***Melitaea britomartis* population structure**

Intermediate report

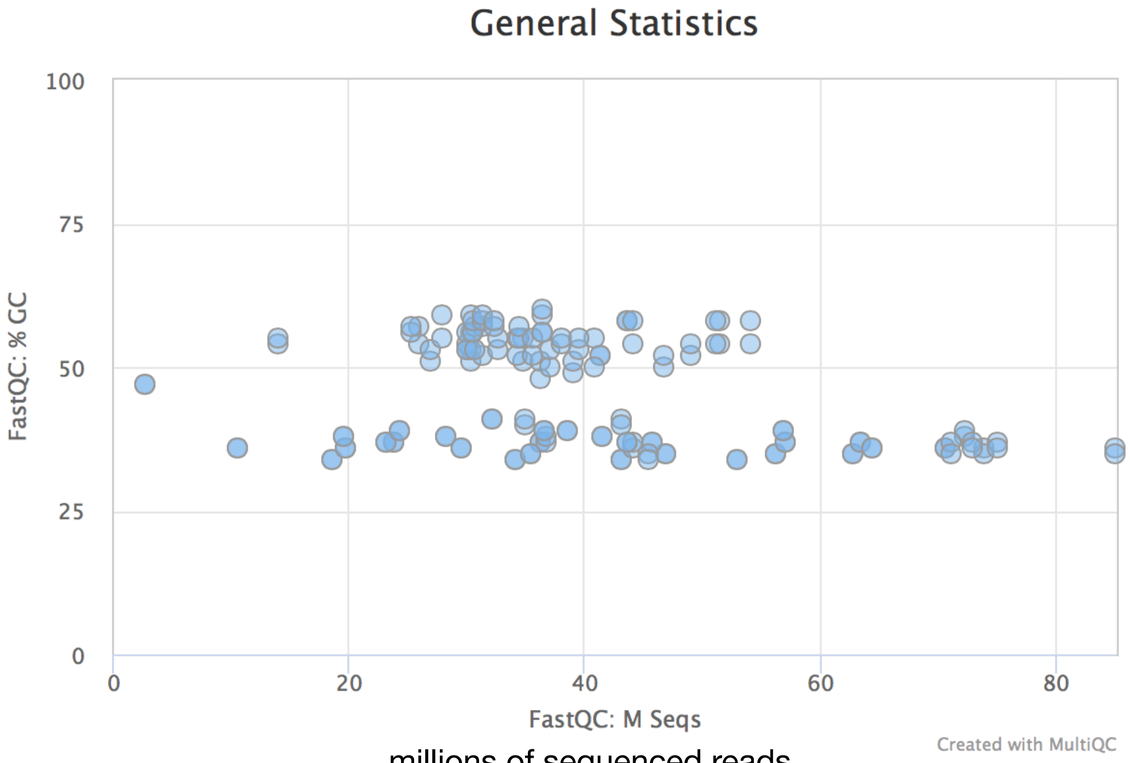
Raw sequencing data QC

Read mapping QC
Variant calling summary
Basic population structure (PCA)
Mitochondrial tree (COI)

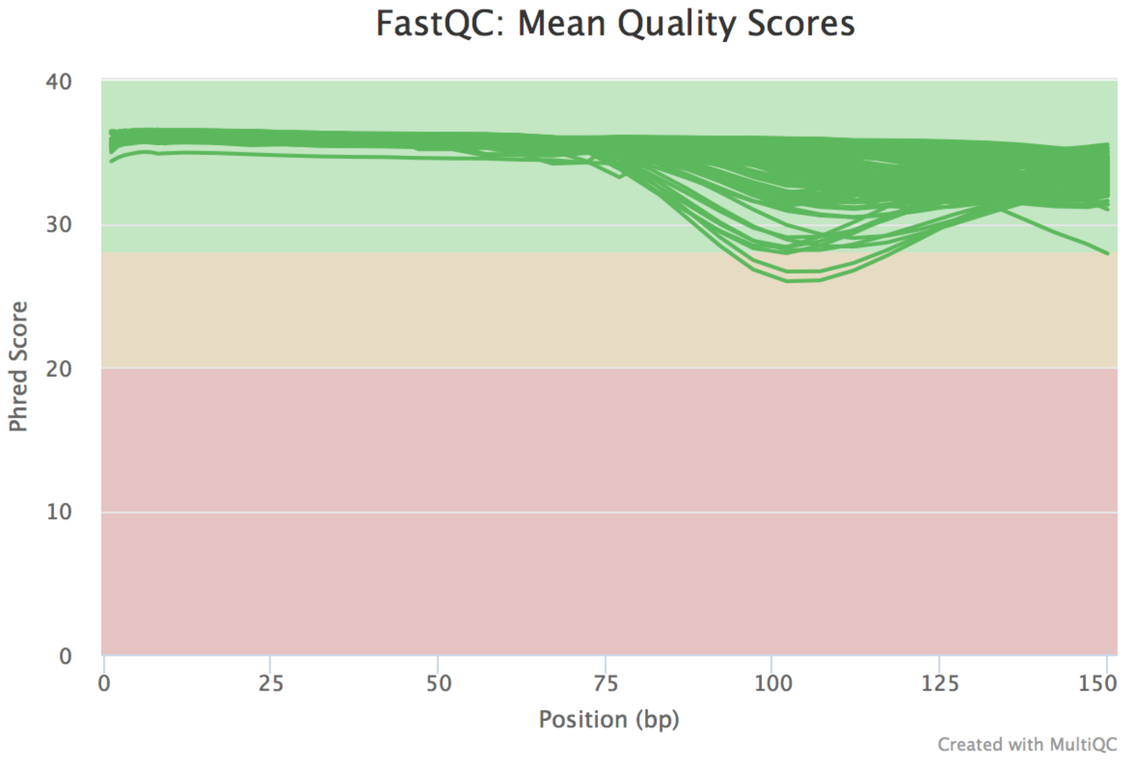
***Melitaea britomartis* population structure**

Raw sequencing data QC

10-85M (single outlier 2.5M) of sequences reads obtained, most of the high quality
GC content in large proportion of the samples appeared abnormal (37% is expected)



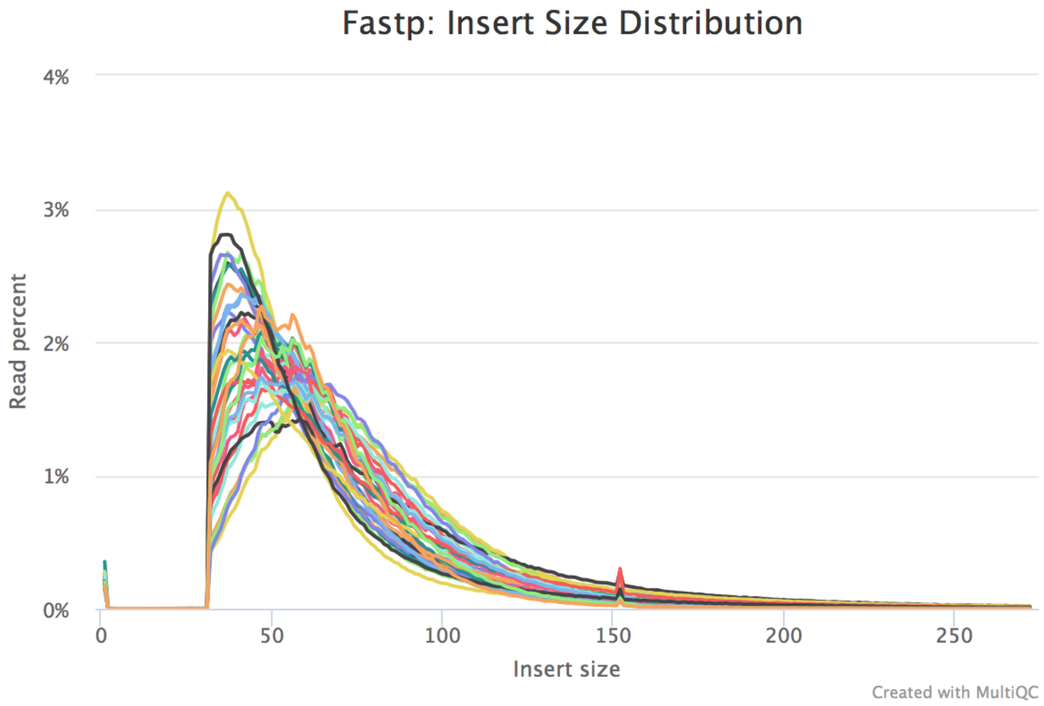
millions of sequenced reads



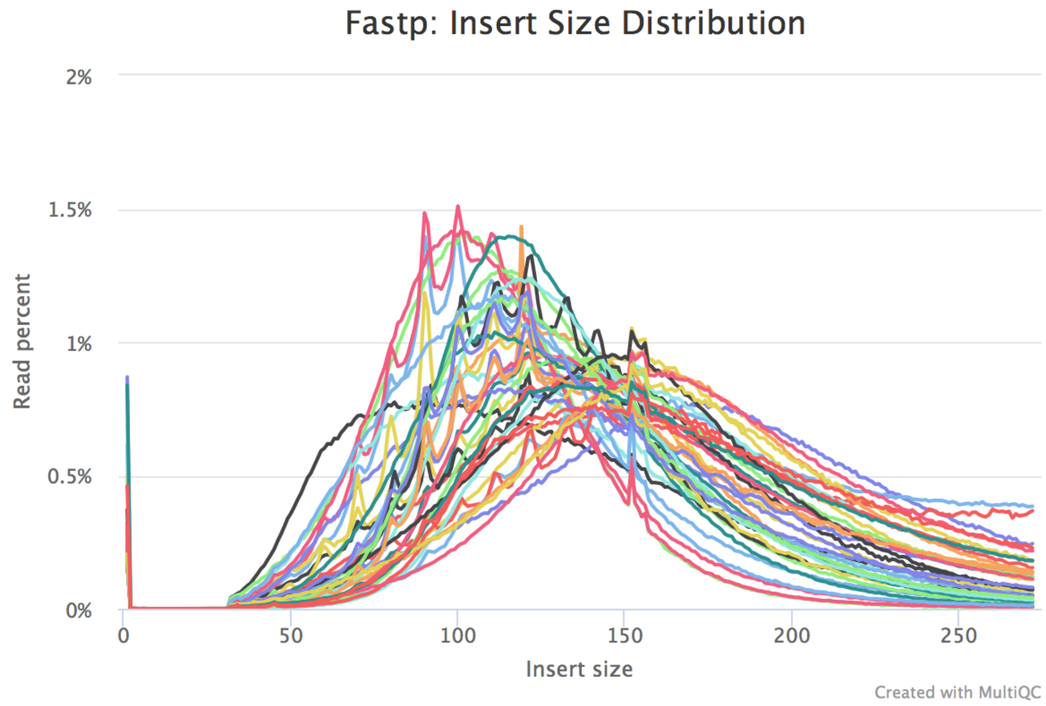
***Melitaea britomartis* population structure**

Read mapping QC

Read mapping revealed difference in insert sizes (length of sequences fragments) between historical and contemporary samples



“Historical” samples
input GC > 50%

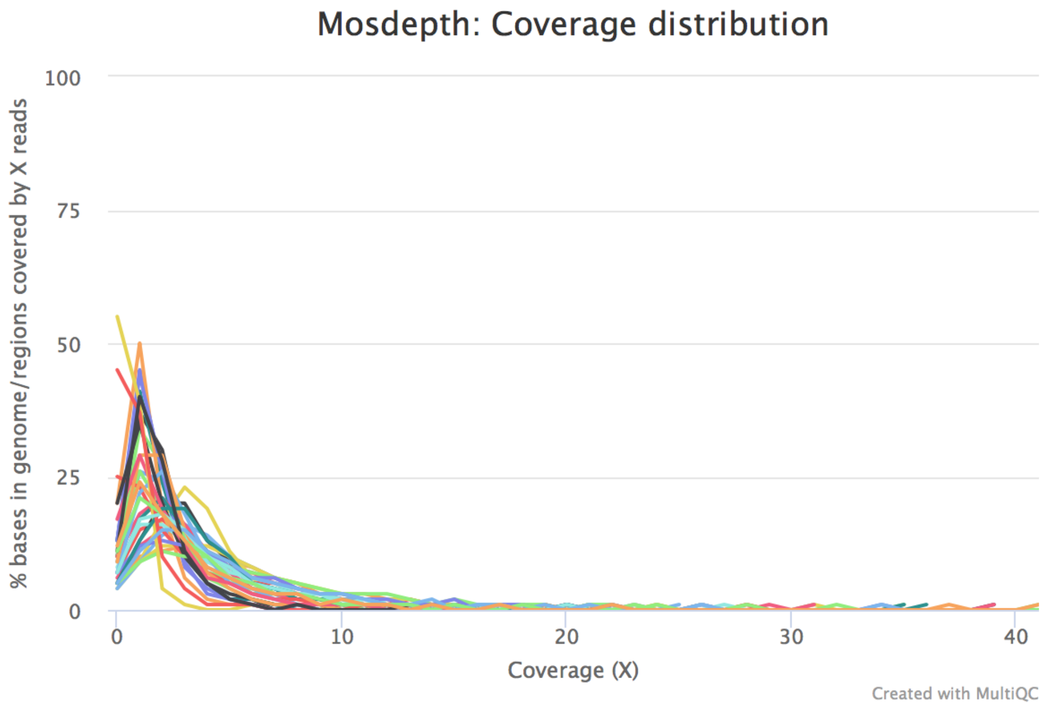


“Contemporary” samples
input GC < 50%

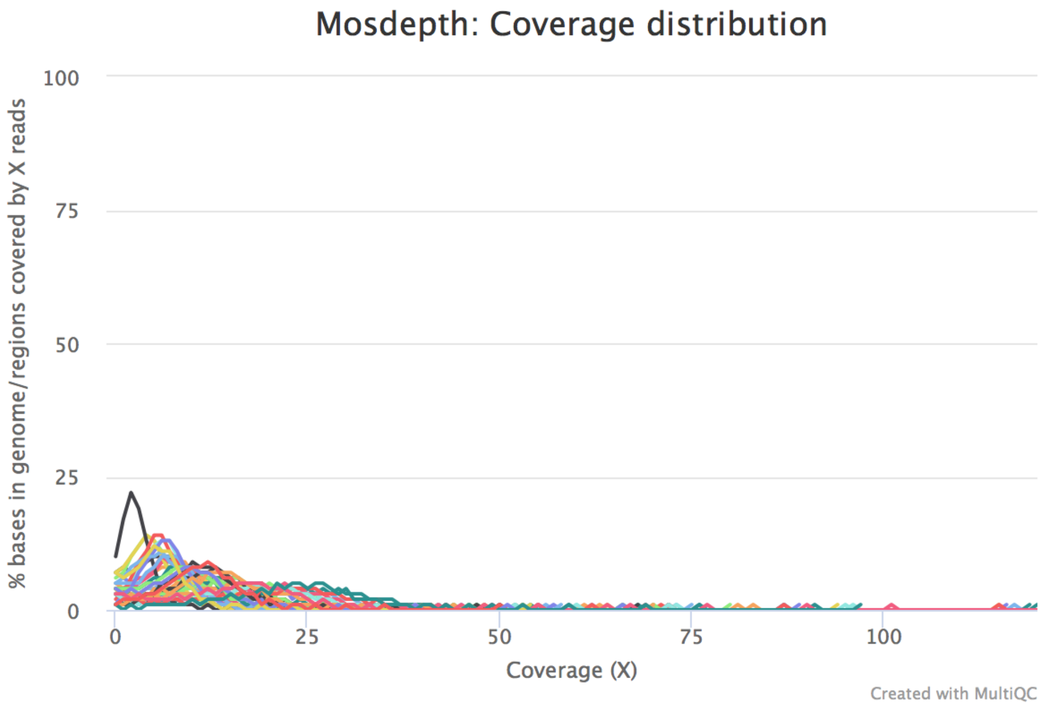
***Melitaea britomartis* population structure**

