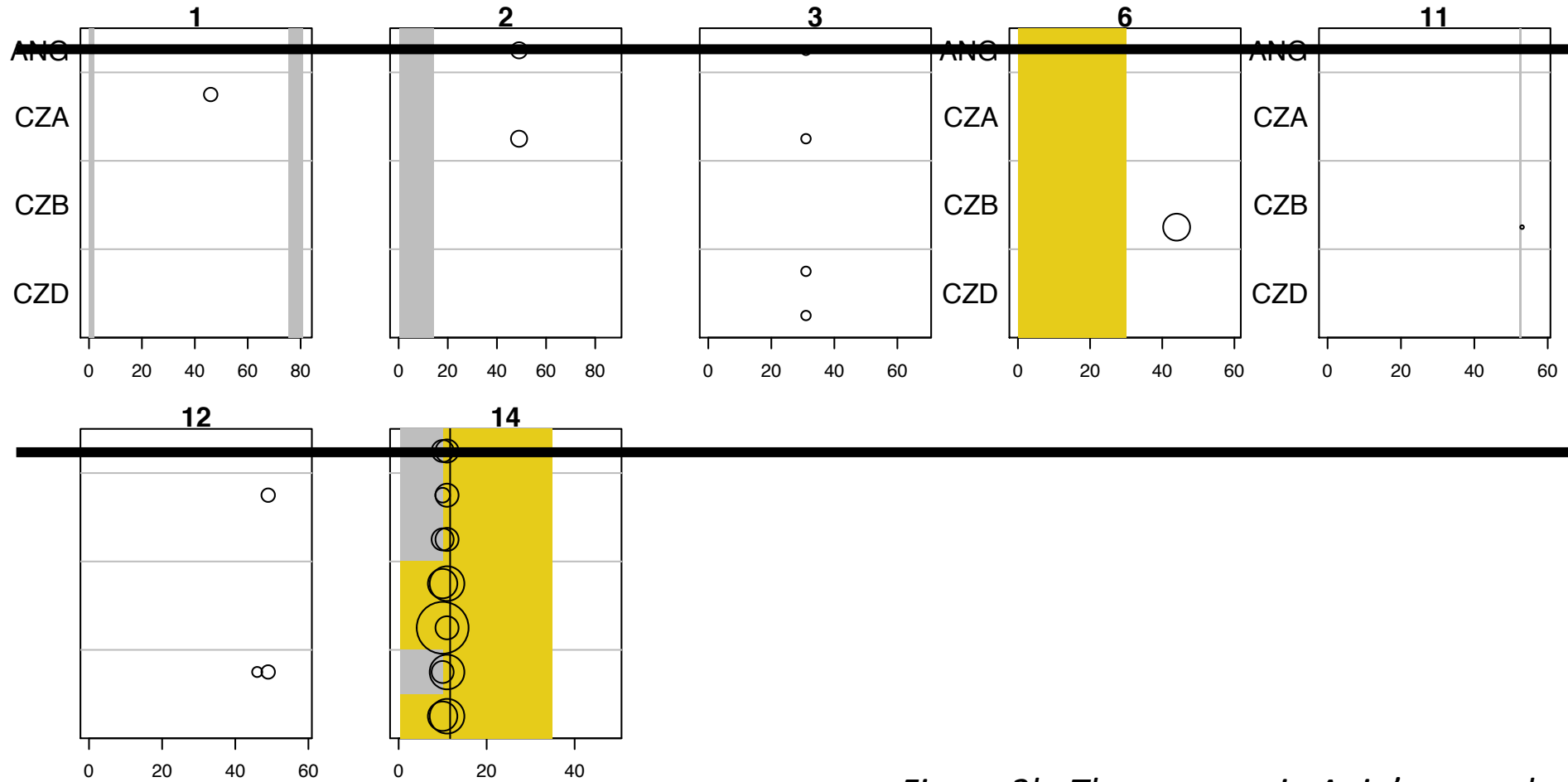


Figure 2b. The same as in Anja's paper but with INDELs. ANG is crossed out because it was a copy of CZA right.

2. Clustering of (different types) markers

- INDELs and SNPs after filtering and cline analysis.
- All six hybrid zones combined.

