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. Distributions of cline parameters

1. Outlier sharing

Total number of SNP: 11225

Total number of INDEL: 1752

	Proportion of SNP with significant clines.	Proportion of INDEL with significant clines.
CZA left:	0.5317595	0.5296804
CZA right:	0.4457016	0.4549087
CZB left:	0.3277506	0.3413242
CZB right:	0.4244989	0.4092466
CZD left:	0.4473942	0.4737443
CZD right:	0.4823163	0.4834475

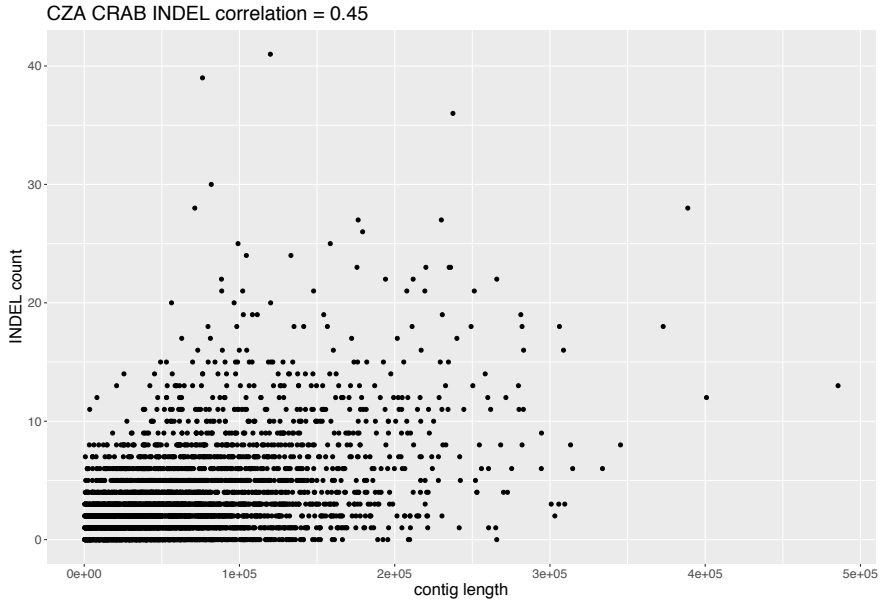
	Proportions of SNP outliers that are shared.	Proportions of INDEL outliers that are shared.
CZA left and right:	0.6160714	0.7058824
CZB left and right:	0.5178571	0.4705882
CZD left and right:	0.6339286	0.6470588
CZA and CZB:	0.359375	0.3529412
CZA and CZD:	0.4107143	0.4117647
CZB and CZD:	0.484375	0.4411765

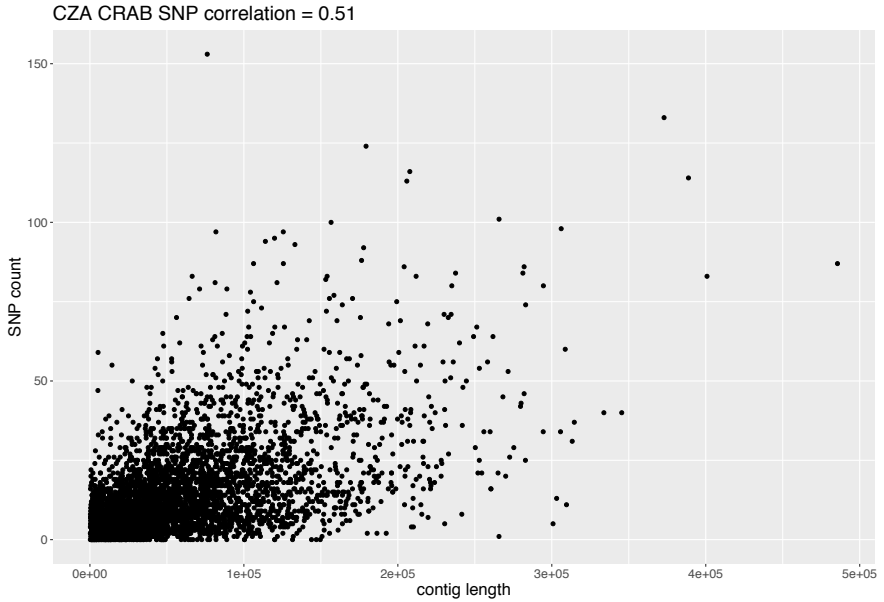
Number of SNP outliers found in 1 hybrid zone(s):	142	24
Number of SNP outliers found in 2 hybrid zone(s):	66	7
Number of SNP outliers found in 3 hybrid zone(s):	29	7
Number of SNP outliers found in 4 hybrid zone(s):	27	5
Number of SNP outliers found in 5 hybrid zone(s):	25	1
Number of SNP outliers found in 6 hybrid zone(s):	13	3

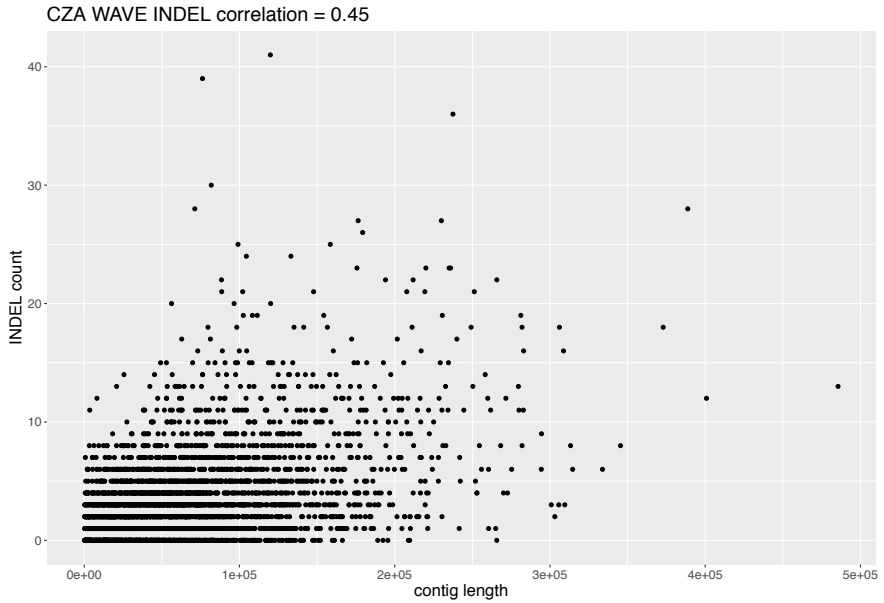
Prop. of SNP outliers in inversions found in 1 zone(s):	0.556	0.625
Prop. of SNP outliers in inversions found in 2 zone(s):	0.636	0.57
Prop. of SNP outliers in inversions found in 3 zone(s):	0.862	0.86
Prop. of SNP outliers in inversions found in 4 zone(s):	0.889	1
Prop. of SNP outliers in inversions found in 5 zone(s):	0.92	1
Prop. of SNP outliers in inversions found in 6 zone(s):	1	1