Read mapping QC

Historical samples showed lower coverage / mapping depth, variant calling strategy adjusted to accommodate



“Historical” samples
input GC > 50%

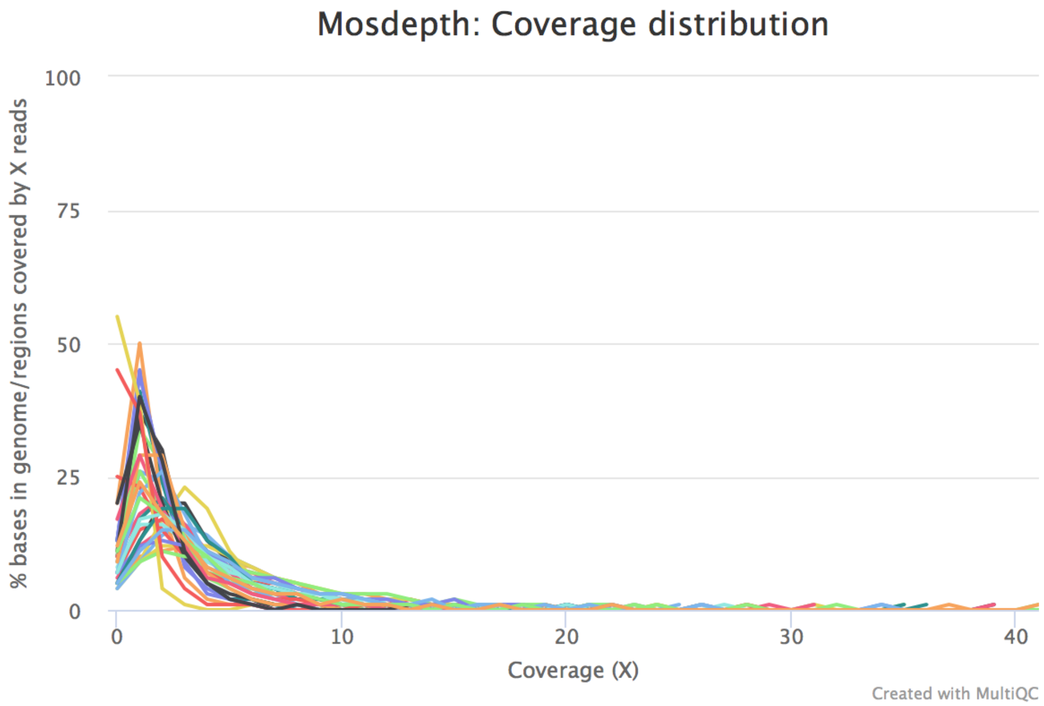


“Contemporary” samples
input GC < 50%

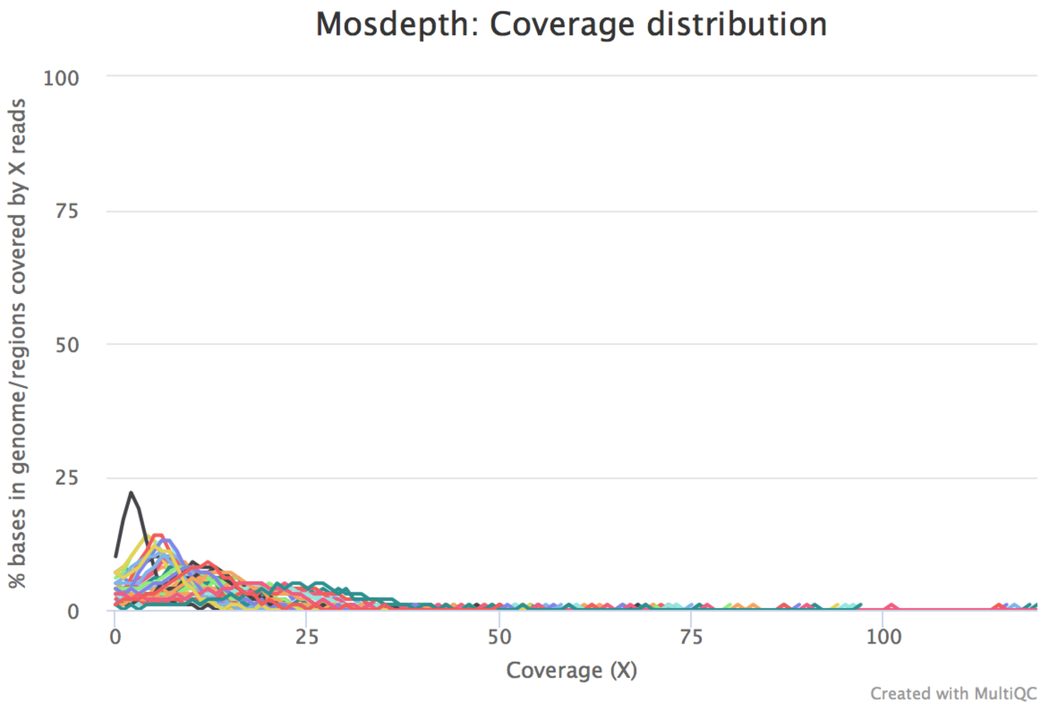
***Melitaea britomartis* population structure**

Read mapping QC

Historical samples showed lower coverage / mapping depth, variant calling strategy adjusted to accommodate



“Historical” samples
input GC > 50%

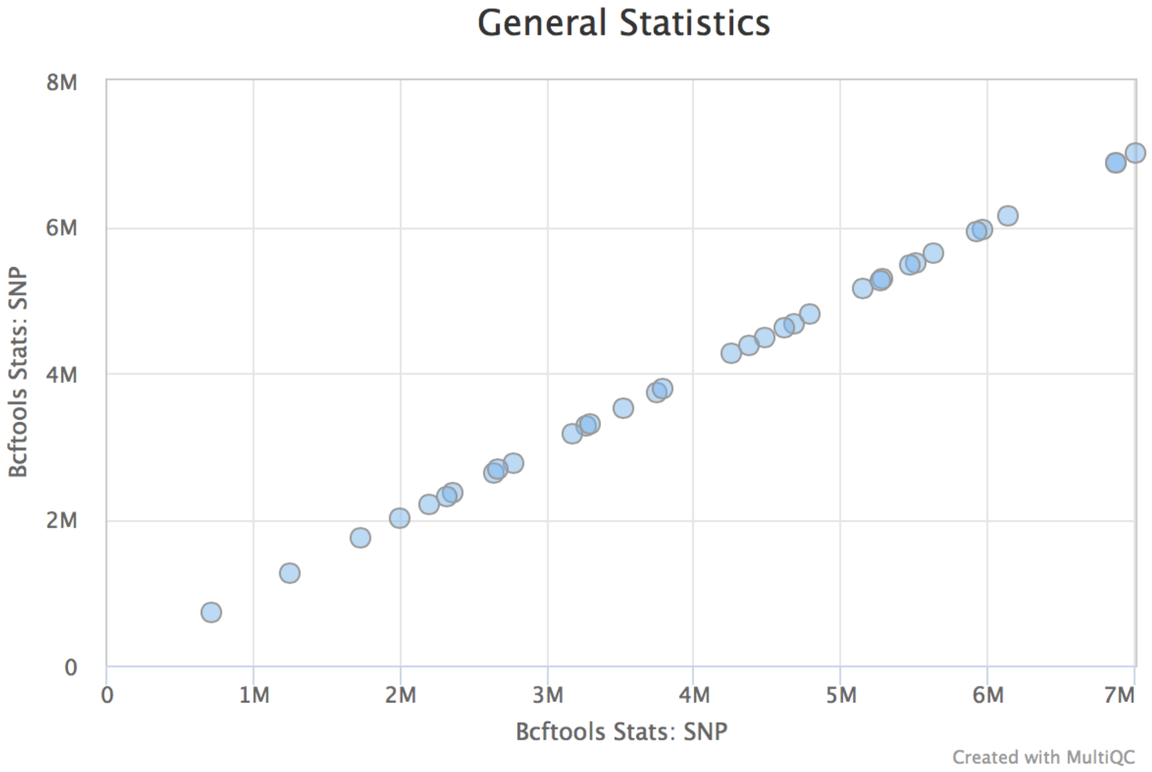


“Contemporary” samples
input GC < 50%

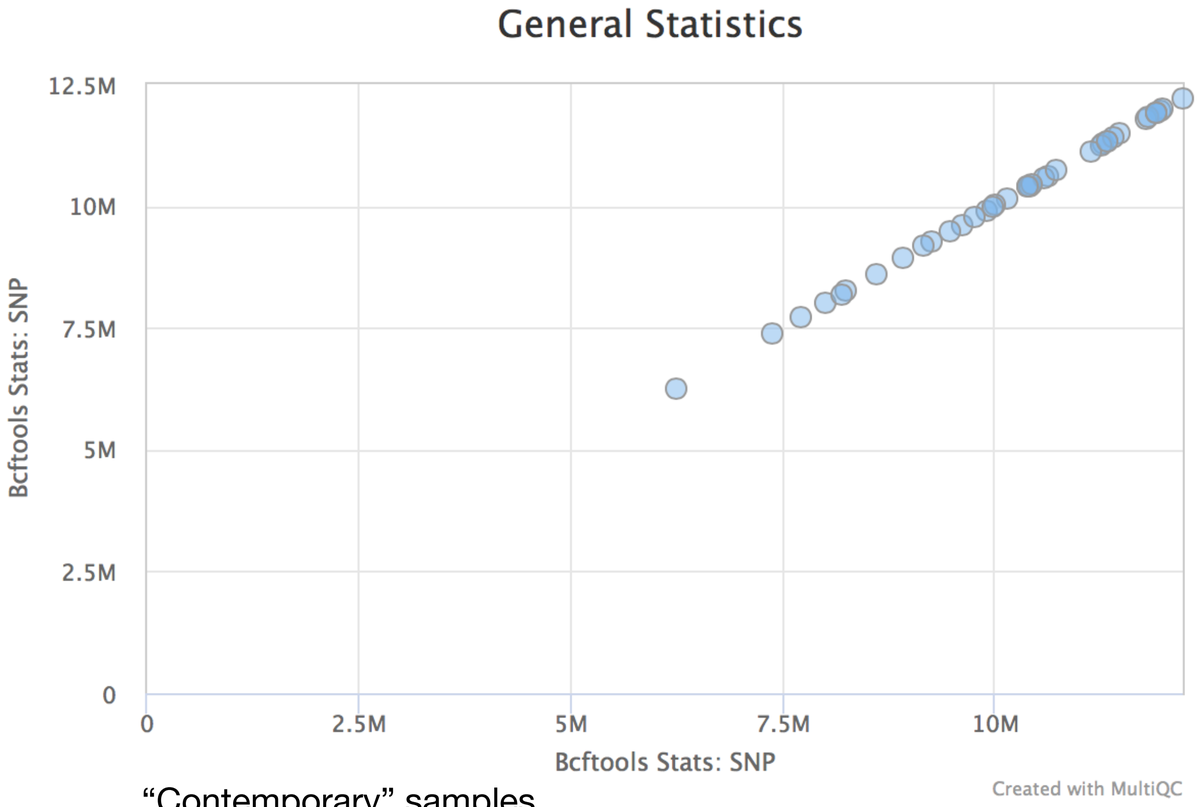
***Melitaea britomartis* population structure**

Variant calling QC

At the first step samples are called individually (freeBayes), sufficient number of SNPs recovered



“Historical” samples
input GC > 50%

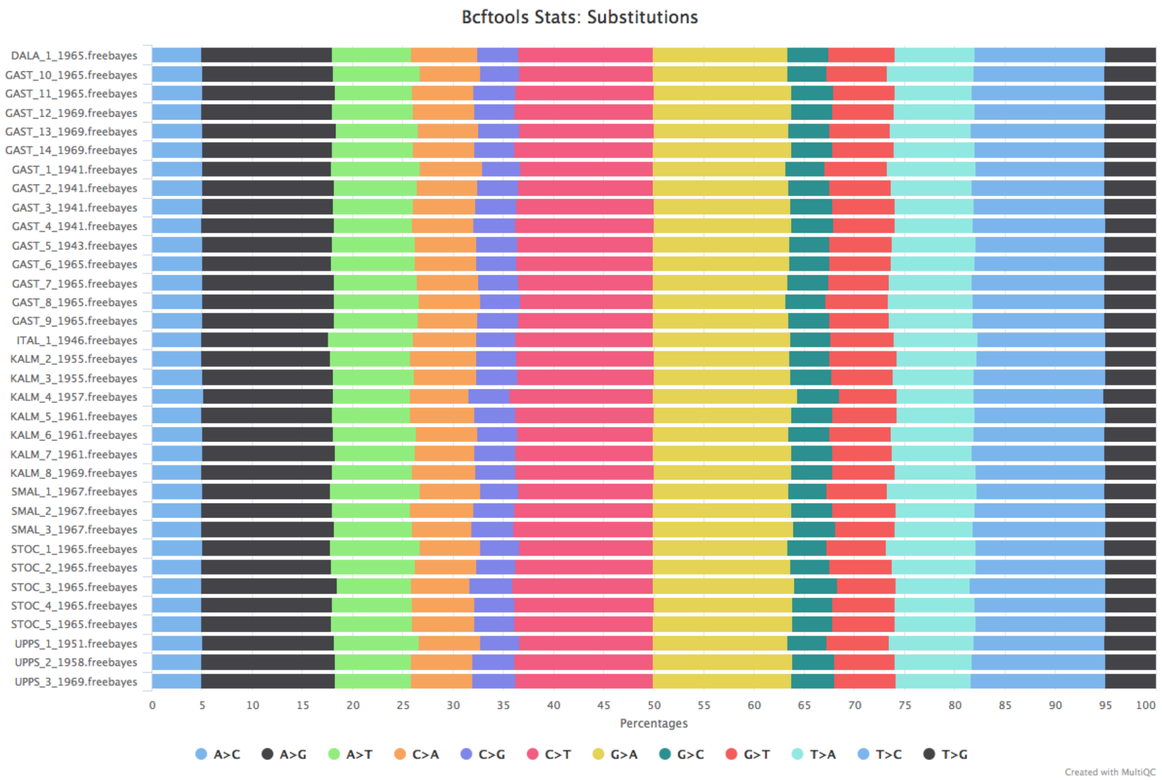


“Contemporary” samples
input GC < 50%

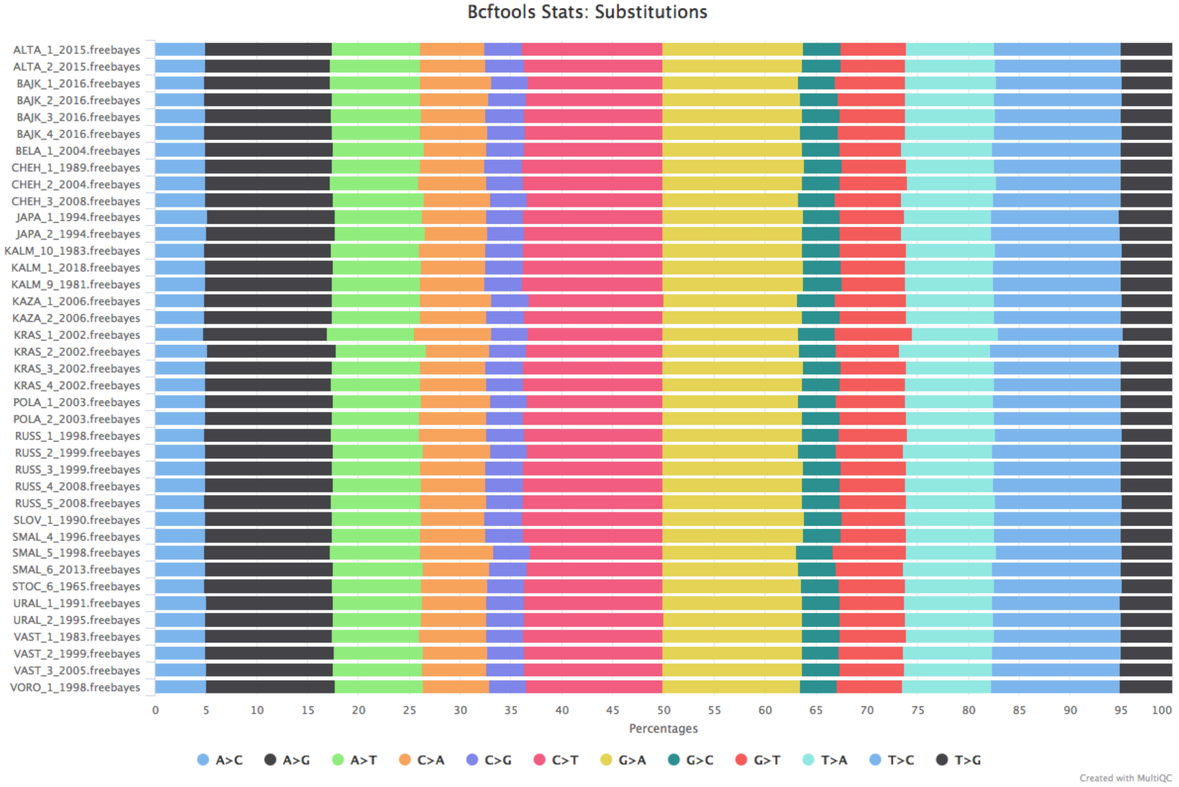
***Melitaea britomartis* population structure**