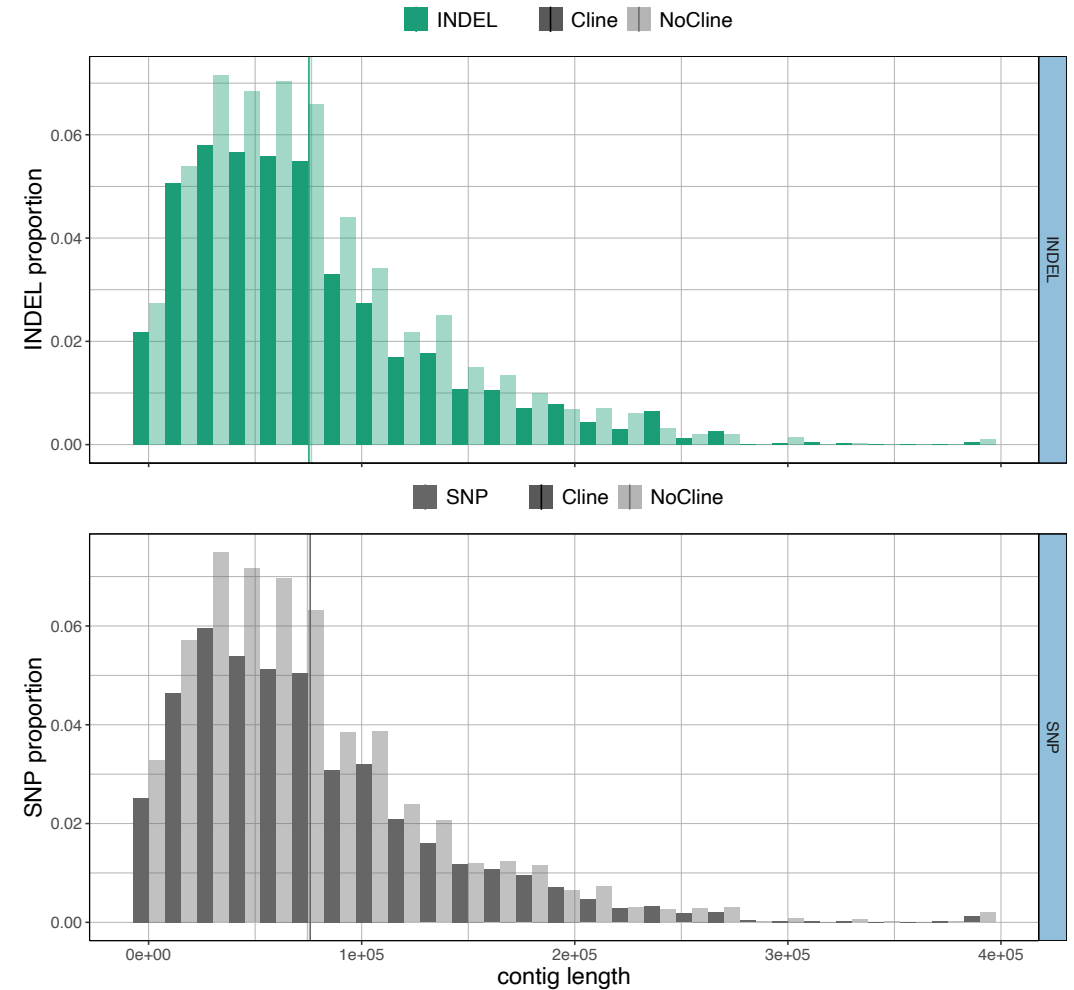
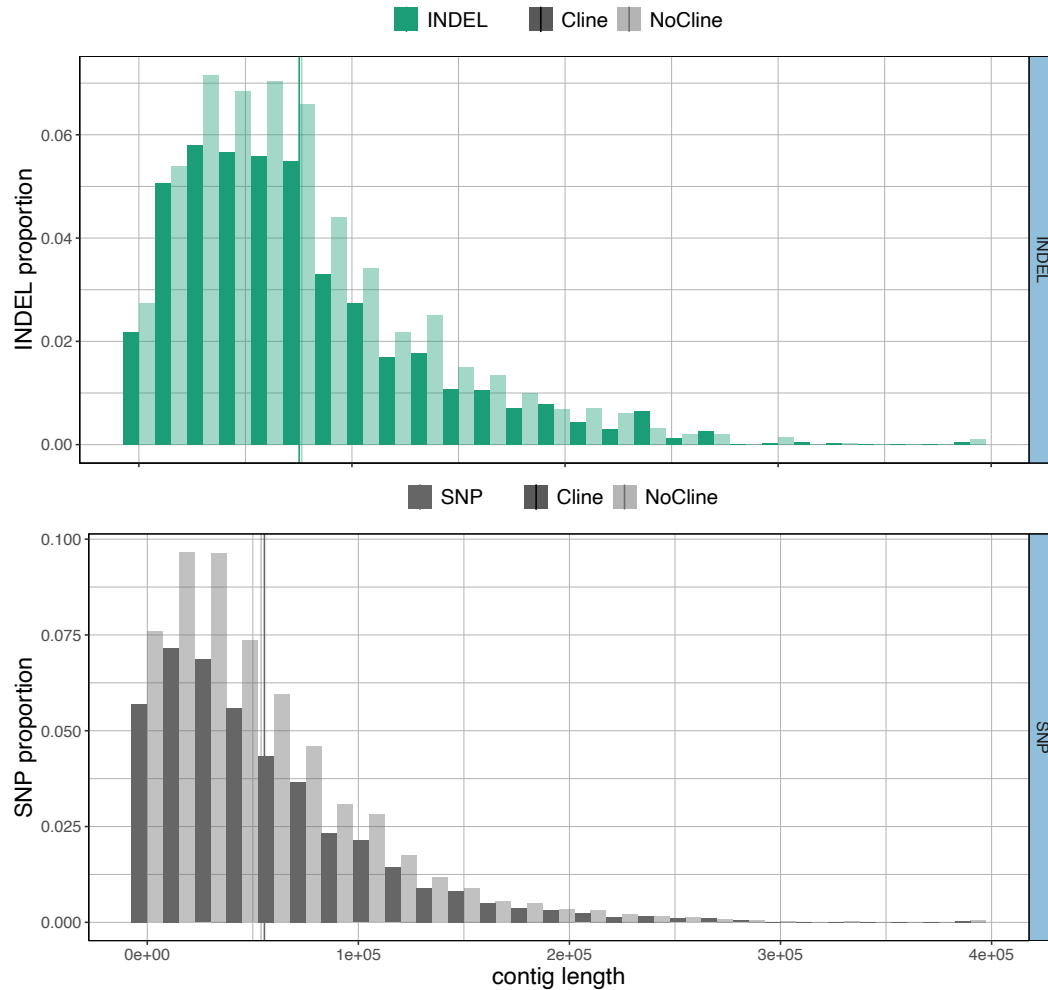