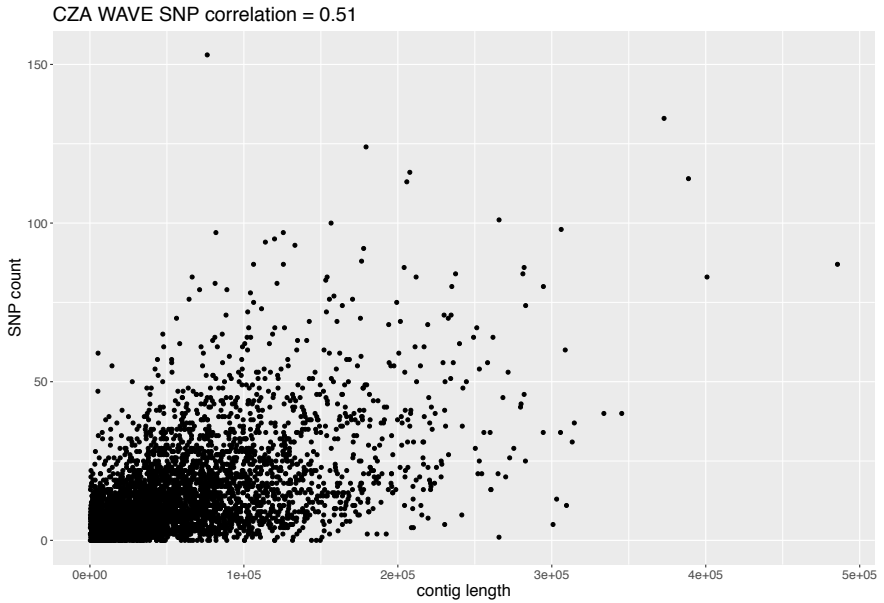
2. Clustering of (different types) markers

CZA: INDELs and SNPs after filtering but before cline analysis.