Variant calling QC

Conclusion: distribution of substitution types doesn't indicate strong signatures of deamination



"Historical" samples
input GC > 50%



"Contemporary" samples
input GC < 50%

***Melitaea britomartis* population structure**

Variant calling QC

Joined variant calling performed with settings from GenErode* pipeline

Stringient filtering applied:

Quality filter: MQ < 30

Missingness filter: variants are required to be present in at least 70 individuals

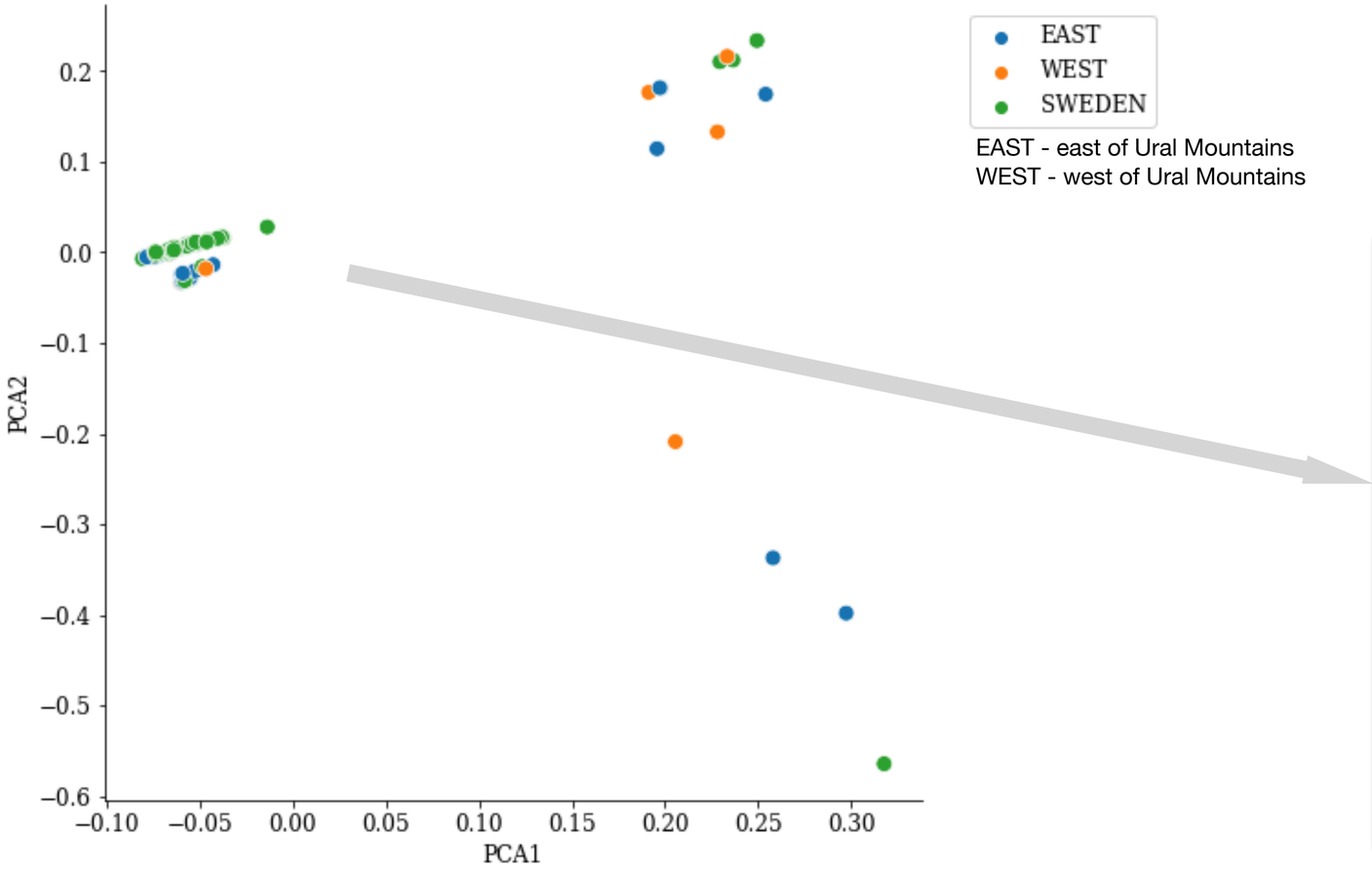
Total number of SNPs: **220,333**

*Kutschera, V.E., Kierczak, M., van der Valk, T. et al. GenErode: a bioinformatics pipeline to investigate genome erosion in endangered and extinct species. BMC Bioinformatics 23, 228 (2022). https://doi.org/10.1186/s12859-022-04757-0

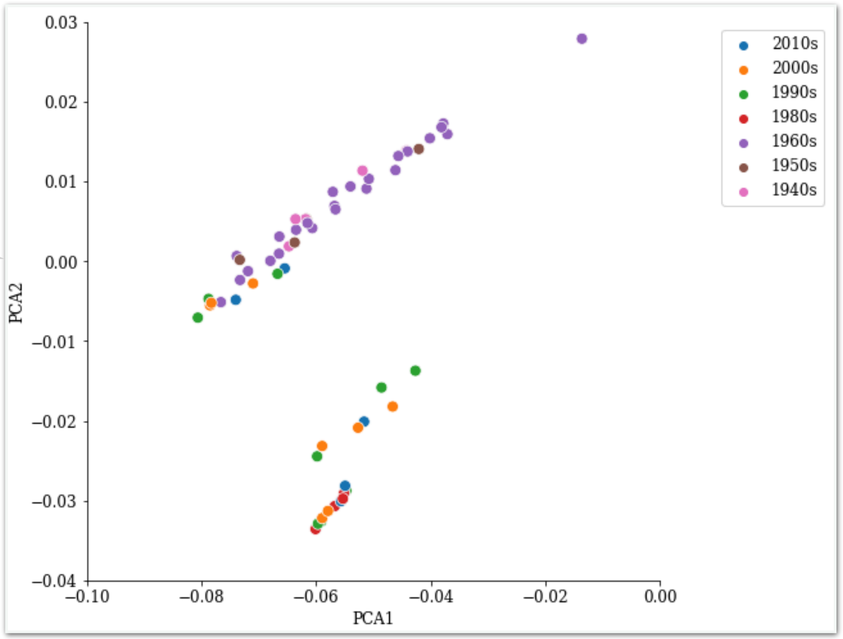
***Melitaea britomartis* population structure**

Population structure analysis

Basic principle component analysis (PCA) revealed unexpected grouping of samples based on geography and sampling year



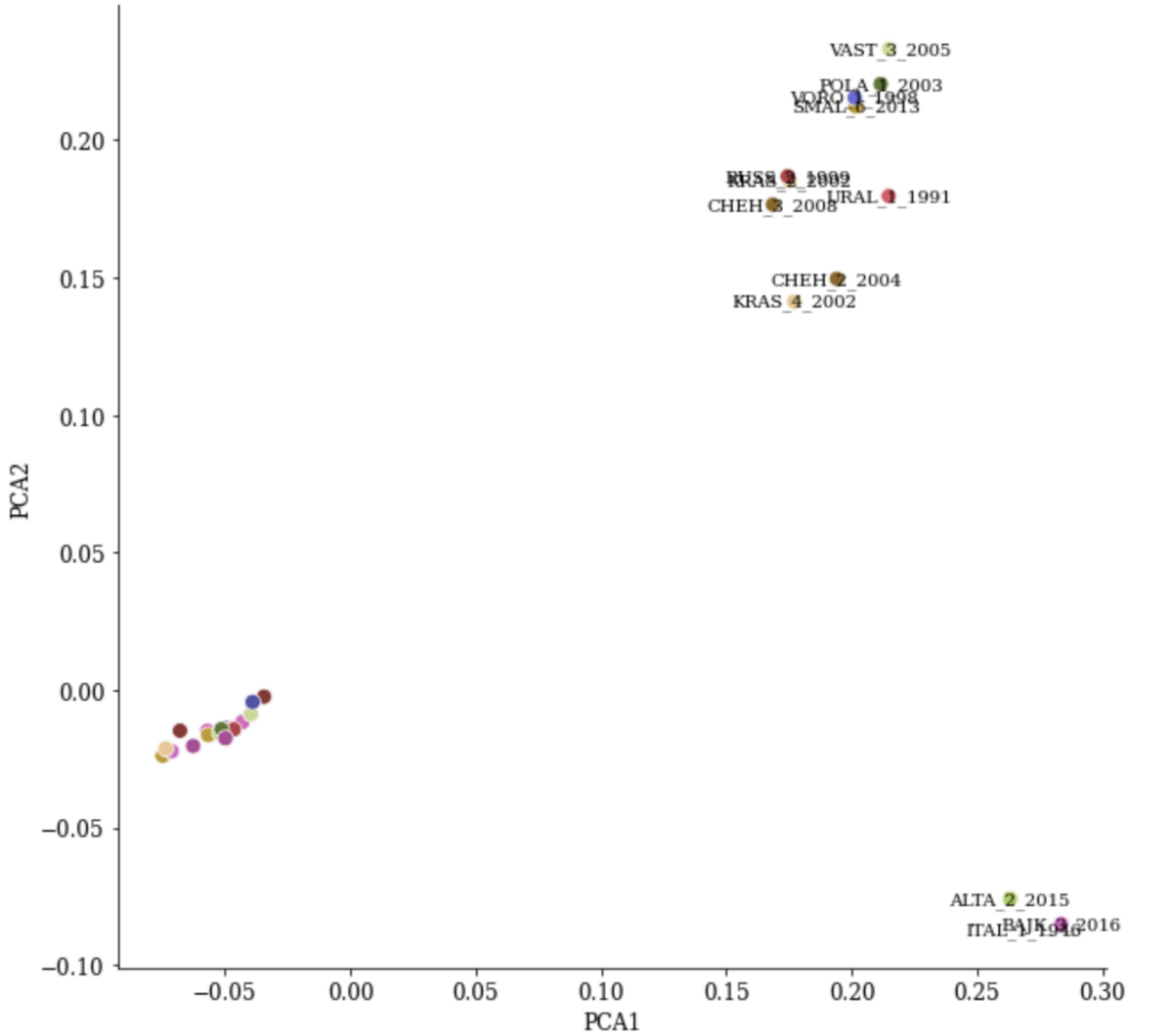
EAST - east of Ural Mountains
WEST - west of Ural Mountains



***Melitaea britomartis* population structure**

Population structure analysis

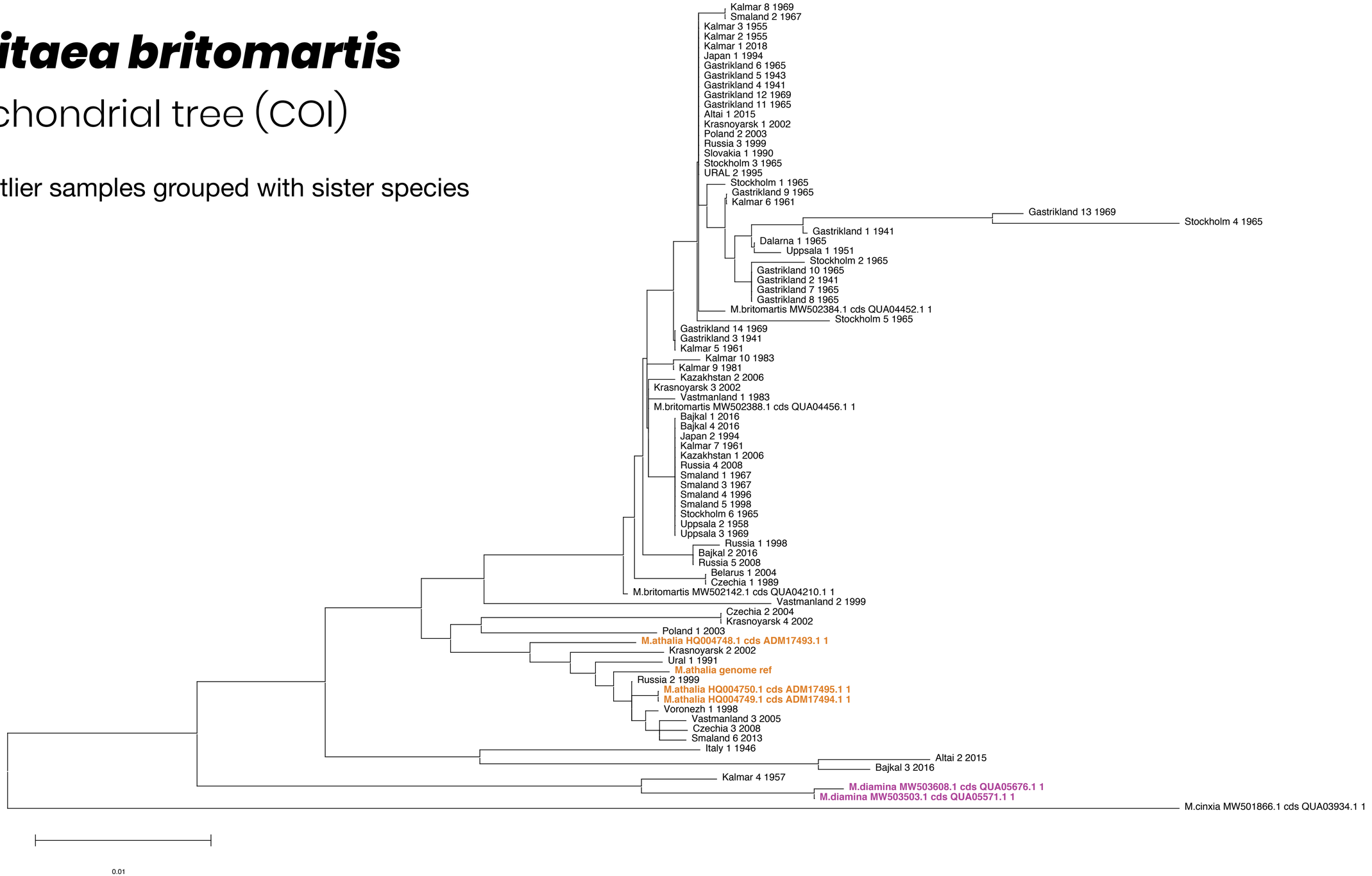
Excluding mtDNA, sex-chromosomes and lower quality (historical) samples did not change overall structure of the plot