Figure 3a. Marker proportions over contig length. Proportion = $\text{count} / \sum \text{count per marker type}$ and bin width = 15000 base pairs. Clinal variants are dark coloured and non-clinal variants are light coloured. Left: SNP call using SAMtools and INDEL call using GATK. Right: both INDELs and SNPs were called with GATK.

2. Clustering of (different types) markers

- All INDELs = clinal + non-clinal

$$n/N = \frac{n_i \text{ marker}}{N \text{ marker}} \quad \begin{array}{l} n_i = \text{number of INDELs in contig } i \\ N = \text{Total number of INDELs} \end{array}$$

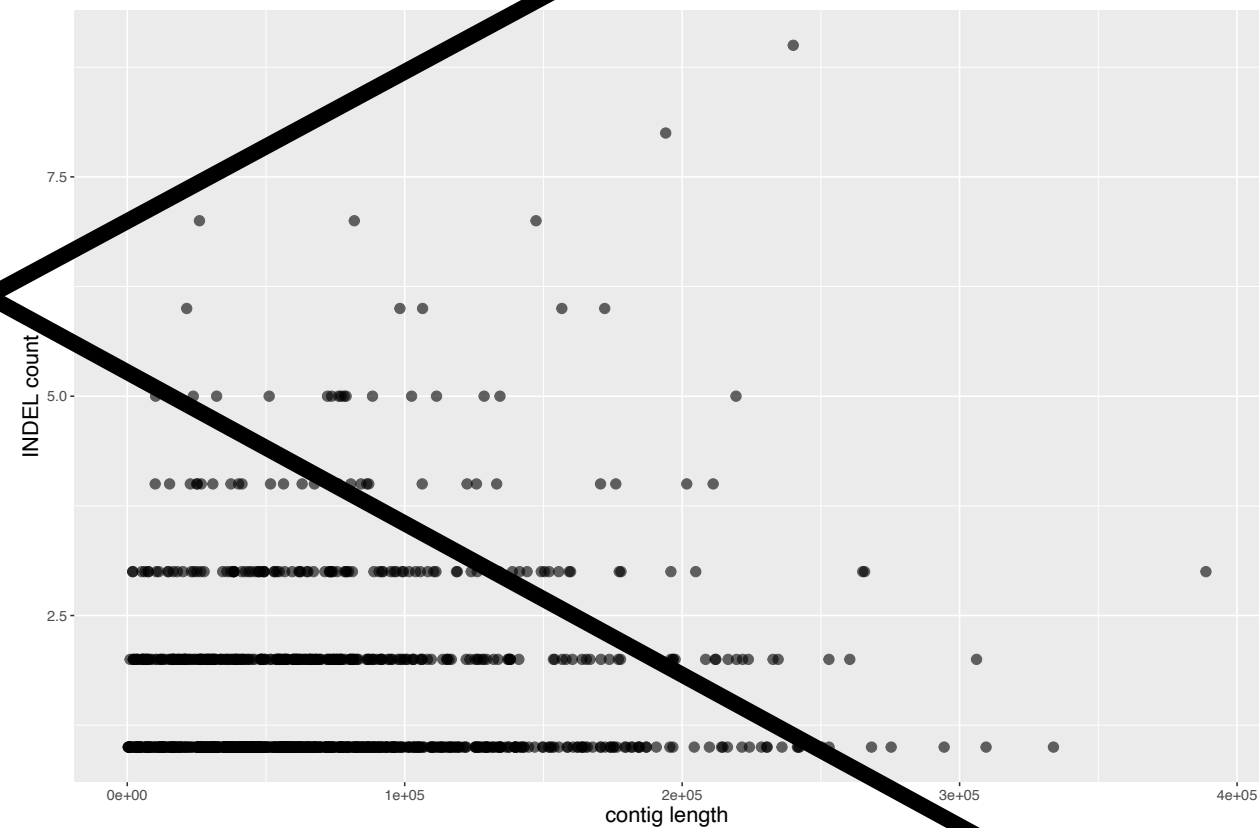
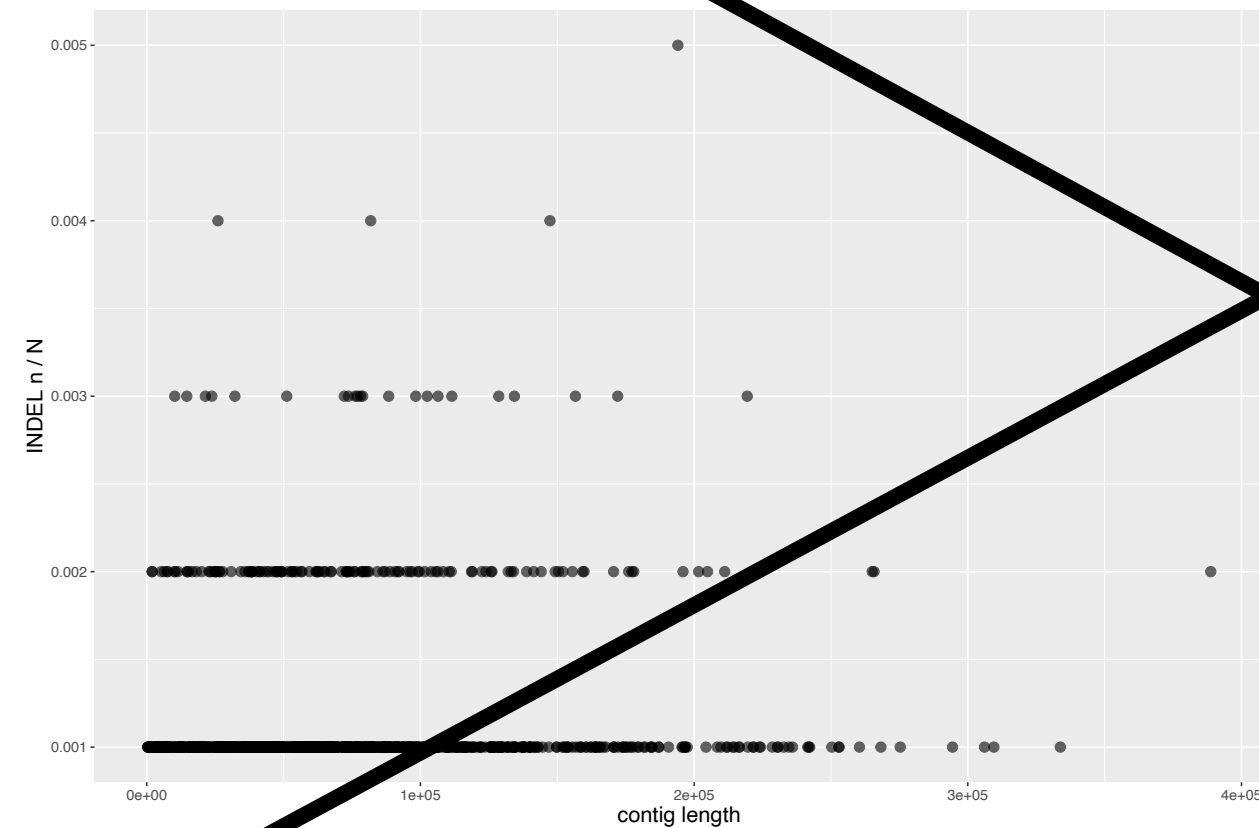


Figure 3a. Proportions (left) and counts (right) of INDELs per contig.