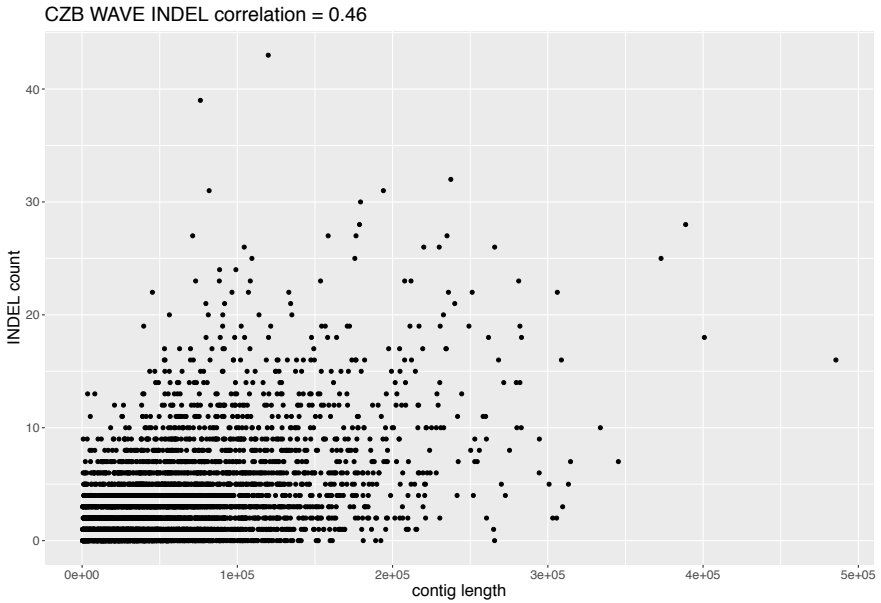


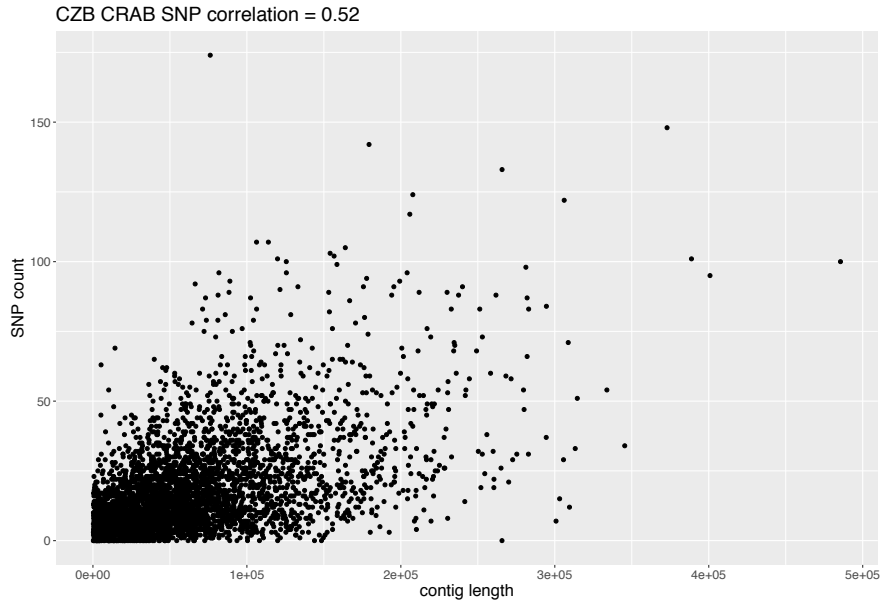


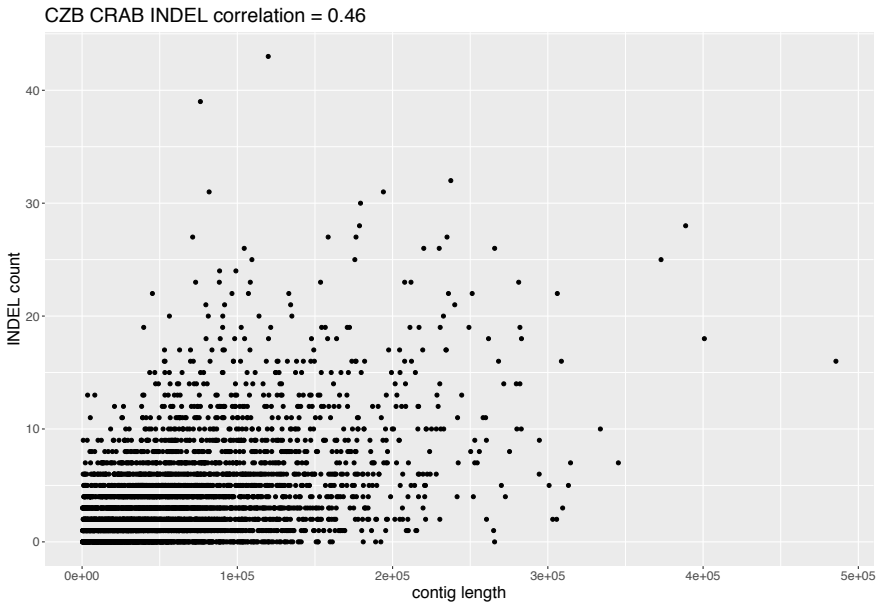


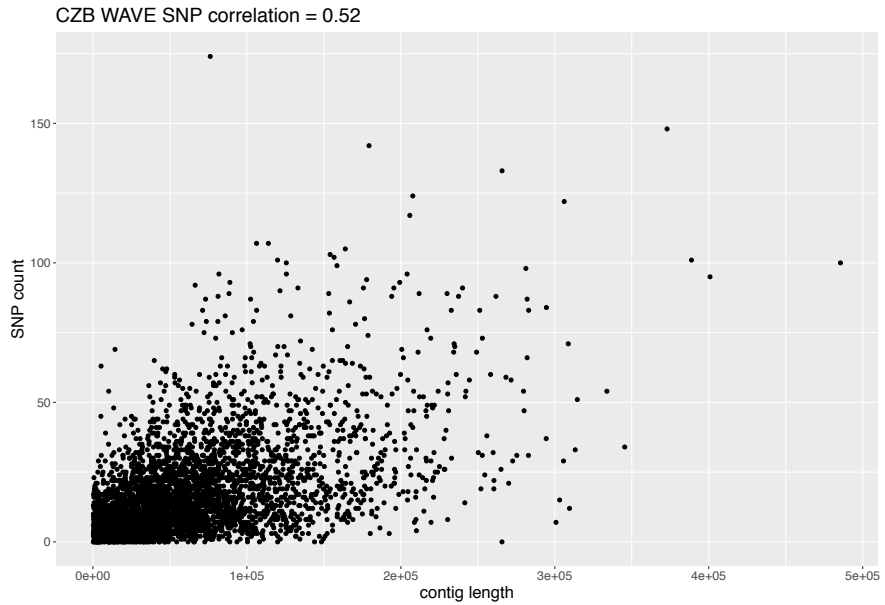
2. Clustering of (different types) markers

CZB: INDELs and SNPs after filtering but before cline analysis.