Melitaea britomartis

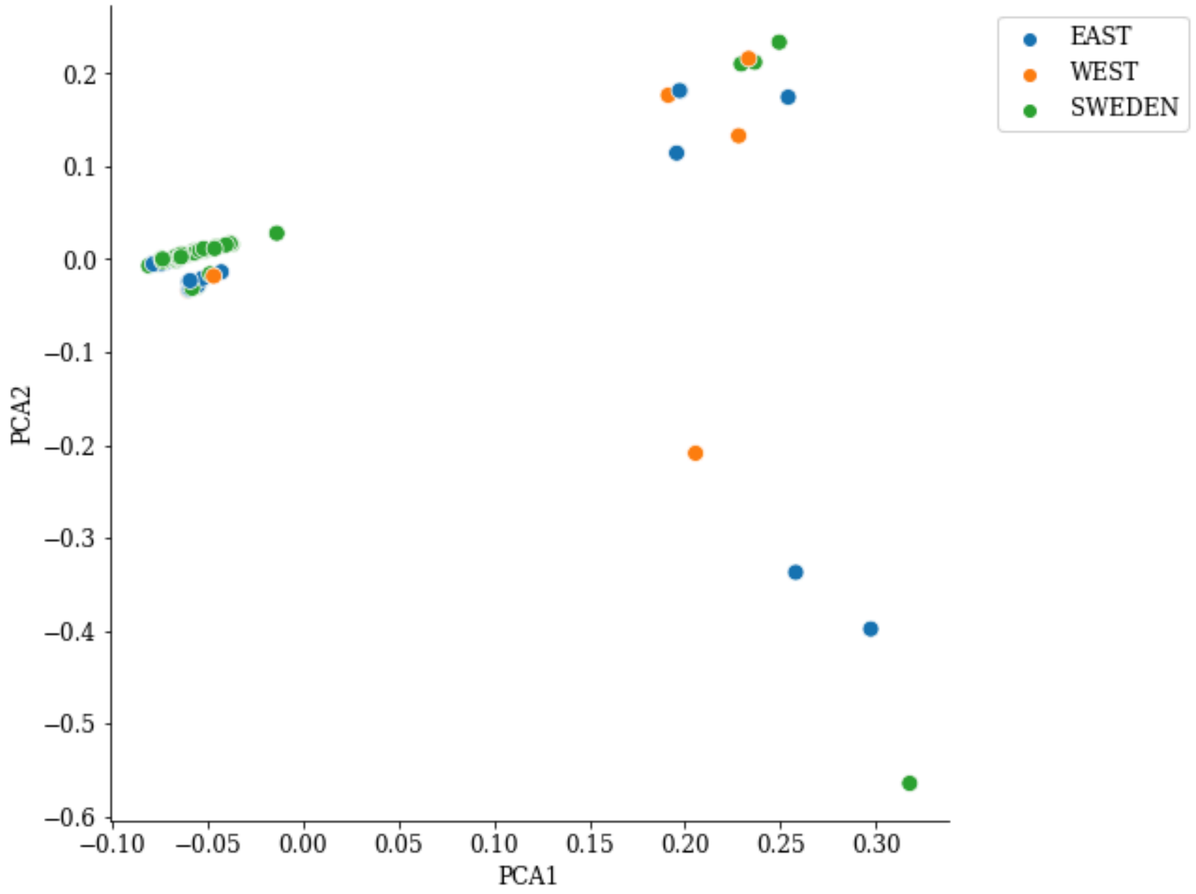
Mitochondrial tree (COI)

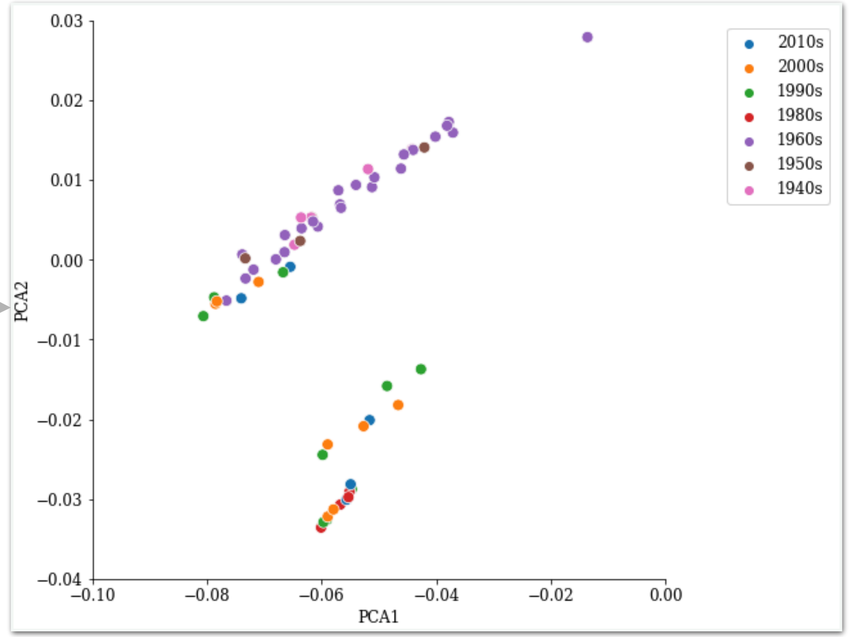
PCA outlier samples grouped with sister species

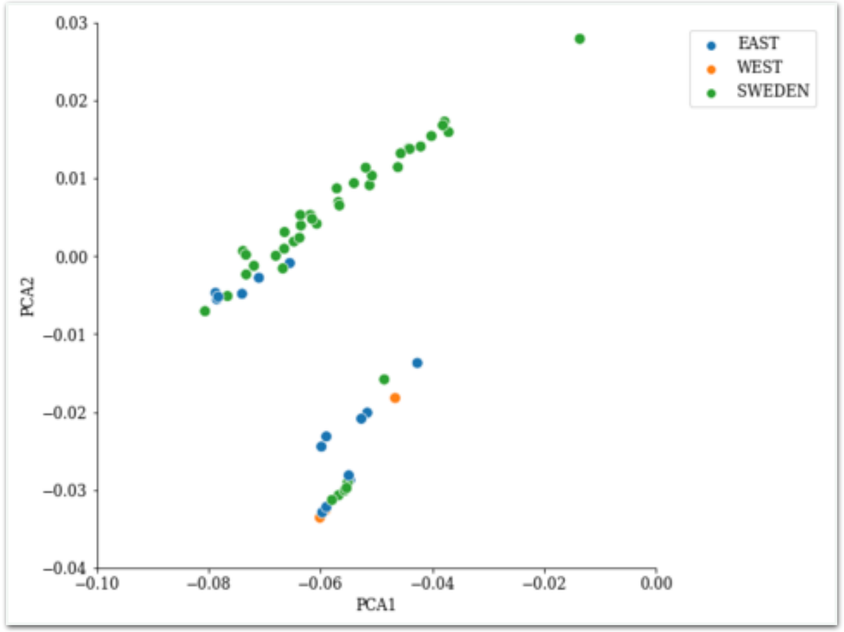


SNP calling #3:

(GenErode + custom settings, whole genome, all individuals)