2. Clustering of (different types) markers

- All SNPs = clinal + non-clinal

$$n/N = \frac{n_i \text{ marker}}{N \text{ marker}} \quad \begin{array}{l} n_i = \text{number of SNPs in contig } i \\ N = \text{Total number of SNPs} \end{array}$$

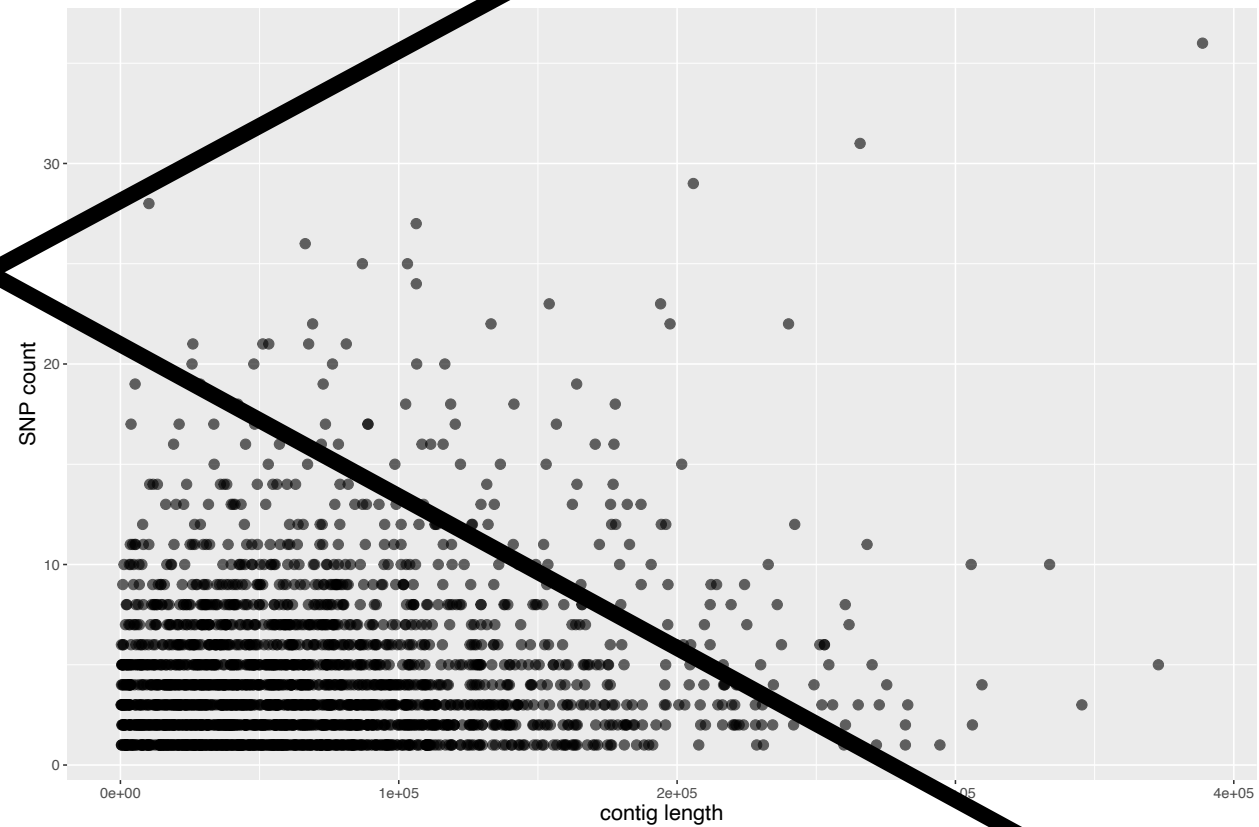
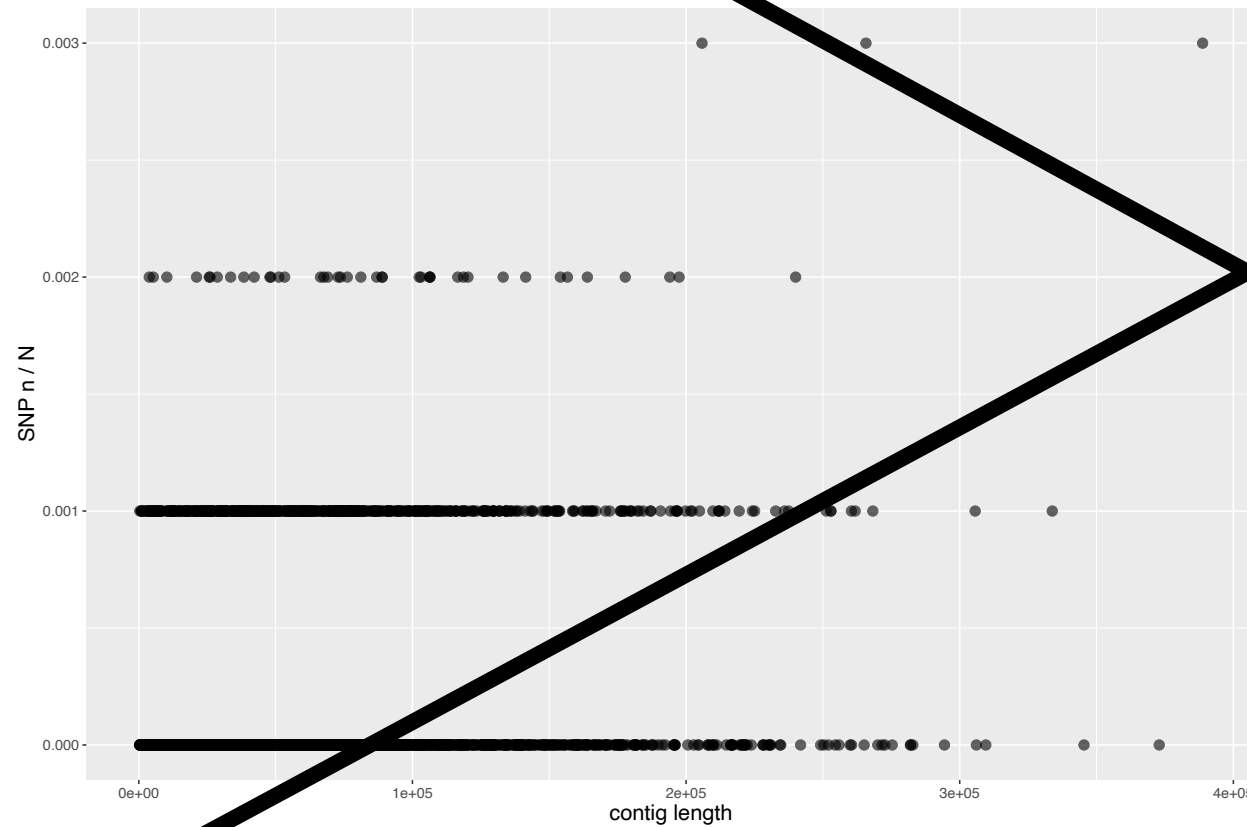