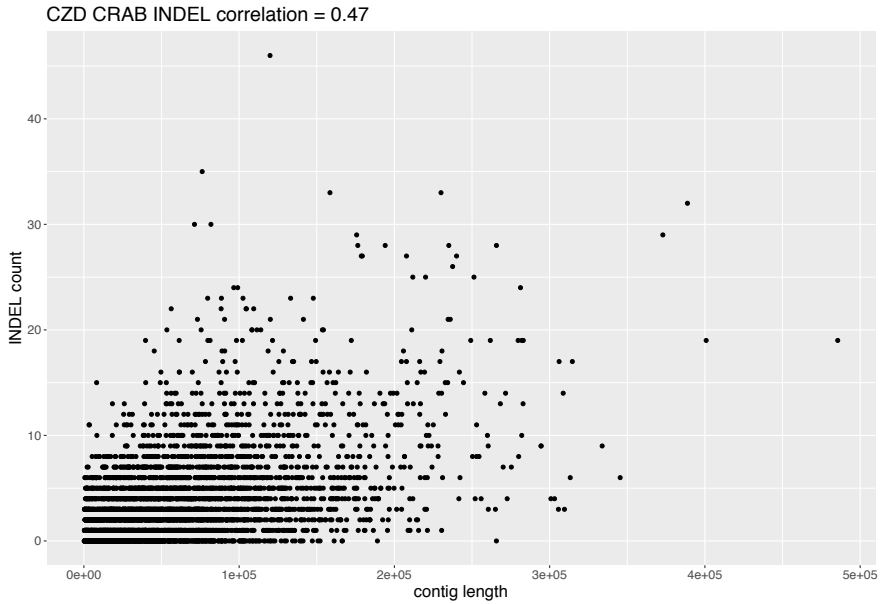


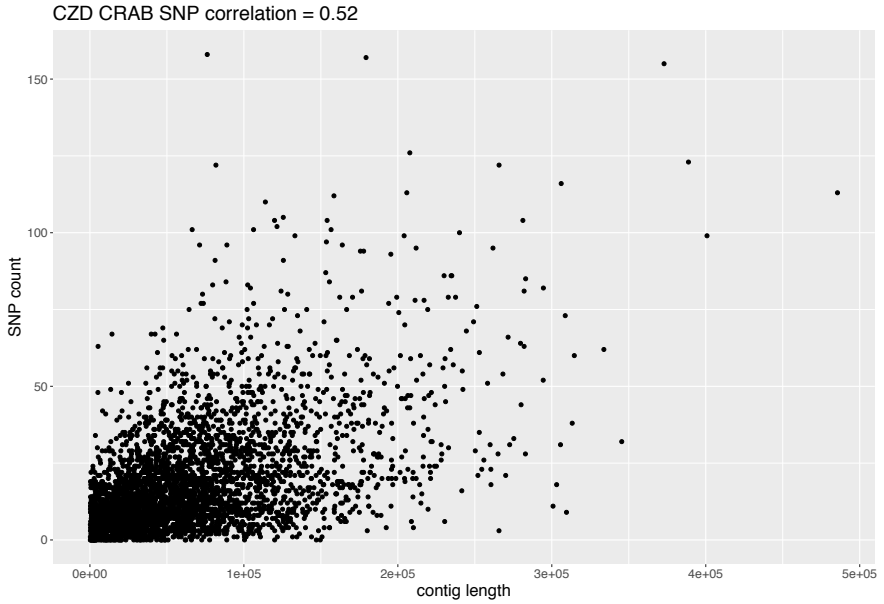


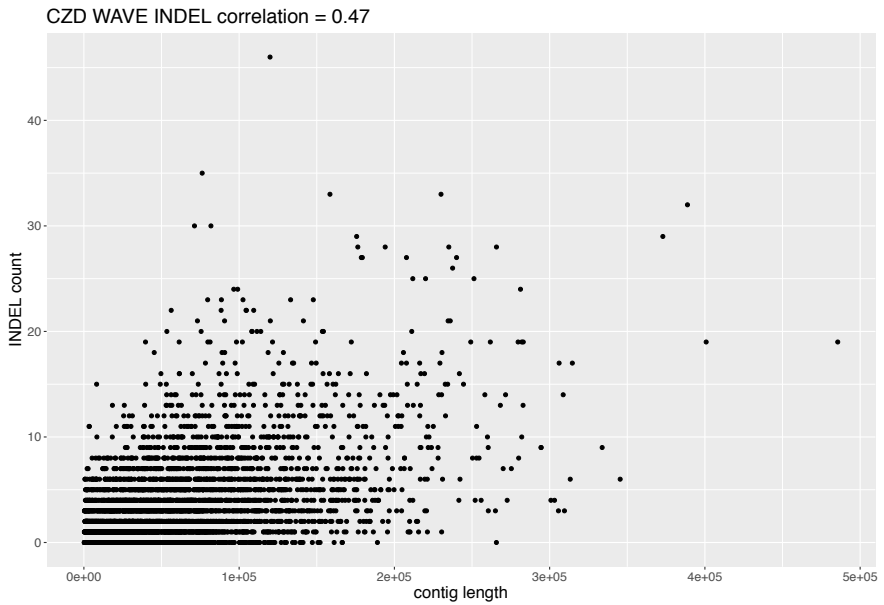


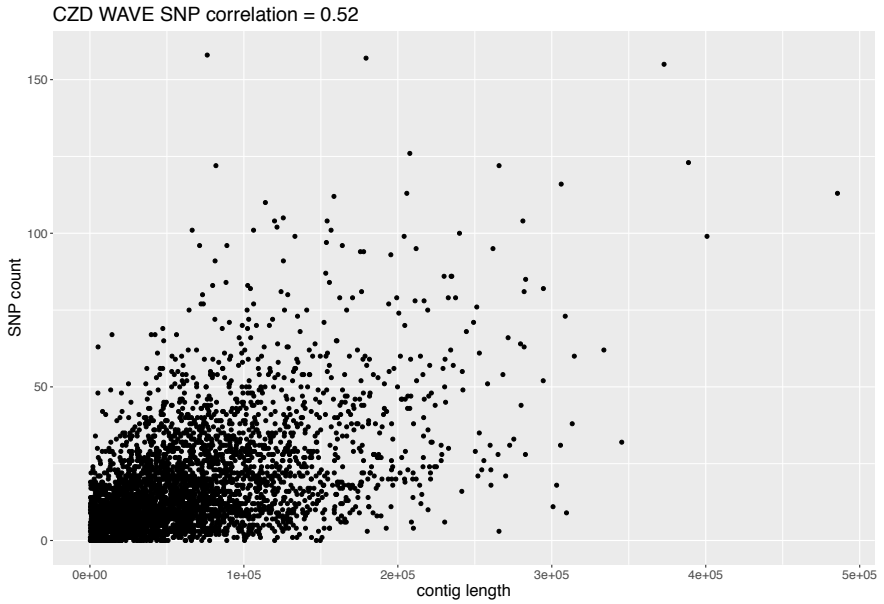
2. Clustering of (different types) markers

CZD: INDELs and SNPs after filtering but before cline analysis.









2. Clustering of (different types) markers

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

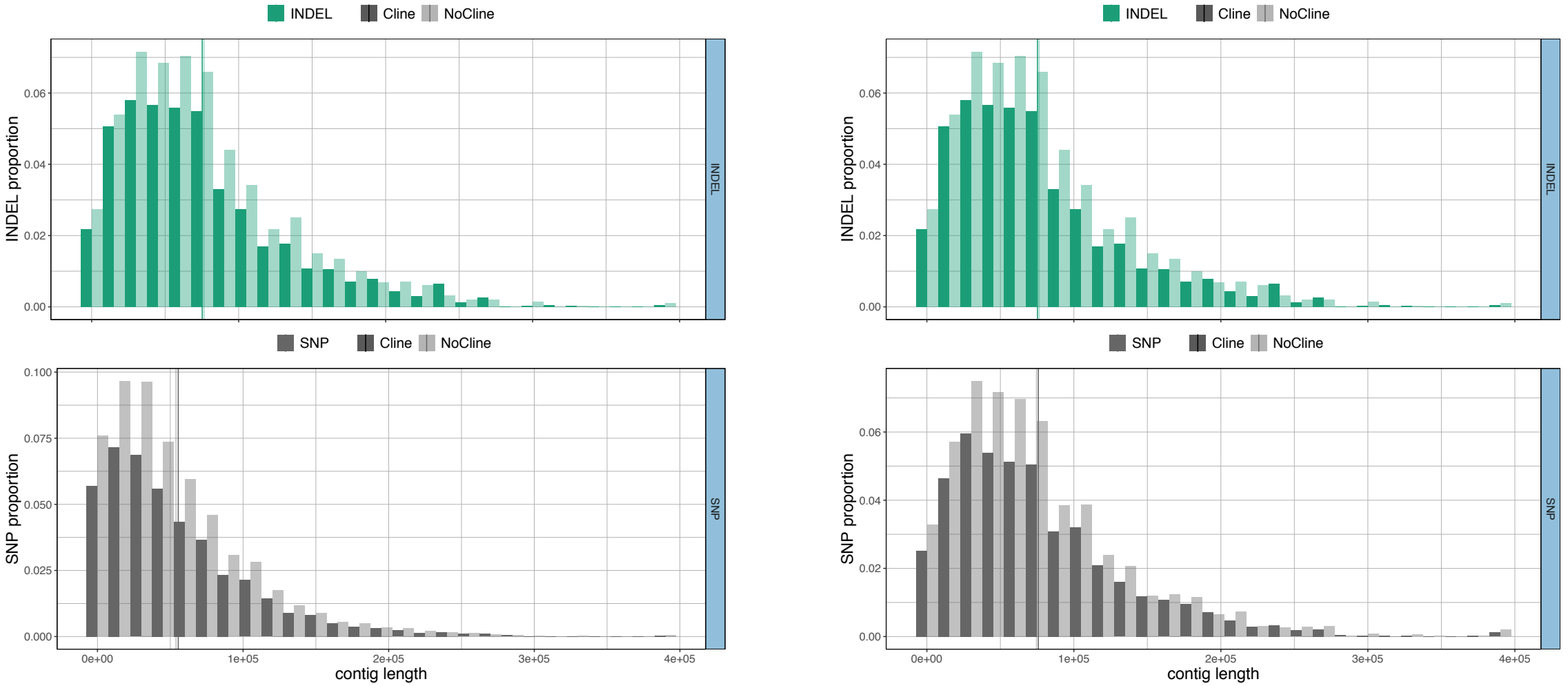


Figure 3a. Marker proportions over contig length. $Proportion = count/\Sigma\, count$ per marker type and bin width = 15000 base pairs. Clinal variants are dark coloured and non-clinal variants are light coloured. Left: SNP call using SAMtools and INDEL call using GATK. Right: both INDELs and SNPs were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

3. Derived allele frequencies - INDELs

- Ancestral state was inferred from called genotypes:

    1. Reference allele = ancestral allele = ref_anc
    compressa is homo for the reference allele (0)

    2. Alternative allele = ancestral allele = alt_anc
    compressa is homo for the alternative allele (2)

    3. Unknown ancestry = het
    compressa is het (1)

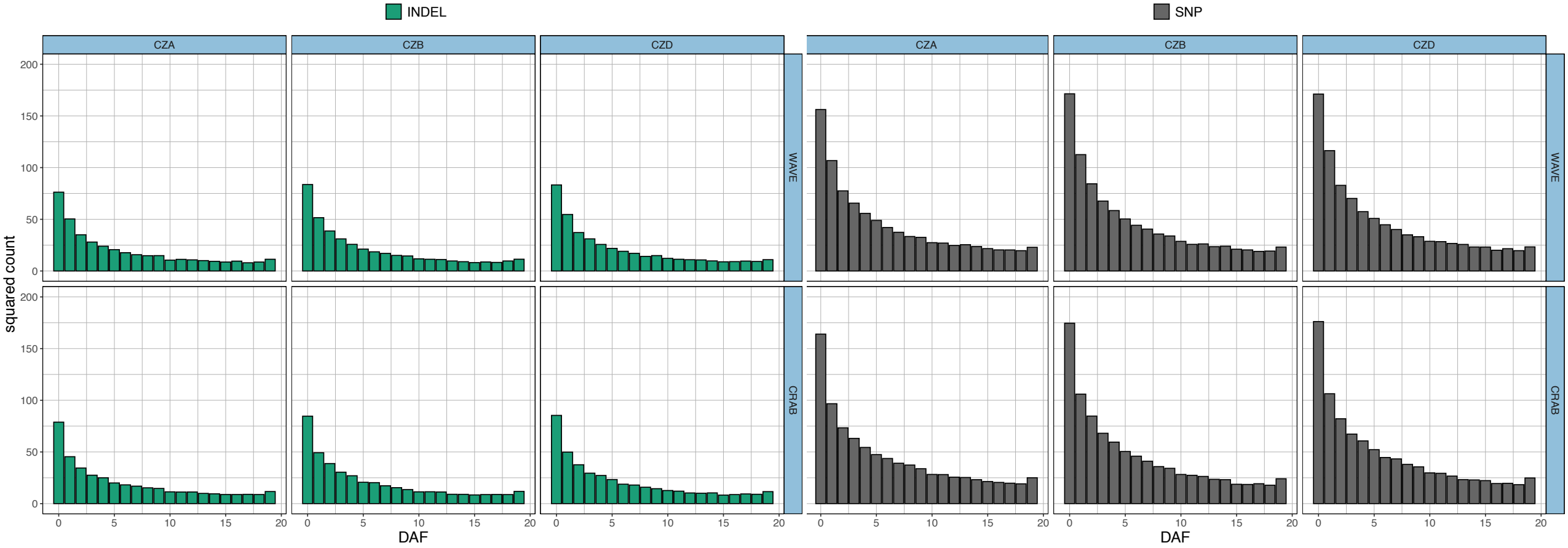
For the rest of the results, kept only variants highlighted in green.

Table 1. Count of INDELs and SNPs for each combination of possible allelic states given one outgroup (L. compressa) with two samples (NE and W). There are two combinations in which the allelic state is concordant in both samples (in green), eight in which the allelic state can only be retrieved from one sample (in yellow) and finally, five in which the allelic state cannot be inferred (in red).

NE_Lcomp	W_Lcomp	INDEL	SNP
alt_anc	alt_anc	5305	27188
alt_anc	het	528	3543
alt_anc	NA	245	1097
alt_anc	ref_anc	511	2195
het	alt_anc	2231	12439
het	het	1577	9691
het	NA	151	627
het	ref_anc	3120	17198
NA	alt_anc	158	765
NA	het	33	267
NA	ref_anc	449	1831
ref_anc	alt_anc	693	3292
ref_anc	het	1422	7462
ref_anc	NA	1003	3675
ref_anc	ref_anc	38884	178715

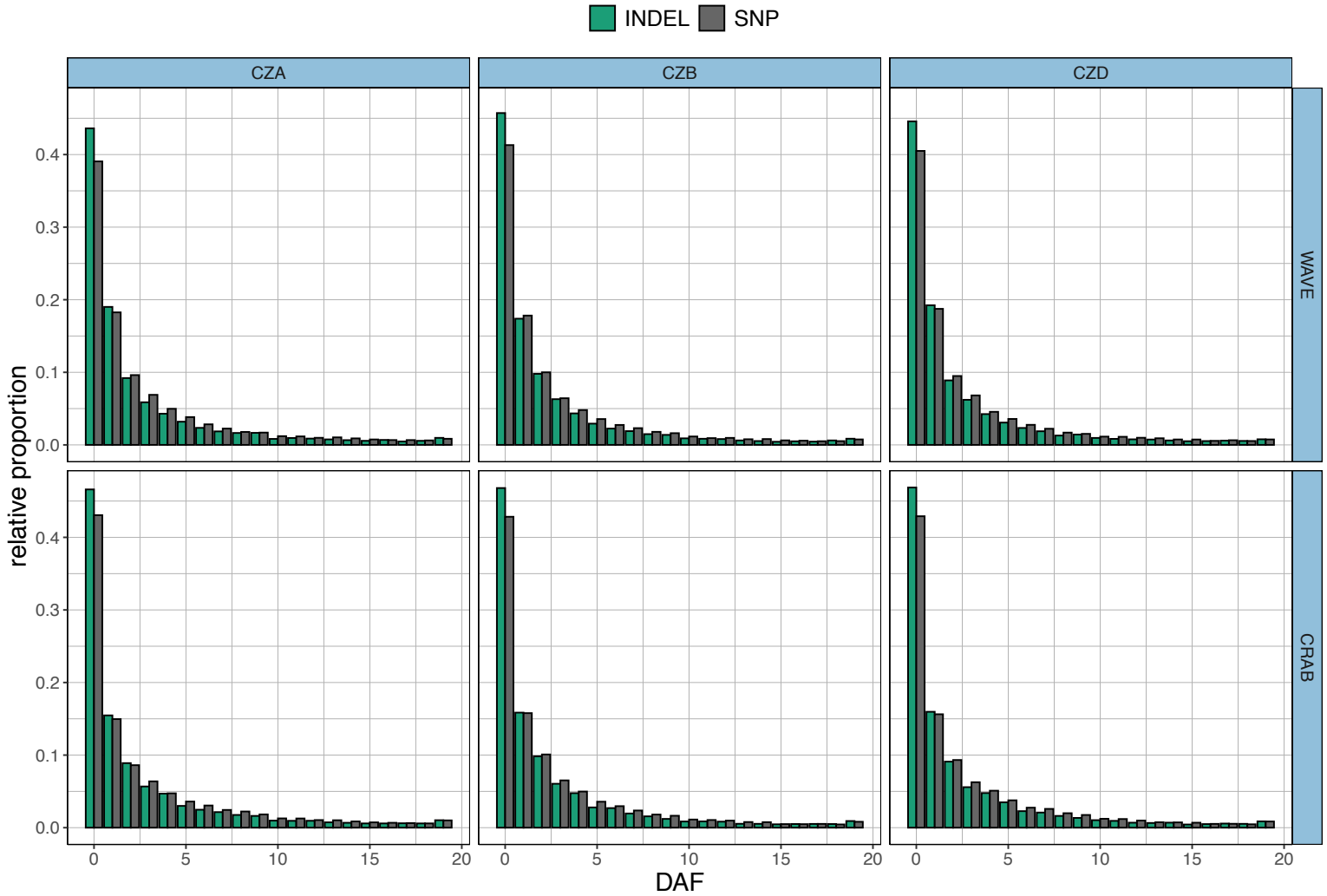
3. Derived allele frequencies (GATK call)

- Square root of count of INDELs and SNPs after filtering but before cline analysis.