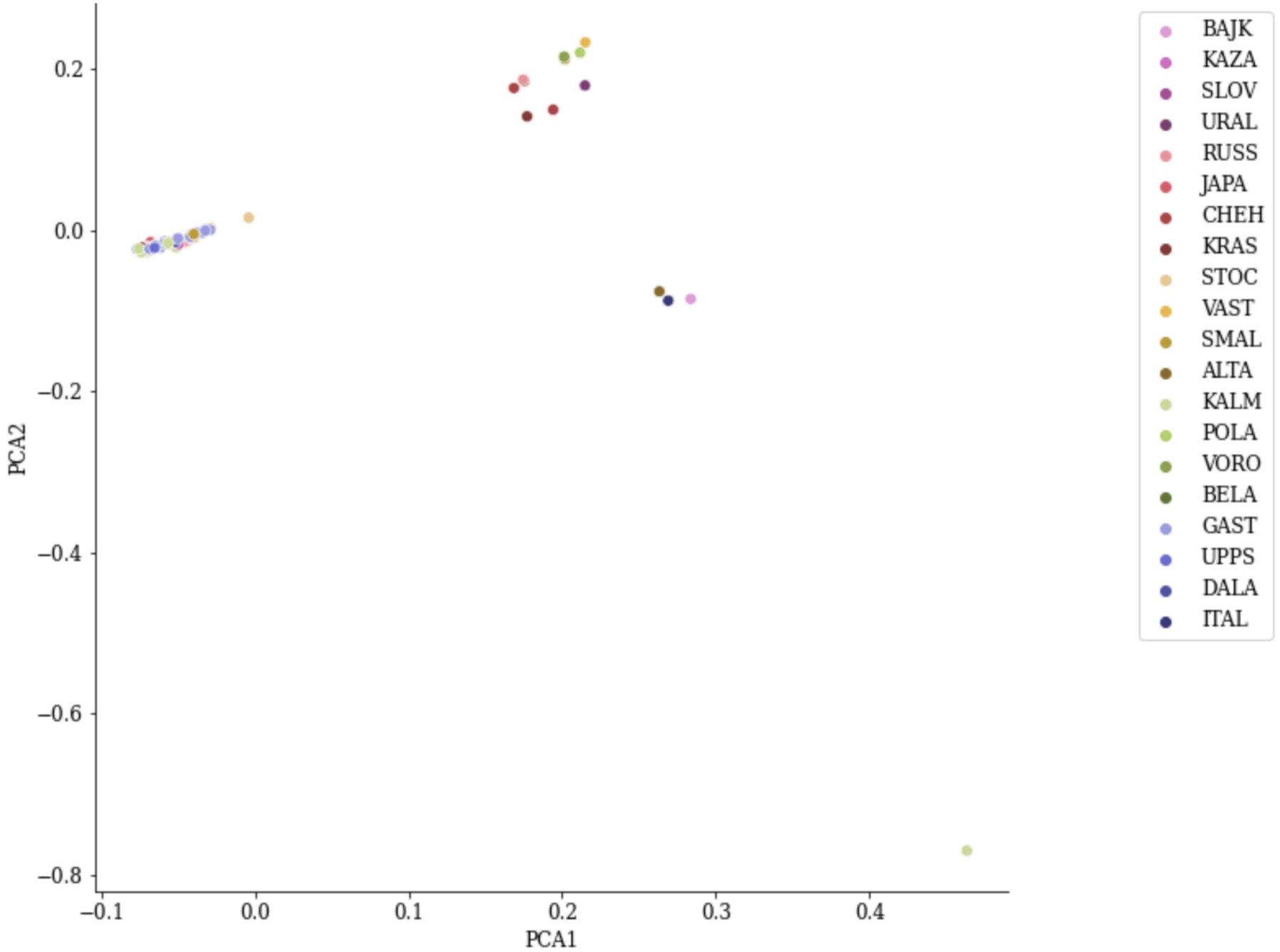


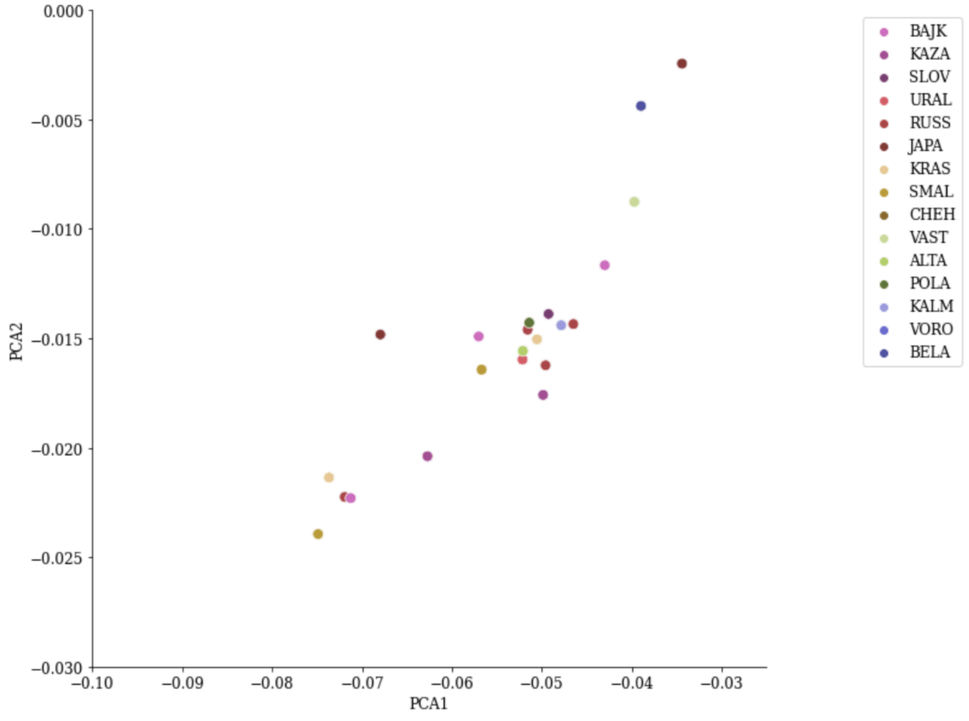


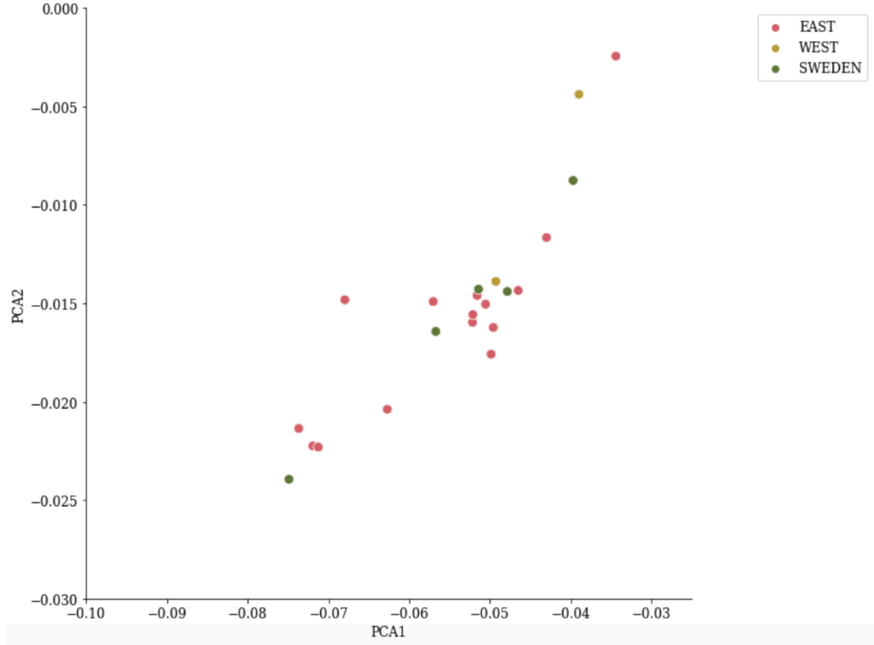


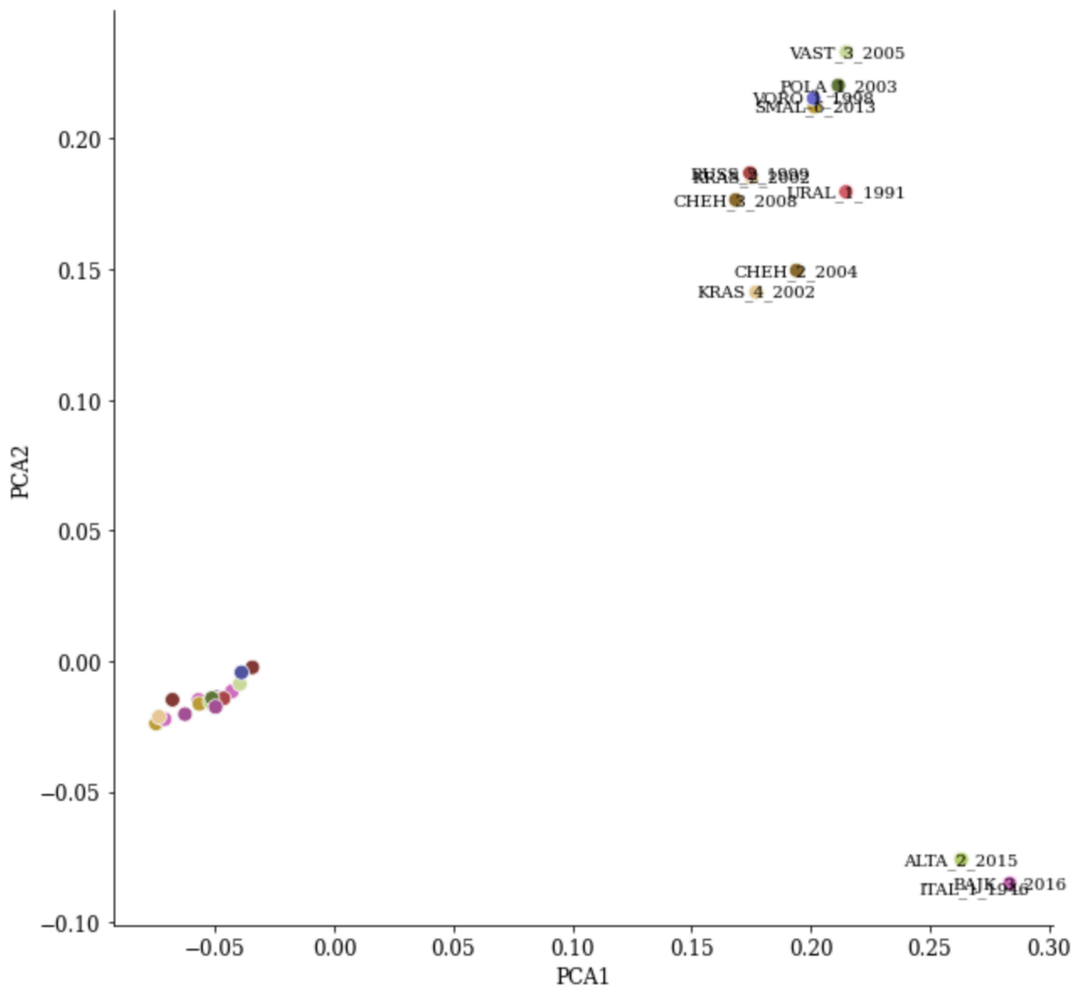
PCA:

- contemporary
- missing max 3
- autosomes only




BAJK
KAZA
SLOV
URAL
RUSS
JAPA
KRAS
SMAL
CHEH
VAST
ALTA
POLA
KALM
VORO
BELA
PCA1
PCA2


EAST
WEST
SWEDEN
PCA1
PCA2


VAST_3_2005
POLA_1_2003
VORO_1_1998
SMAL_5_2013
URAL_1_1991
CHEH_3_2008
CHEH_2_2004
KRAS_4_2002
ALTA_2_2015
PCA1
PCA2

Mapping and quality control

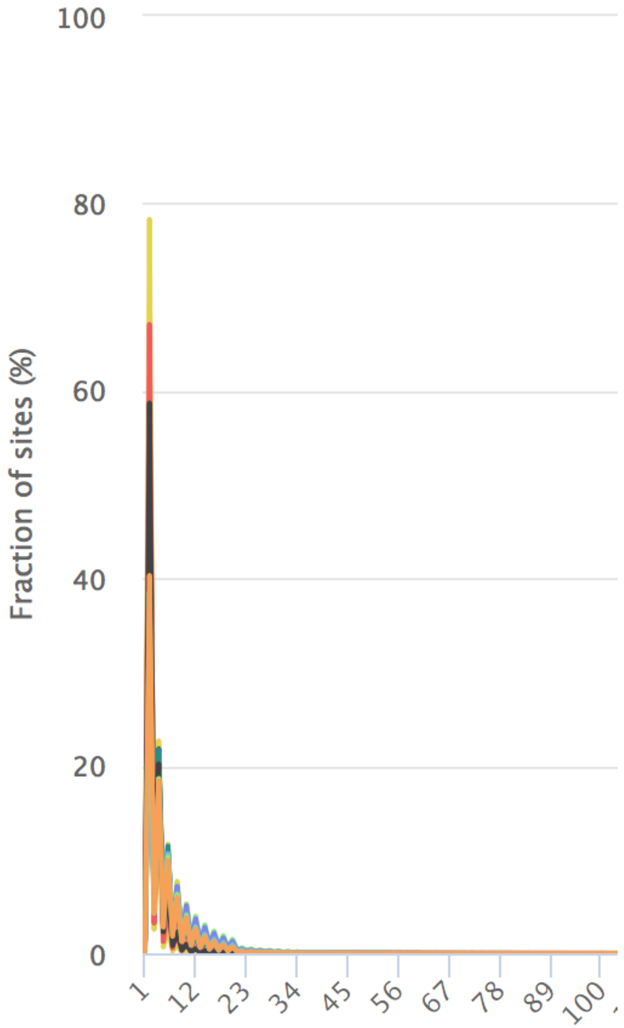
Sarek is used to map and call two sets separately

“Historical” samples
input GC > 50%

“Contemporary” samples
input GC < 50%

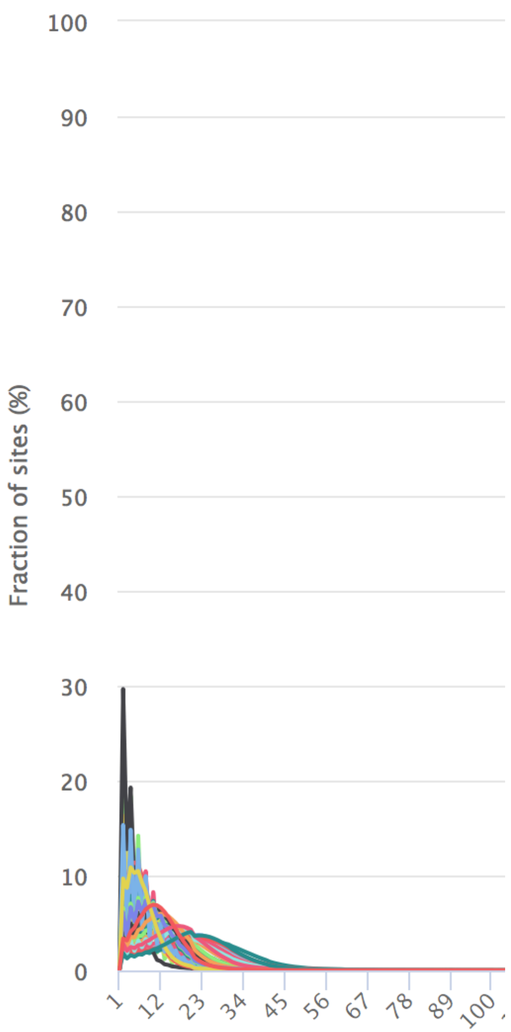
Summary of SNP calling with freebayes

Conclusion: low quality/depth of coverage for historical samples, careful filtering needed



“Historical” samples
input GC > 50%

Variant depth



“Contemporary” samples
input GC < 50%

Variant depth

SNP filtering attempt #1: stringent filtering of merged file

Conclusion: after applying stringent quality filters (below) number of SNPs reduced dramatically

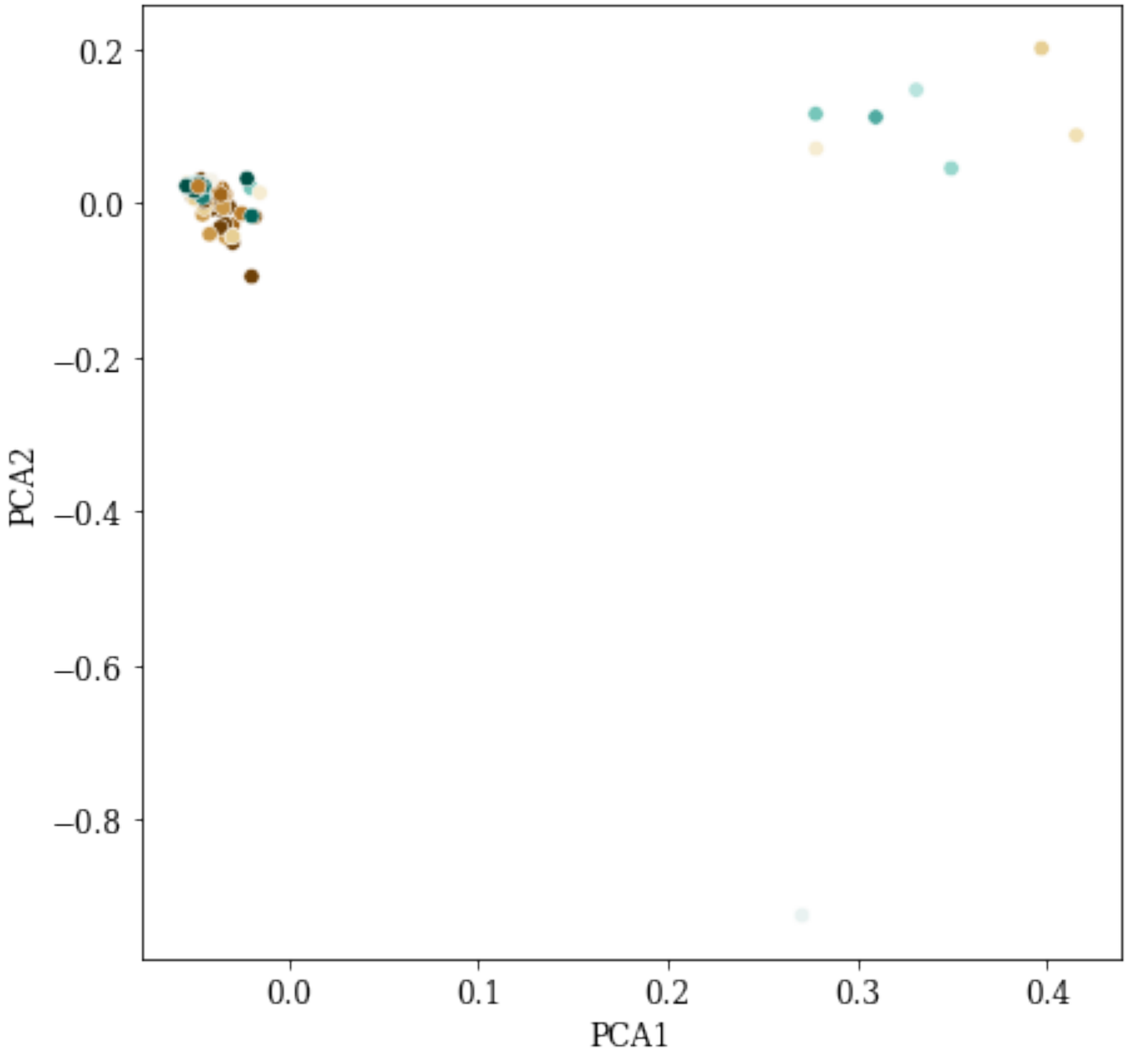
number of samples:	73
number of records:	2203
number of no-ALTs:	0
number of SNPs:	2203
number of MNPs:	0
number of indels:	0
number of others:	0
number of multiallelic sites:	334

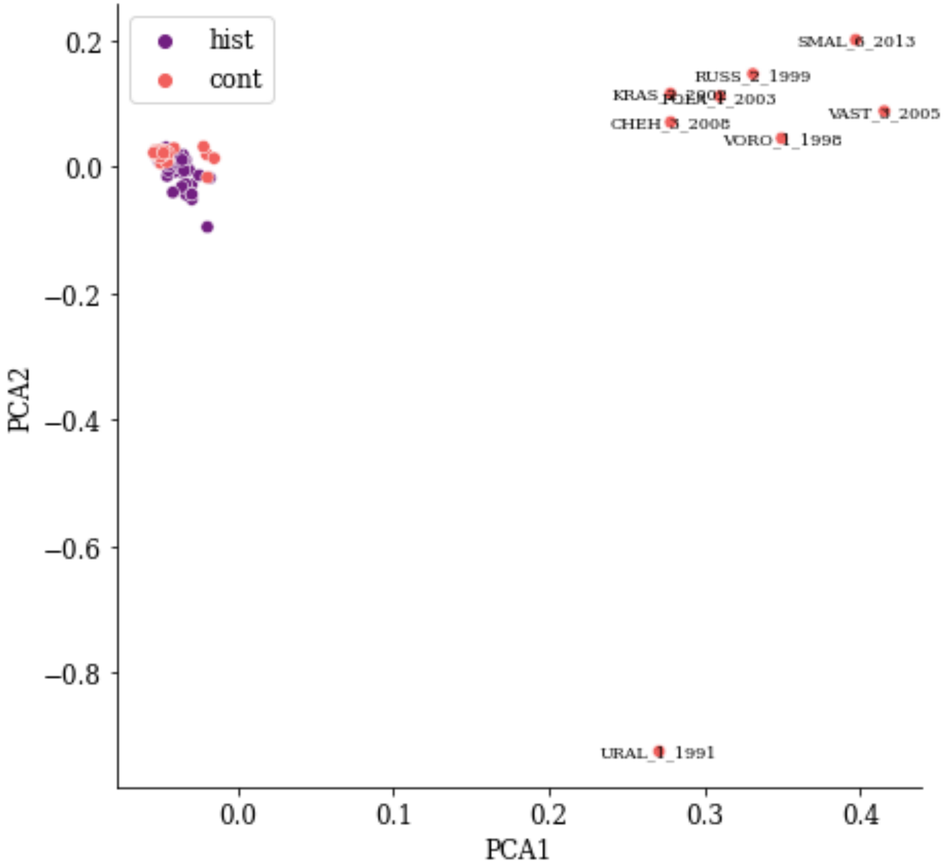
--remove-indels
--max-missing-count 10
--minQ 30
--min-meanDP 10
--max-meanDP 40