Figure 3b. Proportions (left) and counts (right) of SNPs per contig.

2. Clustering of (different types) markers

- All INDELs and SNPs in the same contigs

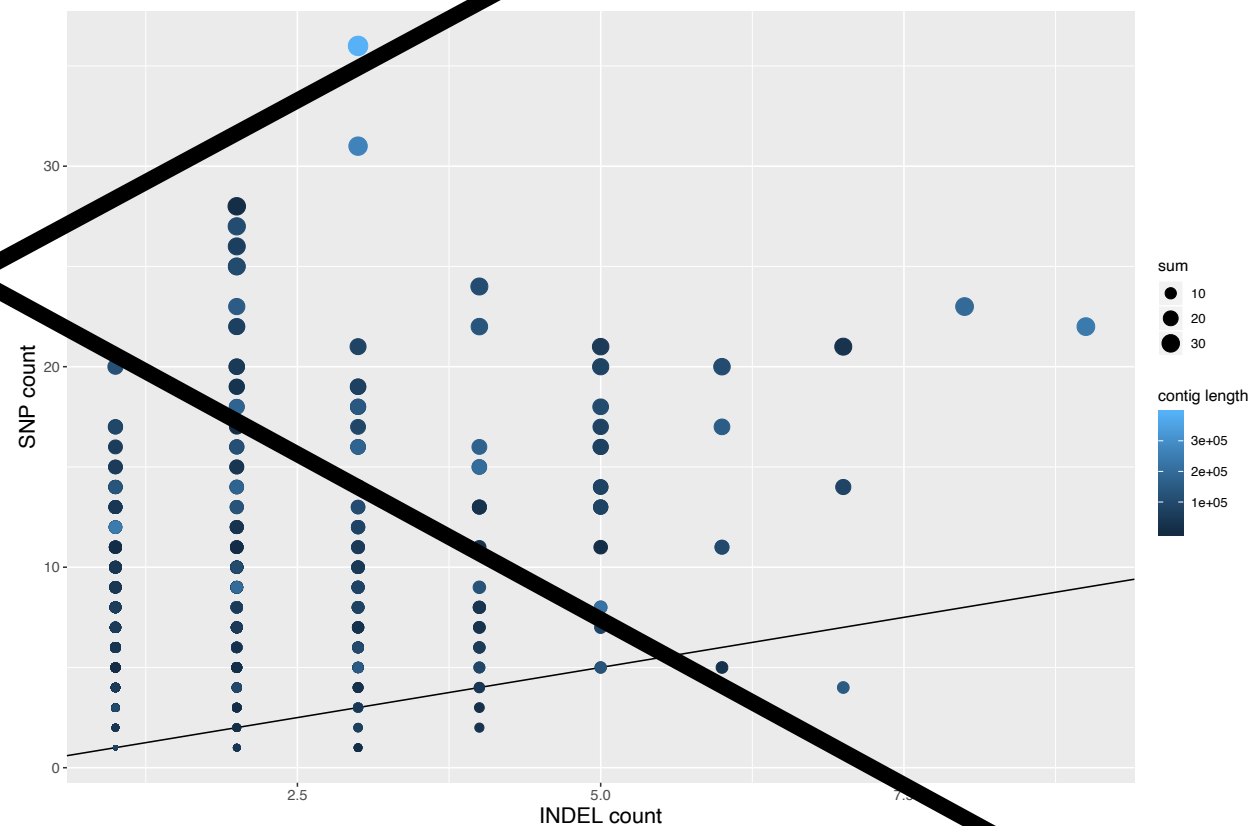
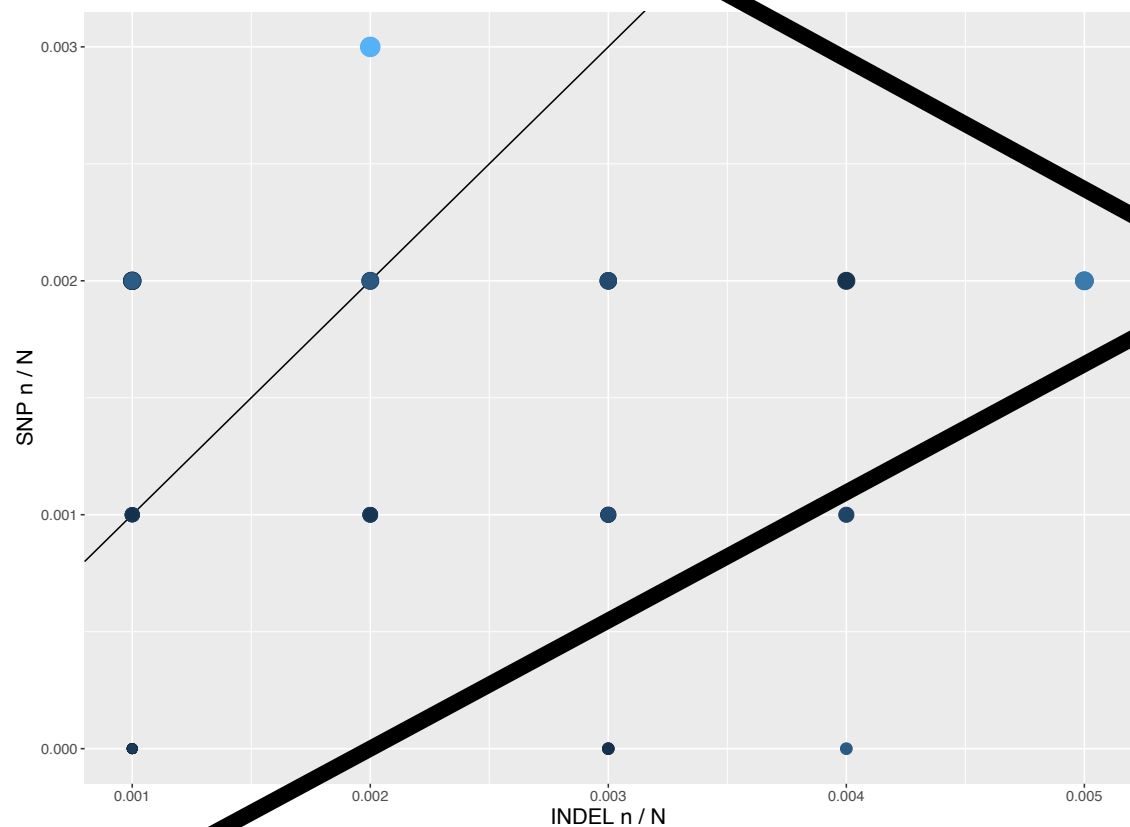


Figure 3c. Relationship between SNPs and INDELs with respect to their proportions (left) and their counts (right) per contig. Proportions and counts of SNPs and INDELs are the same as in Fig. 1-2.

3. Derived allele frequencies (in progress; for now minor allele frequencies Fig. 4-5)

- Ancestral allele was inferred from called genotypes:
 1. Reference allele = ancestral allele
compressa is homo for the reference allele (0)
 2. Alternative allele = ancestral allele
compressa is homo for the alternative allele (2)
 3. Unknown ancestry
compressa is het (1)

3. Minor allele frequencies (GATK call)

- INDELs and SNPs after filtering but before cline analysis.
- All six hybrid zones combined.