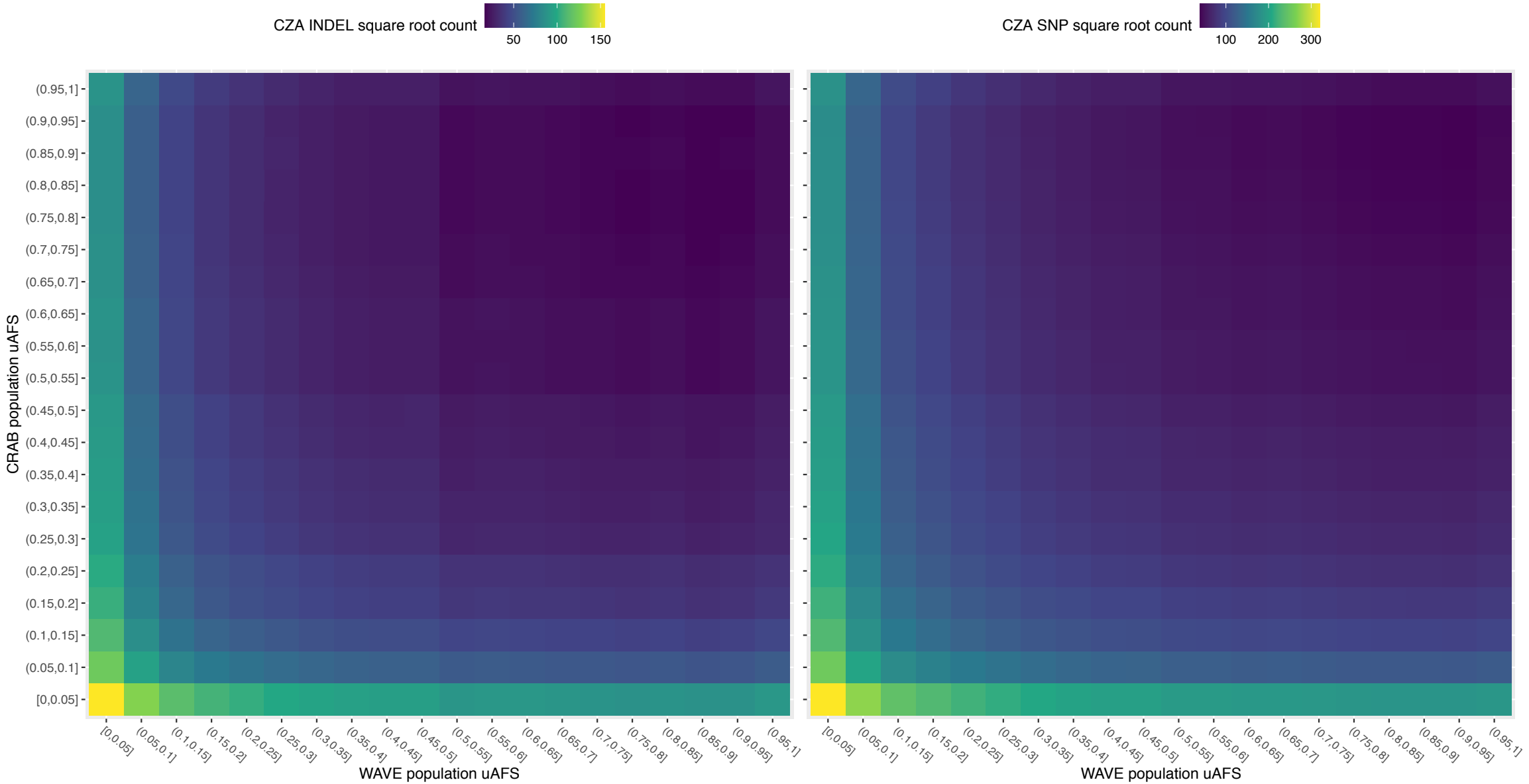
3. Derived allele frequencies (GATK call)

- Relative proportion of INDELs and SNPs after filtering but before cline analysis.



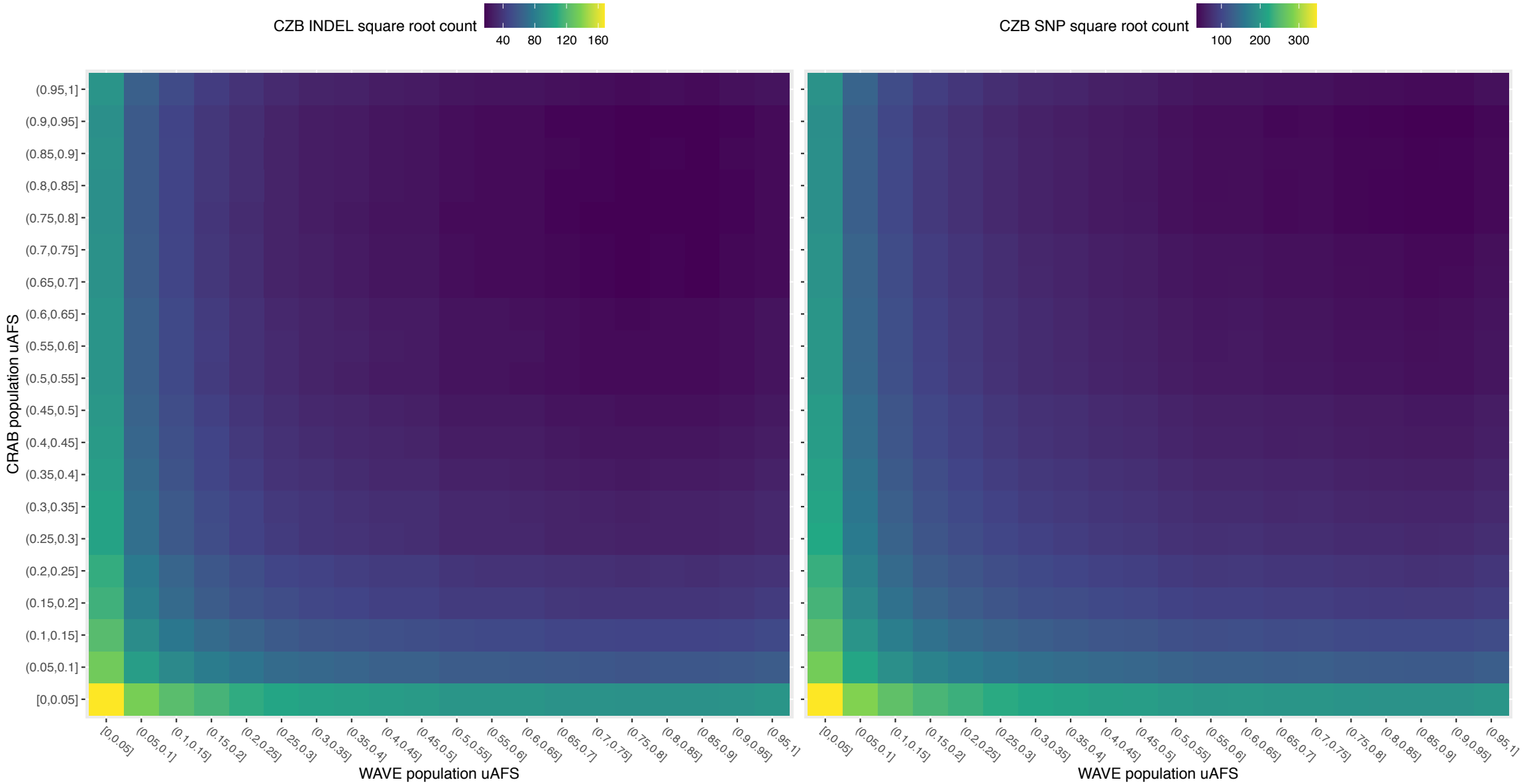
3. Derived allele frequencies (GATK call)