LD pruning applied

SNP filtering attempt #1: stringent filtering of merged file

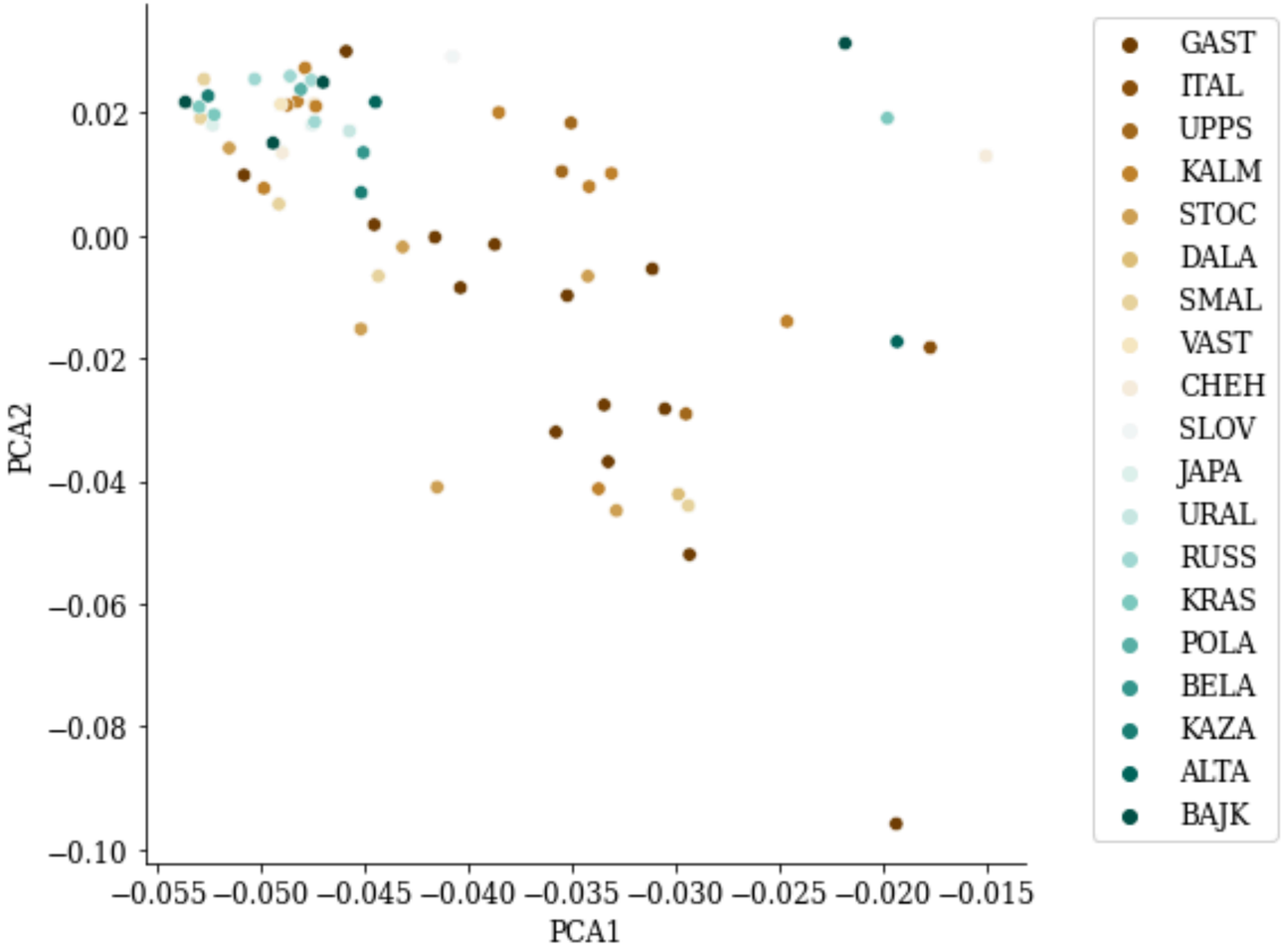
Conclusion: samples group together in unusual pattern, no segregation based on geography, however no bias towards historical samples is observed

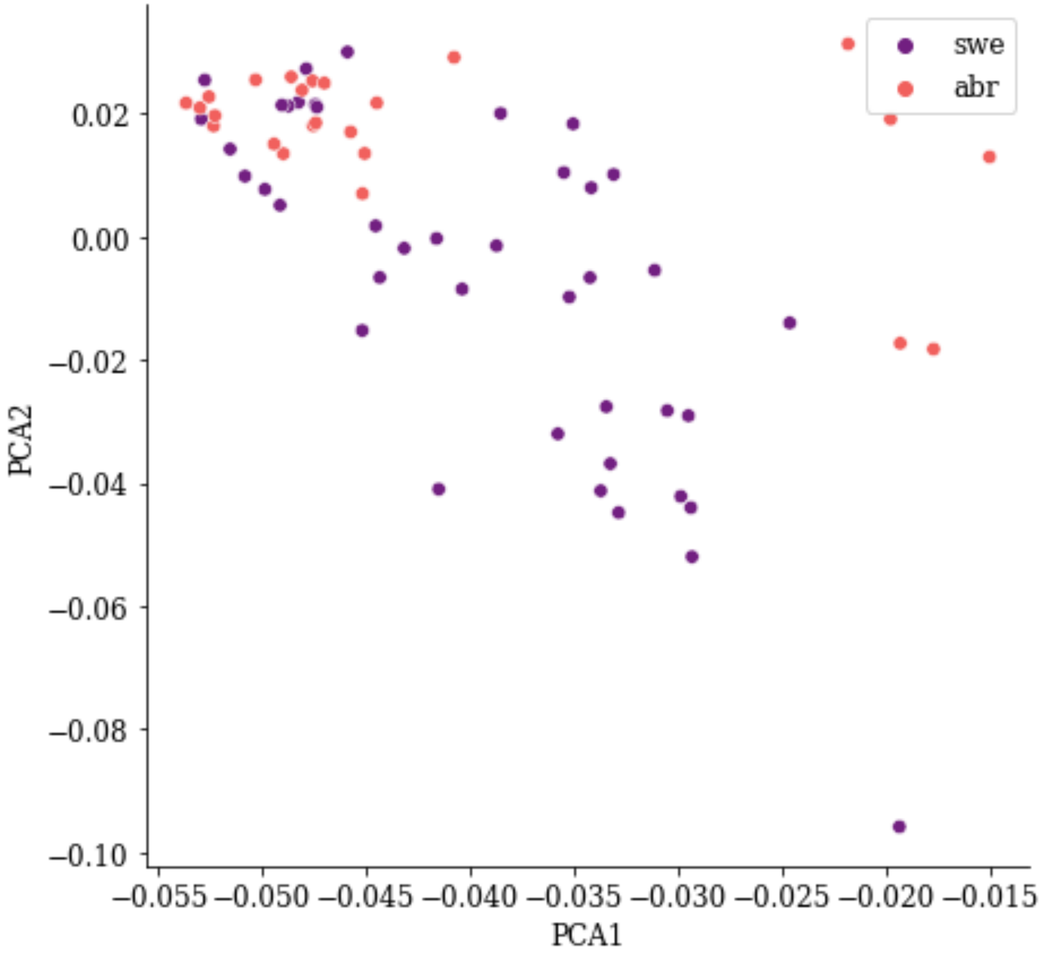




SNP filtering attempt #1: stringent filtering of merged file

Conclusion: closer look at the main cluster confirms conclusions from the previous slide





SNP filtering attempt #2: relaxed filtering of merged and subsampled files

Number of SNPs is low, biological signal might be uninterpretable

Filtring approaches:

1. Removing individuals with low coverage from analysis (contemporary set, chromosome 1, --max-missing-count 10)

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      39
SN      0       number of records:      4071
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 4071
SN      0       number of MNPs: 0
SN      0       number of indels:       0
SN      0       number of others:       0
SN      0       number of multiallelic sites:   611
```

--max-missing-count 10

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      39
SN      0       number of records:      19175
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 19175
SN      0       number of MNPs: 0
SN      0       number of indels:       0
SN      0       number of others:       0
SN      0       number of multiallelic sites:   3137
```

--max-missing-count 30

Conclusion: number of retained SNPs is alarmingly low for high coverage/quality data, technical bias is suspected. Freebayes produced individual vcf files (within sarek) which are merged, possible joint genotyping is needed

SNP calling mpileup, reduced sample set

Effect of calling approach is tested on a reduced sample set: 10 individuals with the best coverage, considering only chromosome 1 (~10Mb)

Filtering (--max-missing-count 3 --min-alleles 2 --max-alleles 2 --minQ 30 + LD prune) produced more realistic results:

```
# SN  [2]id [3]key  [4]value
SN  0 number of samples:  10
SN  0 number of records:  665439
SN  0 number of no-ALTs:  0
SN  0 number of SNPs: 665439
SN  0 number of MNPs: 0
SN  0 number of indels: 0
SN  0 number of others: 0
SN  0 number of multiallelic sites: 0
```

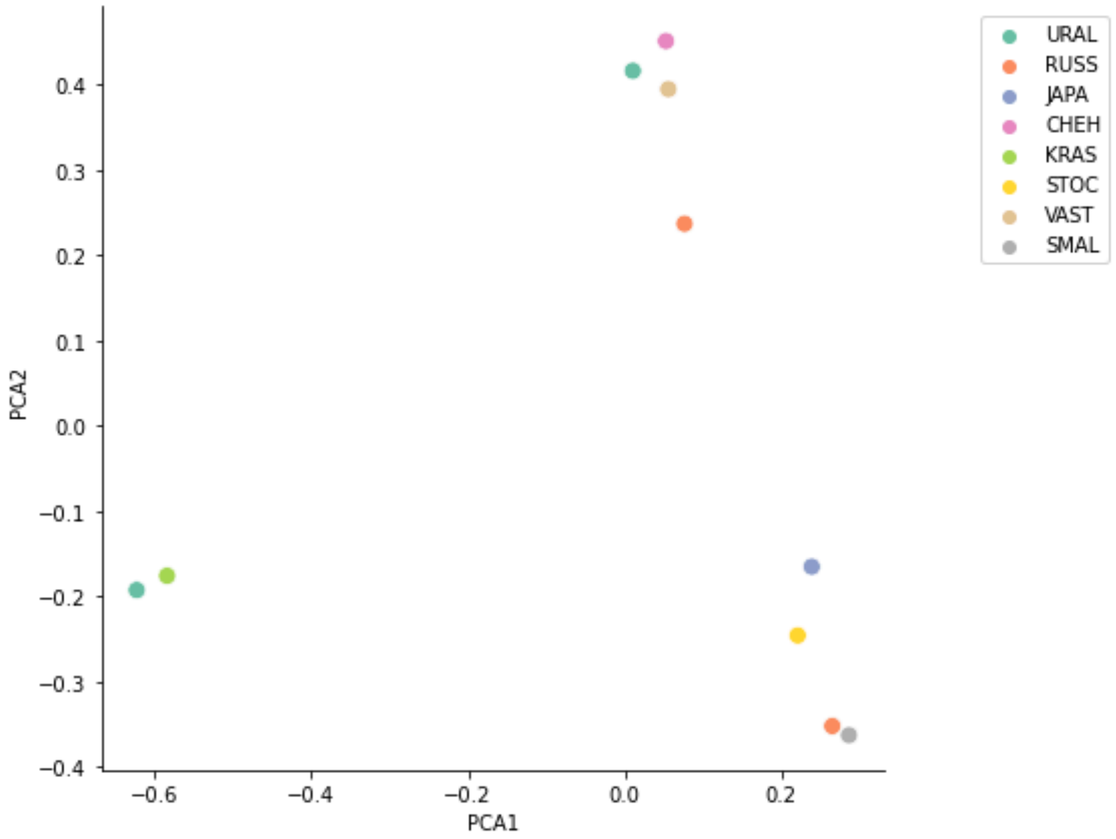
mpileup consensus caller, identical to GenErode

```
# SN  [2]id [3]key  [4]value
SN  0 number of samples:  10
SN  0 number of records:  521907
SN  0 number of no-ALTs:  0
SN  0 number of SNPs: 521907
SN  0 number of MNPs: 0
SN  0 number of indels: 0
SN  0 number of others: 0
SN  0 number of multiallelic sites: 0
```

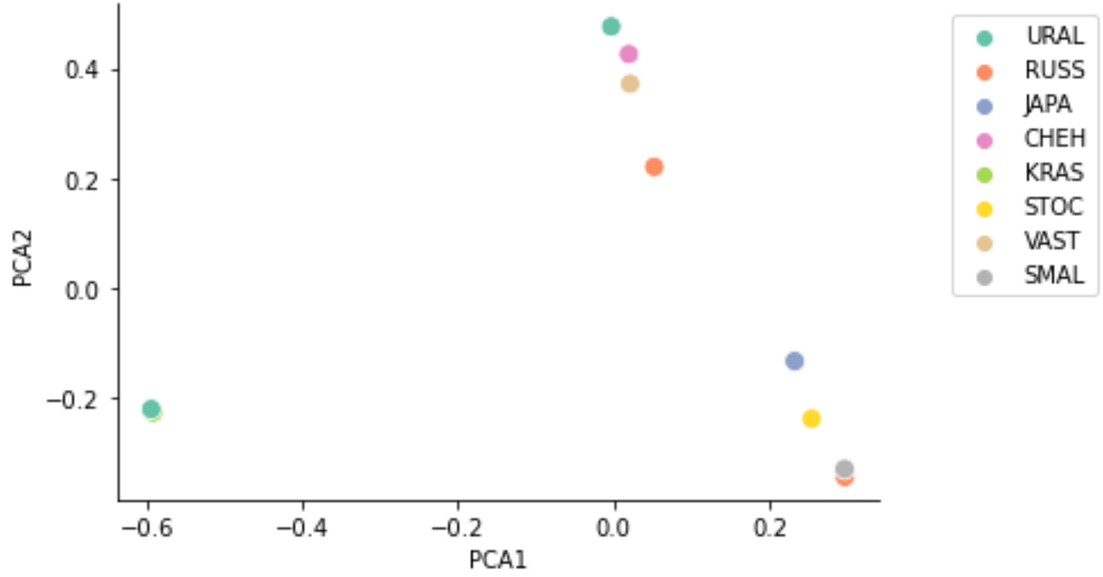
mpileup multiallelic caller, other settings from GenErode

SNP calling mpileup (GenErode settings, reduced sample set)

Conclusion: grouping is still unusual



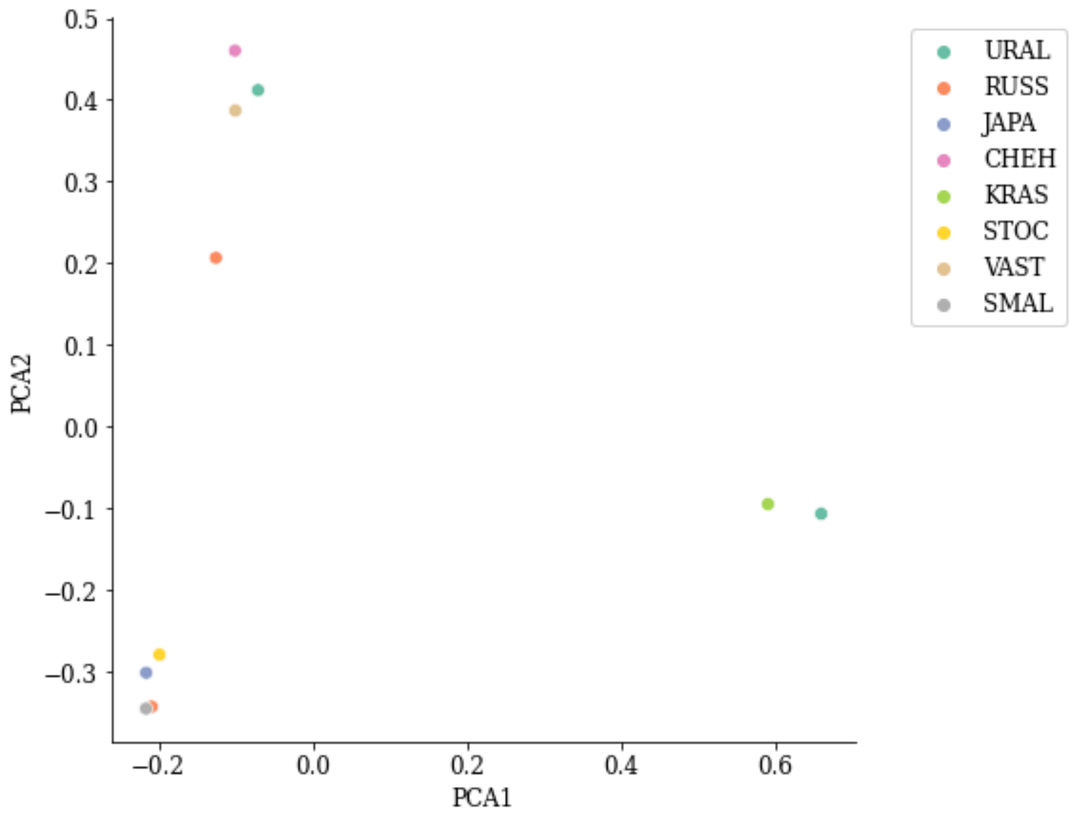
mpileup consensus caller, identical to GenErode