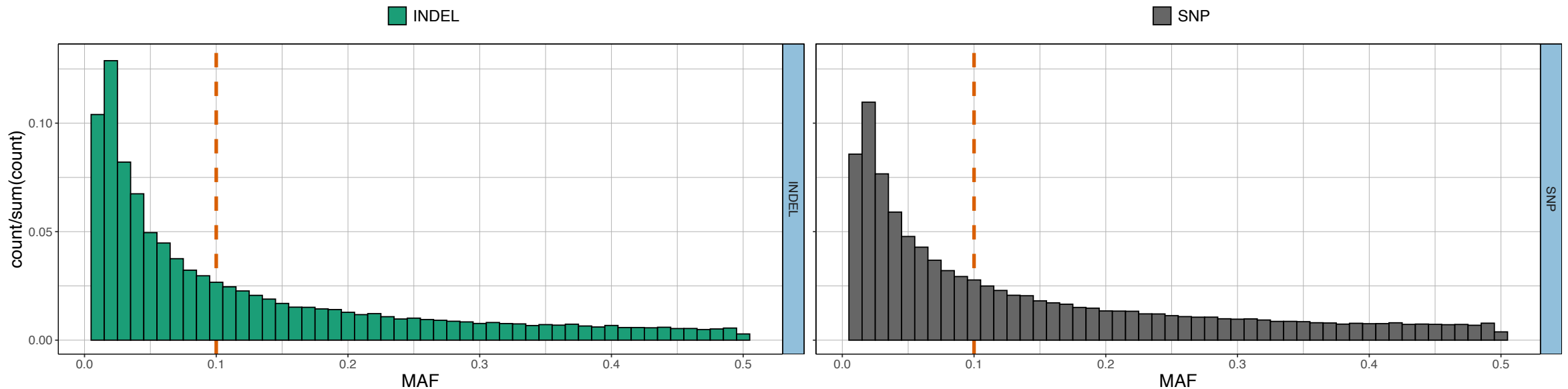


Figure 4a. Proportions of minor allele frequencies of INDELs (left) and SNPs (right) after filtering but before cline analysis. Bin width is 0.01 and orange dashed line marks the maf filter in the cline analysis (0.1).

3. Minor allele frequencies (GATK call)

- INDELs and SNPs after filtering and cline analysis.
- All six hybrid zones combined.

For the joint AFS between INDELs and SNPs, see (attached) file [short_indels.nb.html](#)

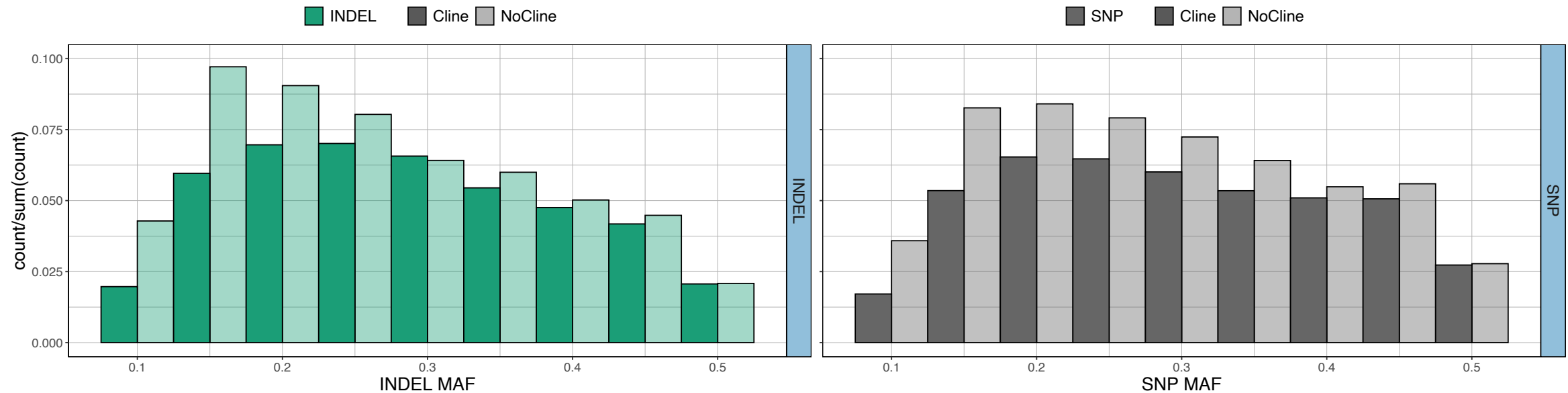


Figure 5a. Proportions of minor allele frequencies of INDELs (left) and SNPs (right) after filtering and cline analysis. Bin width is 0.05. Clinal variants are dark coloured and non-clinal variants are light coloured.