CZA CRAB-WAVE joint allele frequency spectra: square root count.



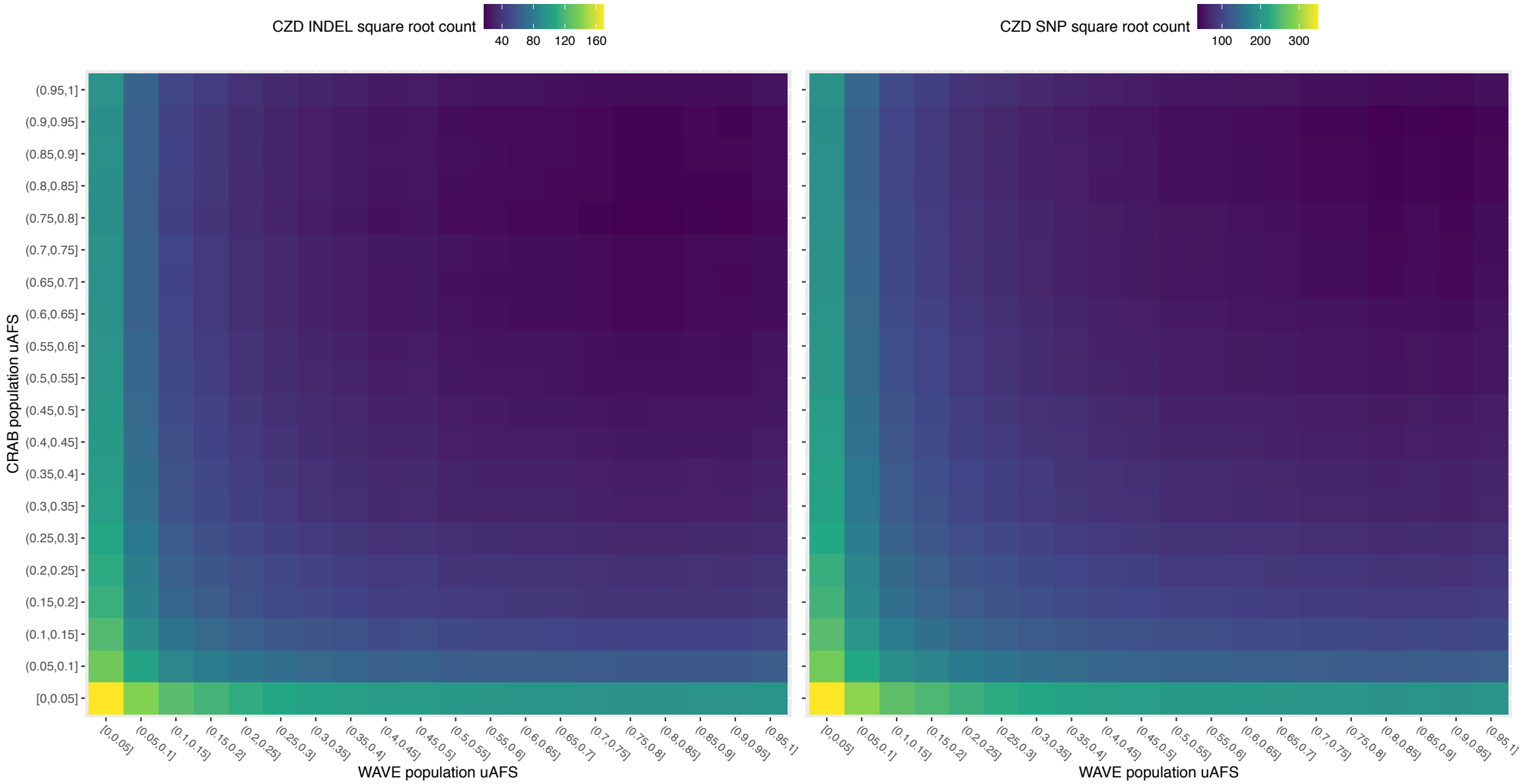
3. Derived allele frequencies (GATK call)

CZB CRAB-WAVE joint allele frequency spectra: square root count.