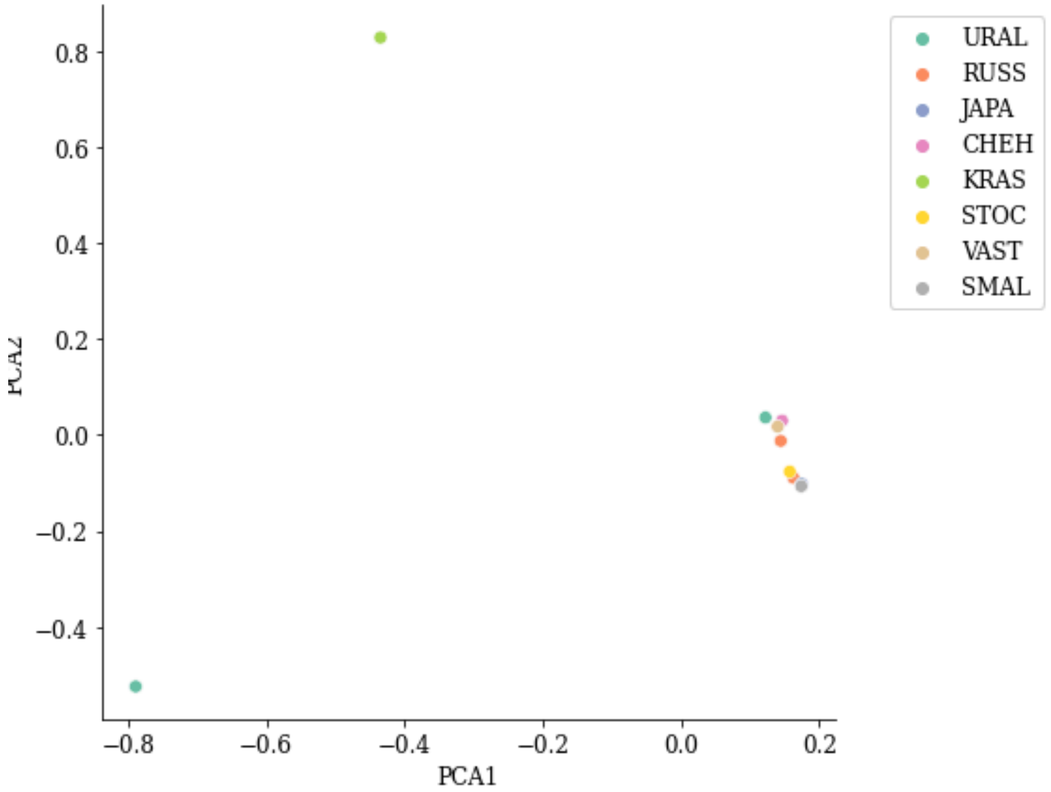


mpileup multiallelic caller, other settings from GenErode

SNP calling mpileup #2 (GenErode settings, multiple chromosomes)

Conclusion: grouping is still unusual



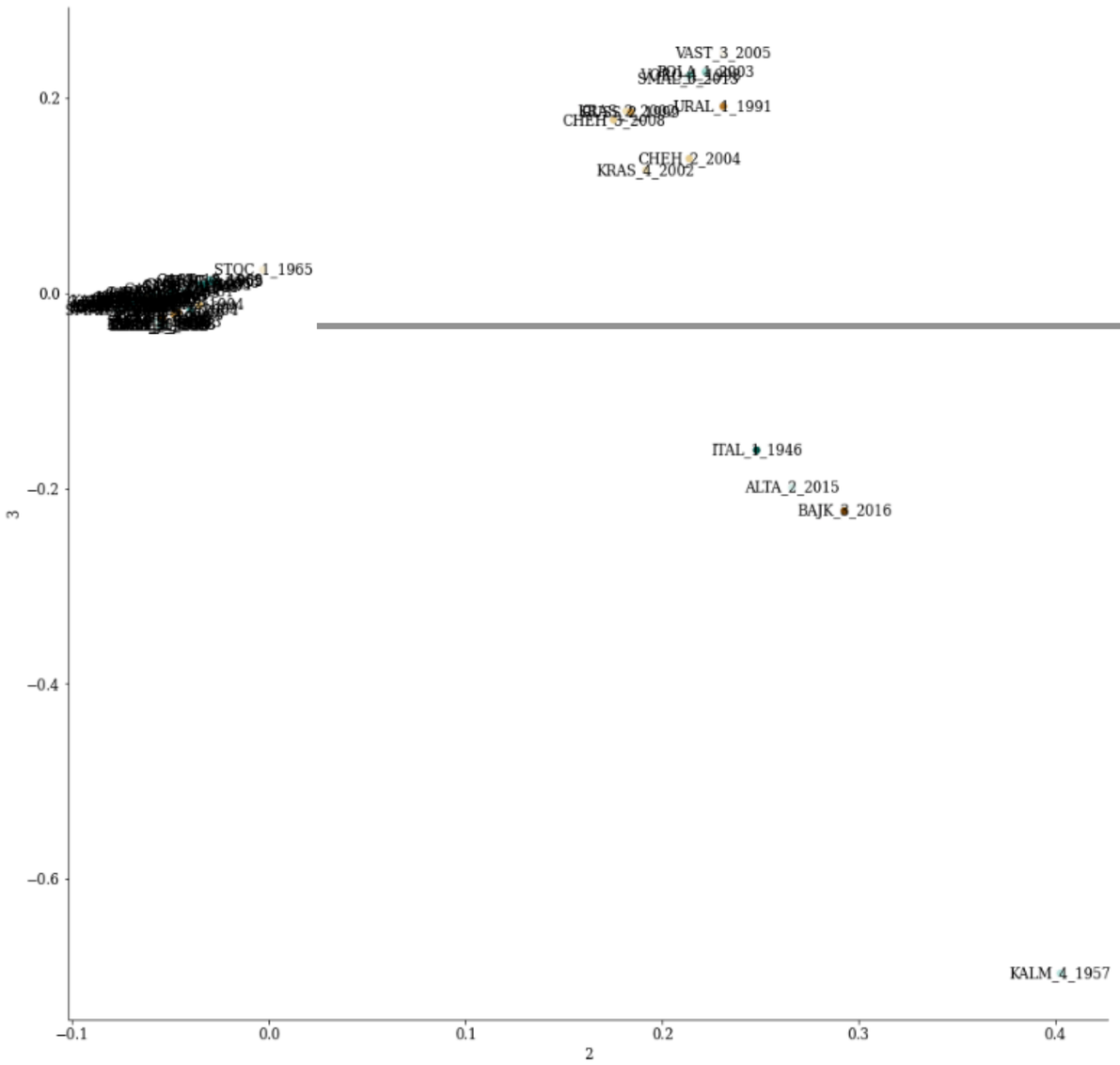


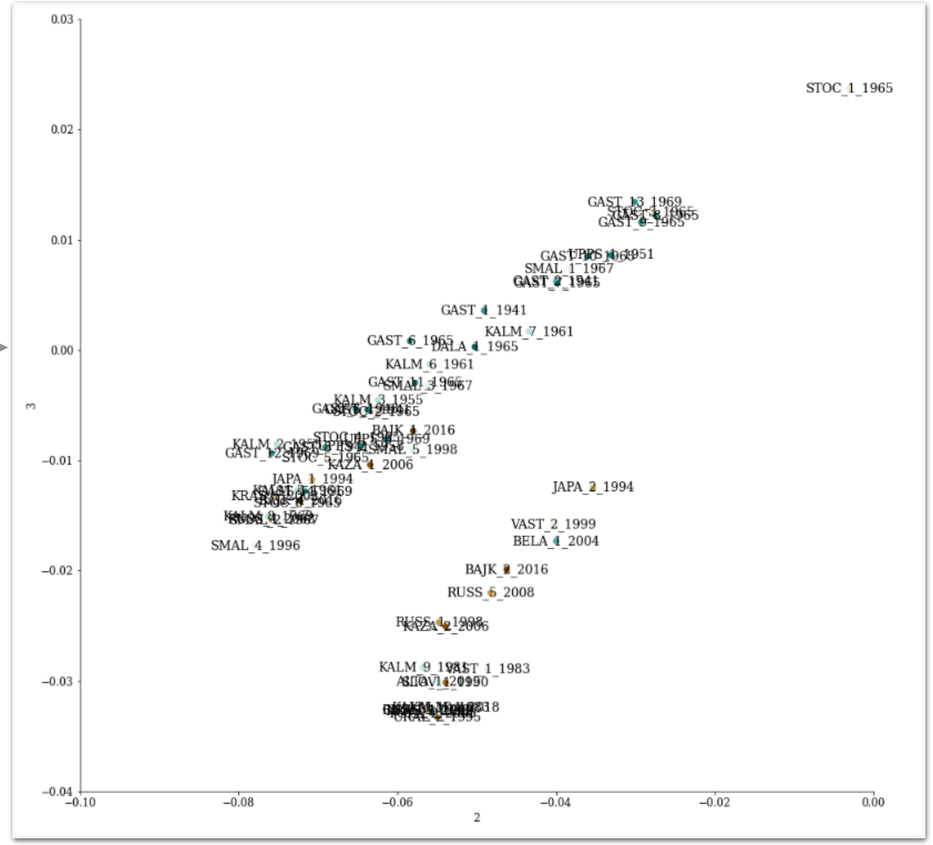
mpileup multiallelic caller, other settings from GenErode

Calling strategy accepted!

SNP calling #3:

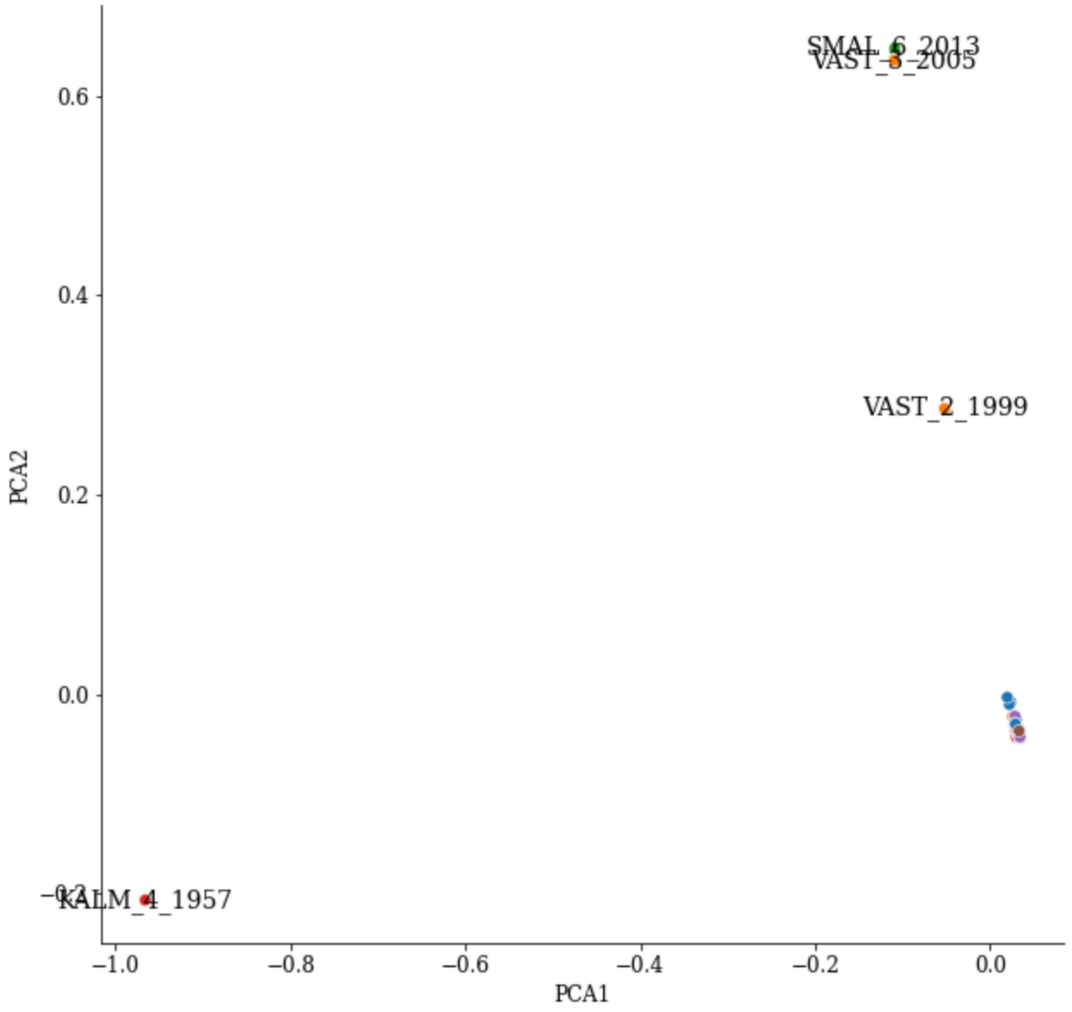
(GenErode + custom settings, whole genome, all individuals)



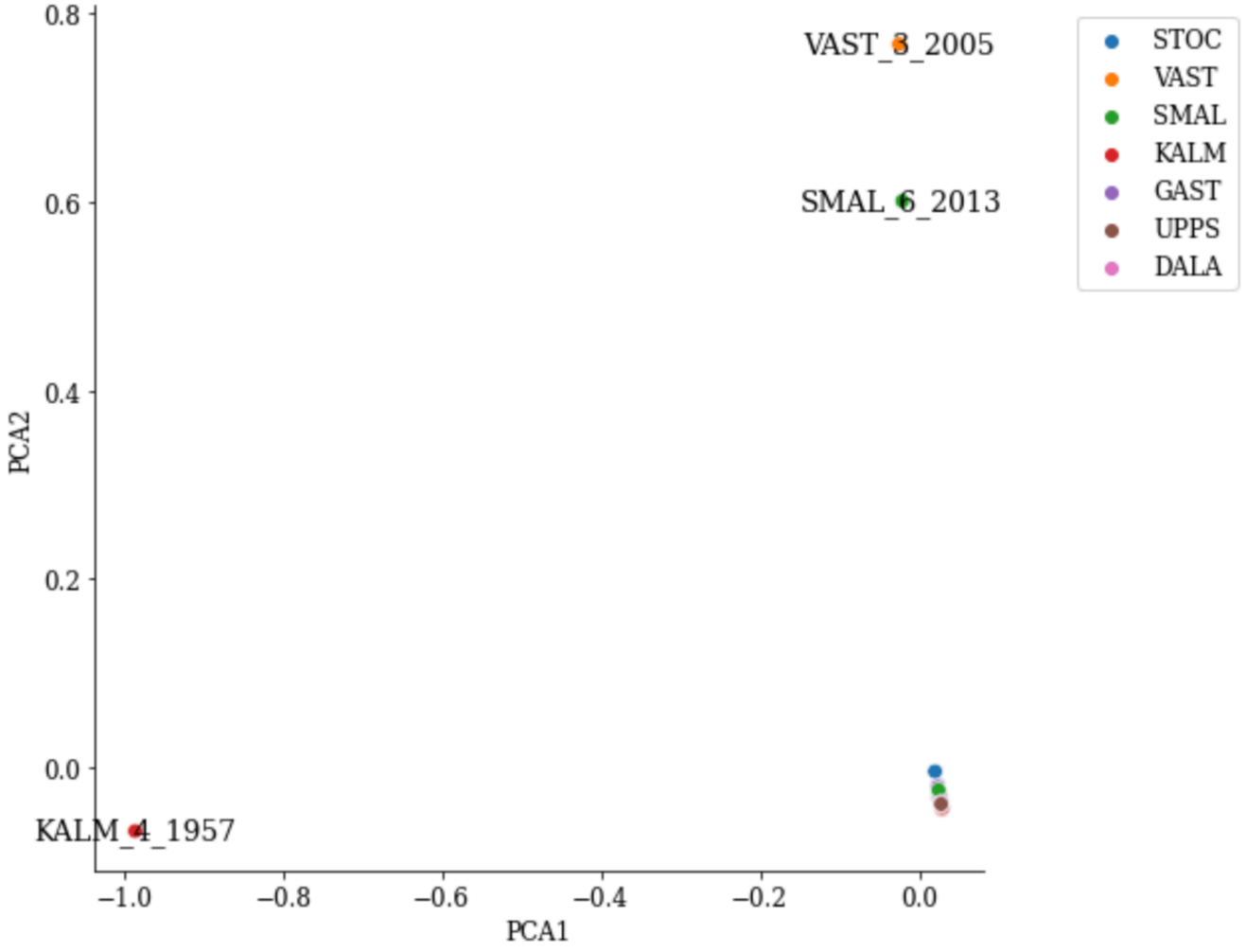


Testing PCA bias:

(PCA with only Swedish samples)



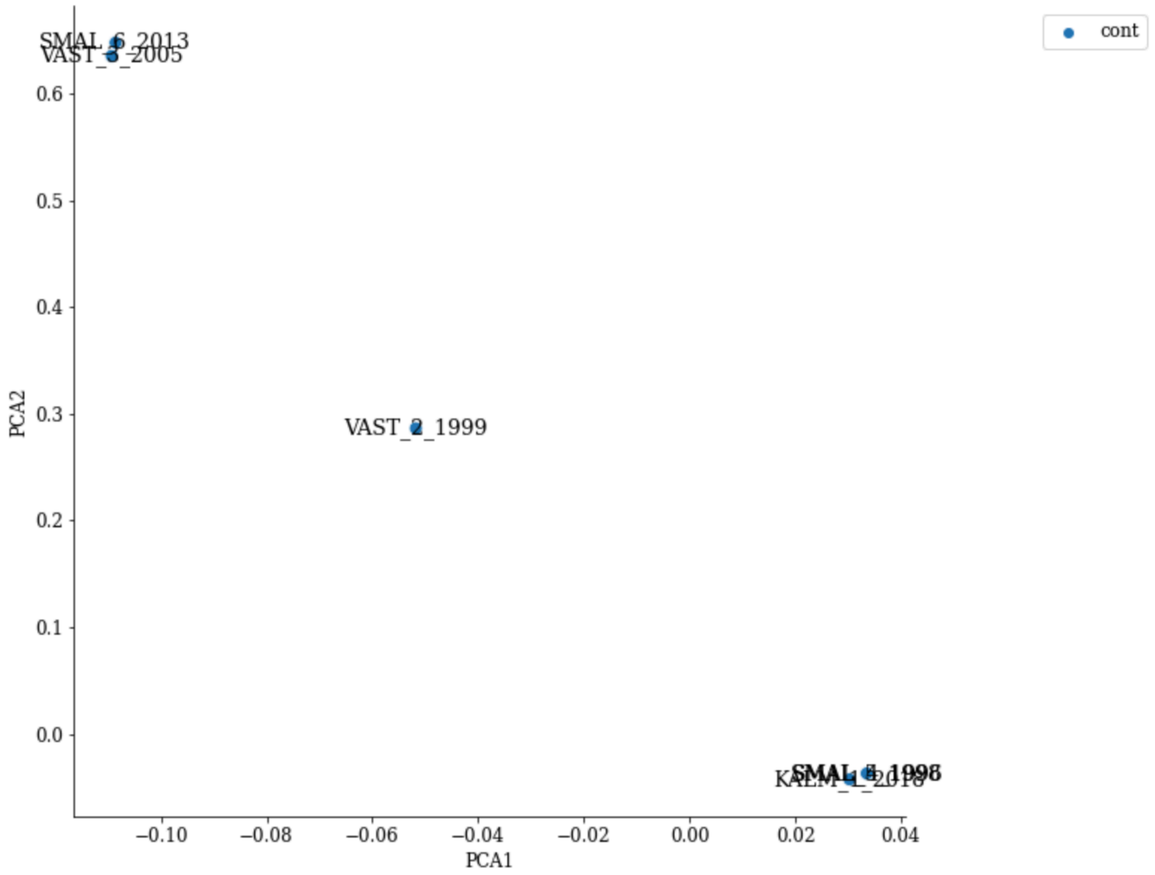
mtDNA

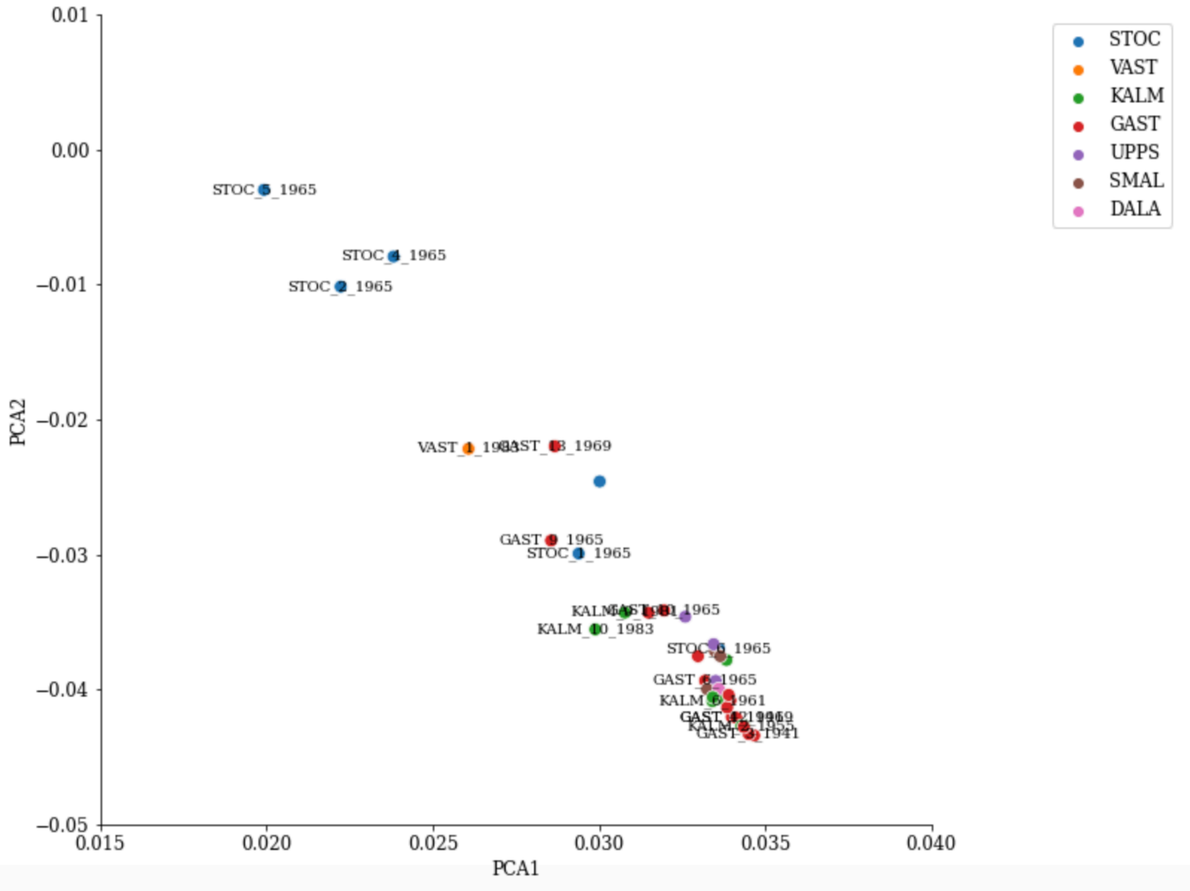


Autosomes

Testing PCA bias:

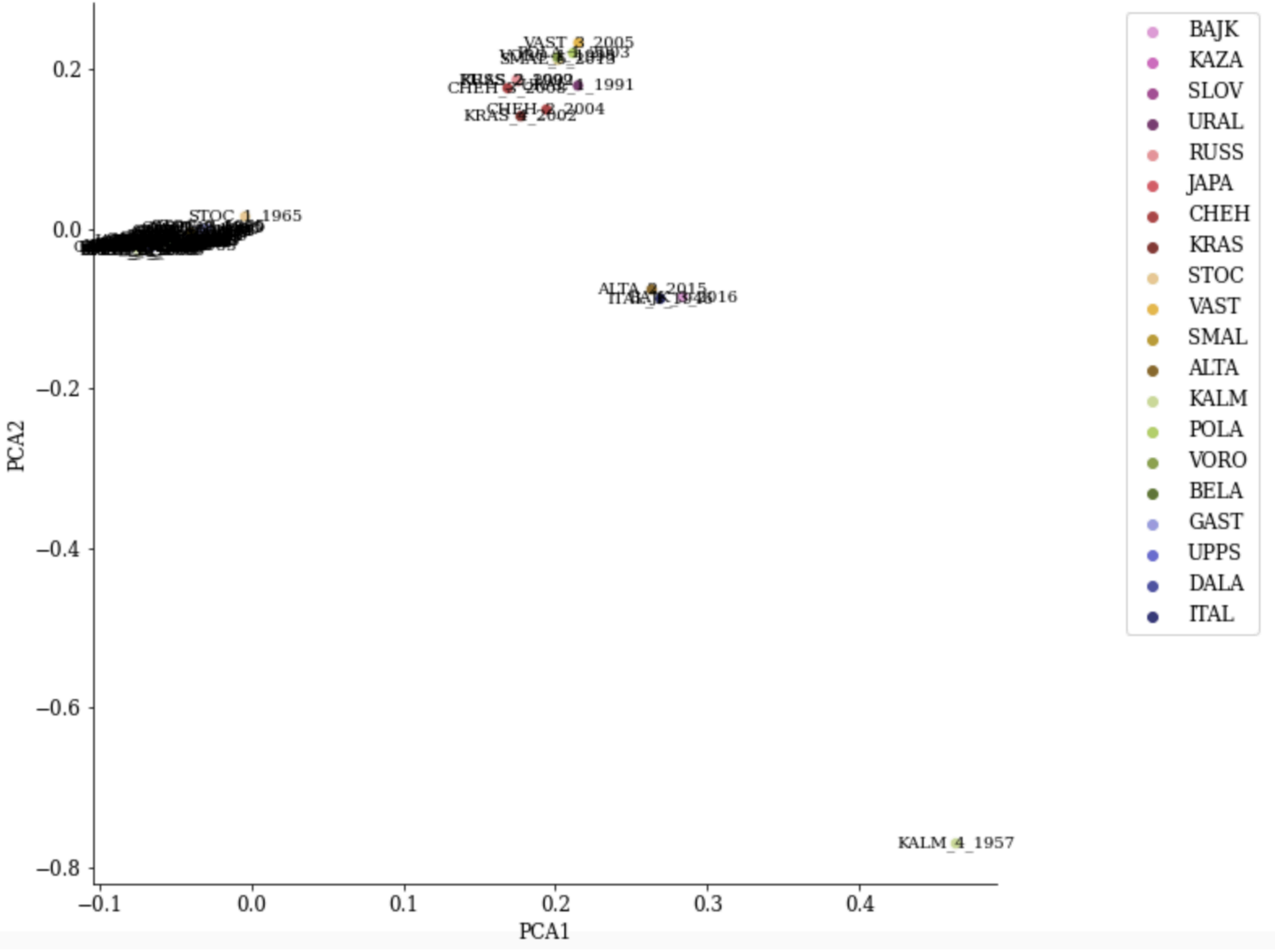
(contemporary and historical are plotted separately, mtDNA)

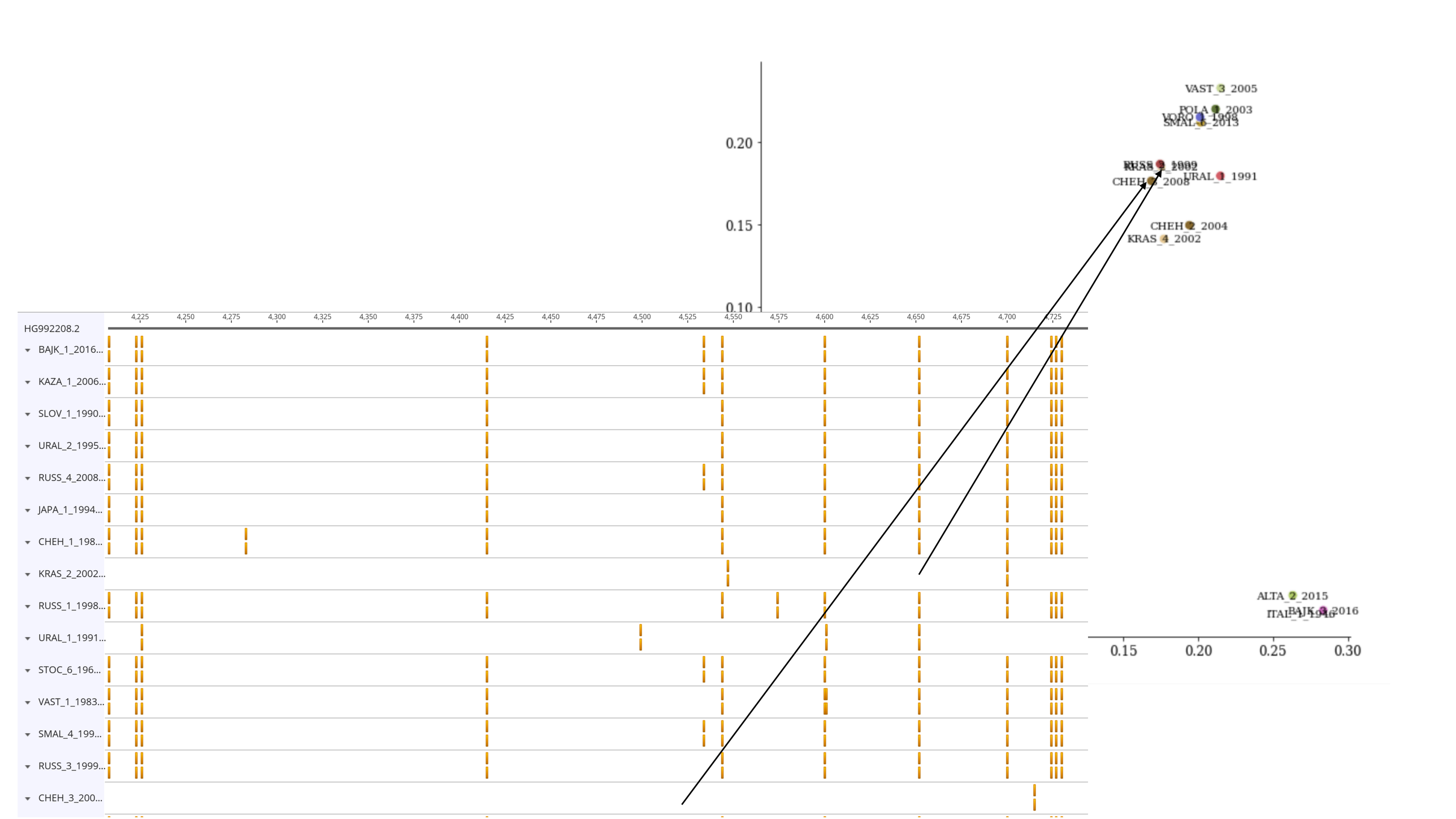




PCA:

- contemporary and historical
- missing max 3
- autosomes only
- ~100.000 SNP




0.20
0.15
0.10
VAST_3_2005
POLA_1_2003
SMAL_5_2013
KRAS_3_2002
URAL_1_1991
CHEH_3_2008
CHEH_2_2004
KRAS_4_2002
ALTA_2_2015
BAJK_3_2016
0.15
0.20
0.25
0.30
HG992208.2
4,225
4,250
4,275
4,300
4,325
4,350
4,375
4,400
4,425
4,450
4,475
4,500
4,525
4,550
4,575
4,600
4,625
4,650
4,675
4,700
4,725
BAJK_1_2016...
KAZA_1_2006...
SLOV_1_1990...
URAL_2_1995...
RUSS_4_2008...
JAPA_1_1994...
CHEH_1_198...
KRAS_2_2002...
RUSS_1_1998...
URAL_1_1991...
STOC_6_196...
VAST_1_1983...
SMAL_4_199...
RUSS_3_1999...
CHEH_3_200...