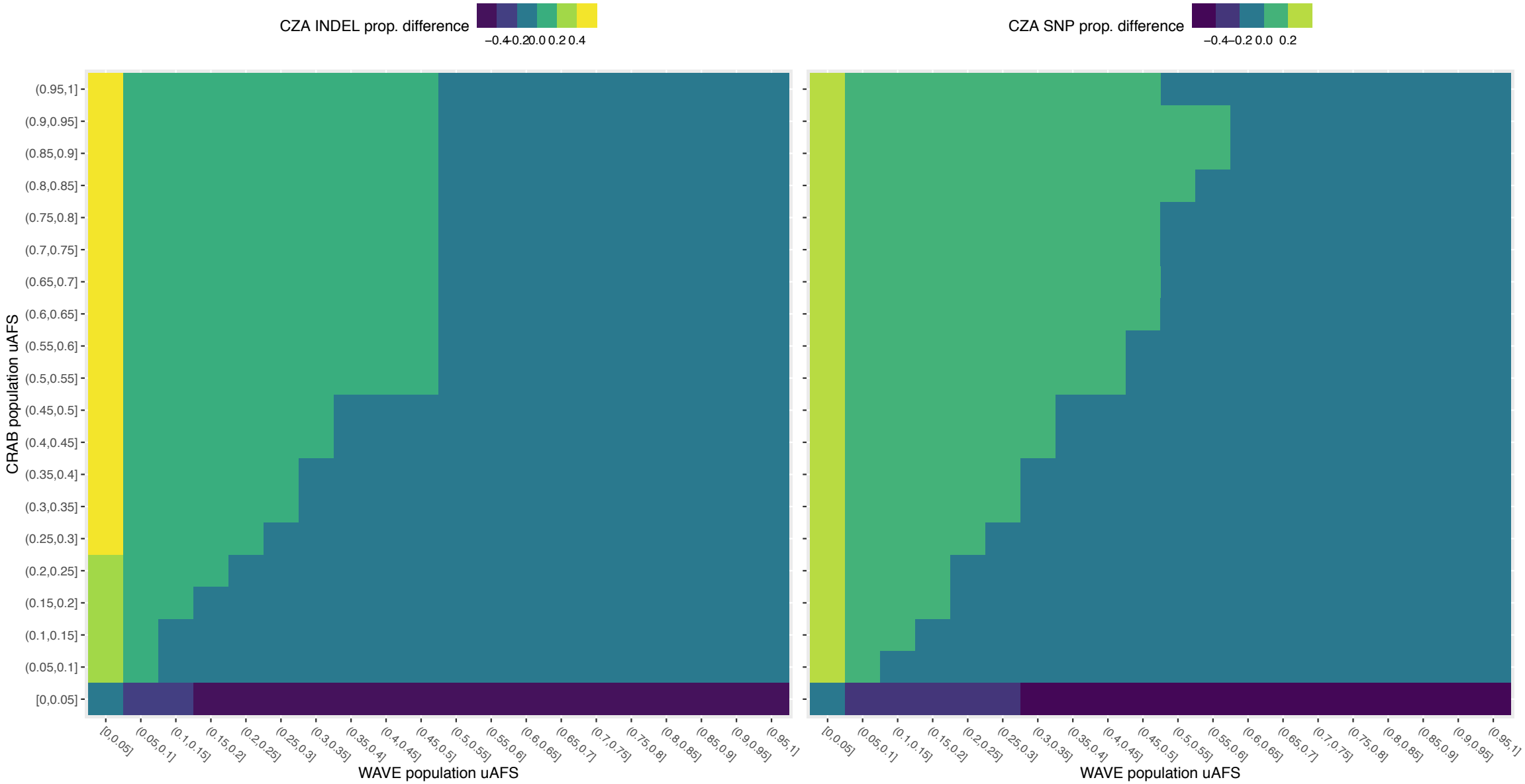
3. Derived allele frequencies (GATK call)

CZD CRAB-WAVE joint allele frequency spectra: square root count.



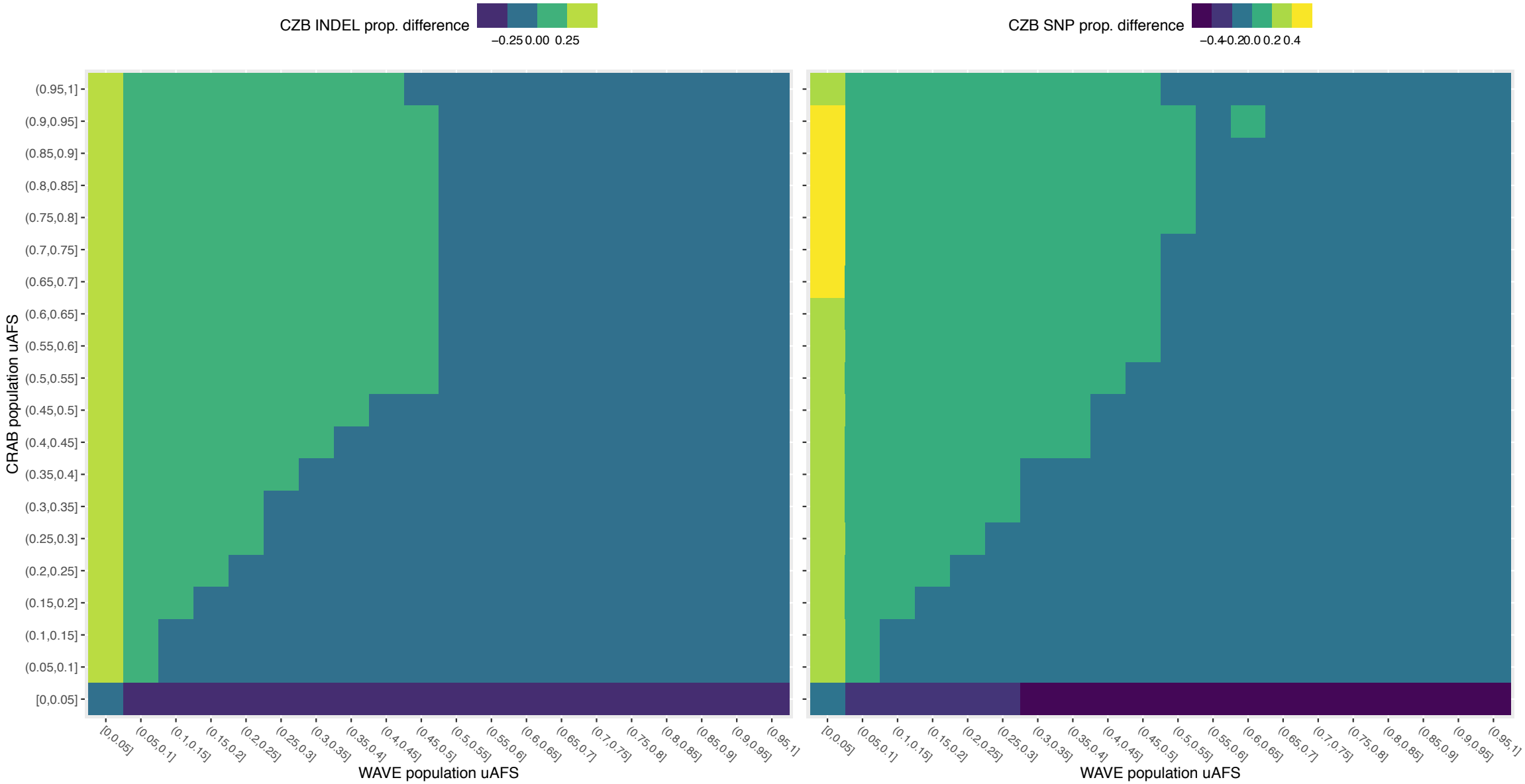
3. Derived allele frequencies (GATK call)

CZA CRAB-WAVE joint allele frequency spectra: difference in relative proportions.



3. Derived allele frequencies (GATK call)

CZB CRAB-WAVE joint allele frequency spectra: difference in relative proportions.



3. Derived allele frequencies (GATK call)

CZD CRAB-WAVE joint allele frequency spectra: difference in relative proportions.