Christian-Albrechts-Universität zu Kiel

Performance Comparisons between MGA software

Corinn Small

Biol 607

Dec - Feb. 2018

Previous studies point to high error rates in MGA software

- Multiple genome alignments produce many false positives (variants)

- Incorrect conclusions can significantly influence:
  - Evolutionary pattern or candidate gene inference
  - Direction of future studies

Markova-Raina, 2011

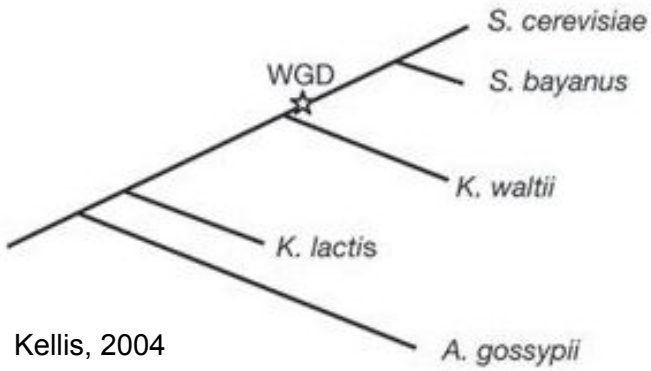
Alignment comparisons necessary to reduce false positives

False positives can arise from

- Duplications
- Indels
- Transposable elements
- Repeat rich regions

```
score=700
Sp_CBS432.CP020258.1           7566 86 + 71482 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_W7.scaffold1420_size3763     464 86 +  3763 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Y6_5.scaffold344_size800     465 86 +   800 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Q95_3.scaffold307_size775    440 86 -   775 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_T21_4.scaffold419_size555    223 86 +   555 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Y9_6.scaffold505_size729       0 62 +   729 -TATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Q59_1.scaffold245_size843    511 86 -   843 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_N44.scaffold368_size5437    3828 74 +  5437 ATATAAT-TA-----CC-ATAATAA-ATT---AATTTTT
Sp_Y8_5.scaffold386_size1105      0 69 +  1105 -TATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_S36_7.scaffold413_size1167   414 86 -  1167 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Z1.scaffold647_size1173      168 86 +  1173 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_YPS138.scaffold420_size4720  252 83 -  4720 -TATTATATATTTTTTTTATTATAATATTTTAAATAAAT
Sp_Y7.scaffold254_size1697      434 86 -  1697 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
Sp_Z1_1.scaffold212_size475     143 86 -   475 ATATAATATATTTTTTC-ATTATAATATTTTAAATAAAT
```

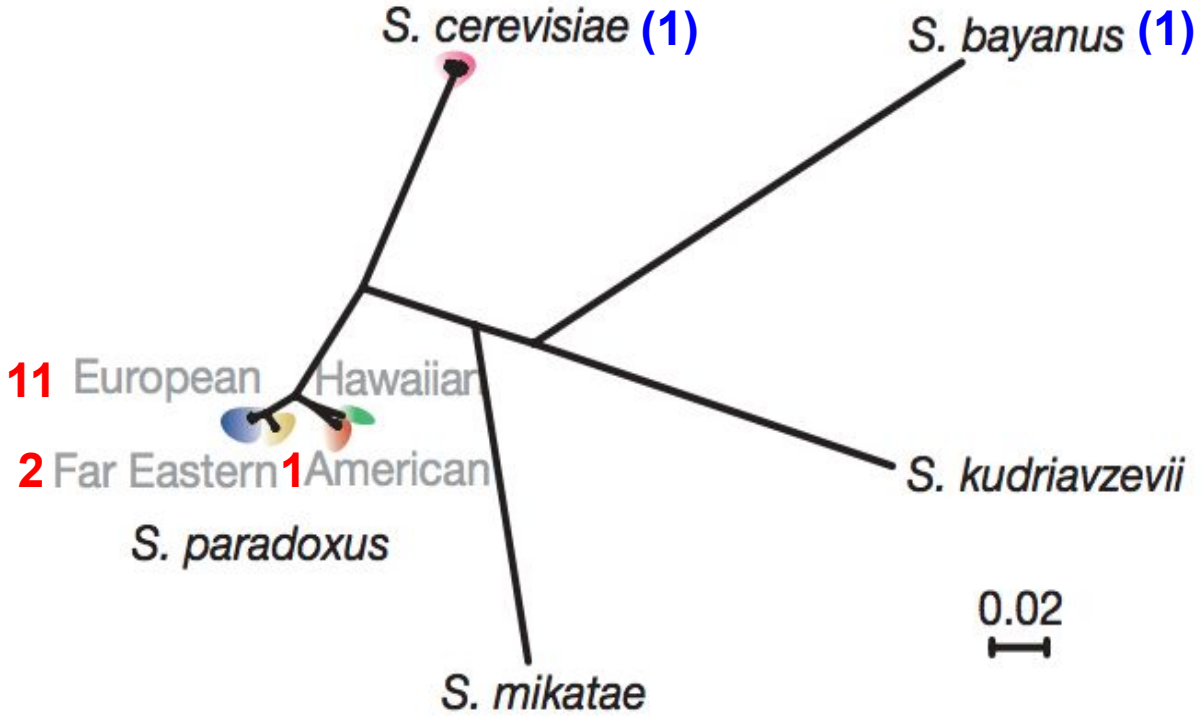
A whole-genome duplication (WGD) event occurred leading to *Saccharomyces* genus



Kellis, 2004

We analyzed 2 different datasets

1) 14 strains *Saccharomyces paradoxus*
2) 14 + 2 outgroups



Liti, 2009

We focused on the 14 *S. paradoxus* strain dataset

13

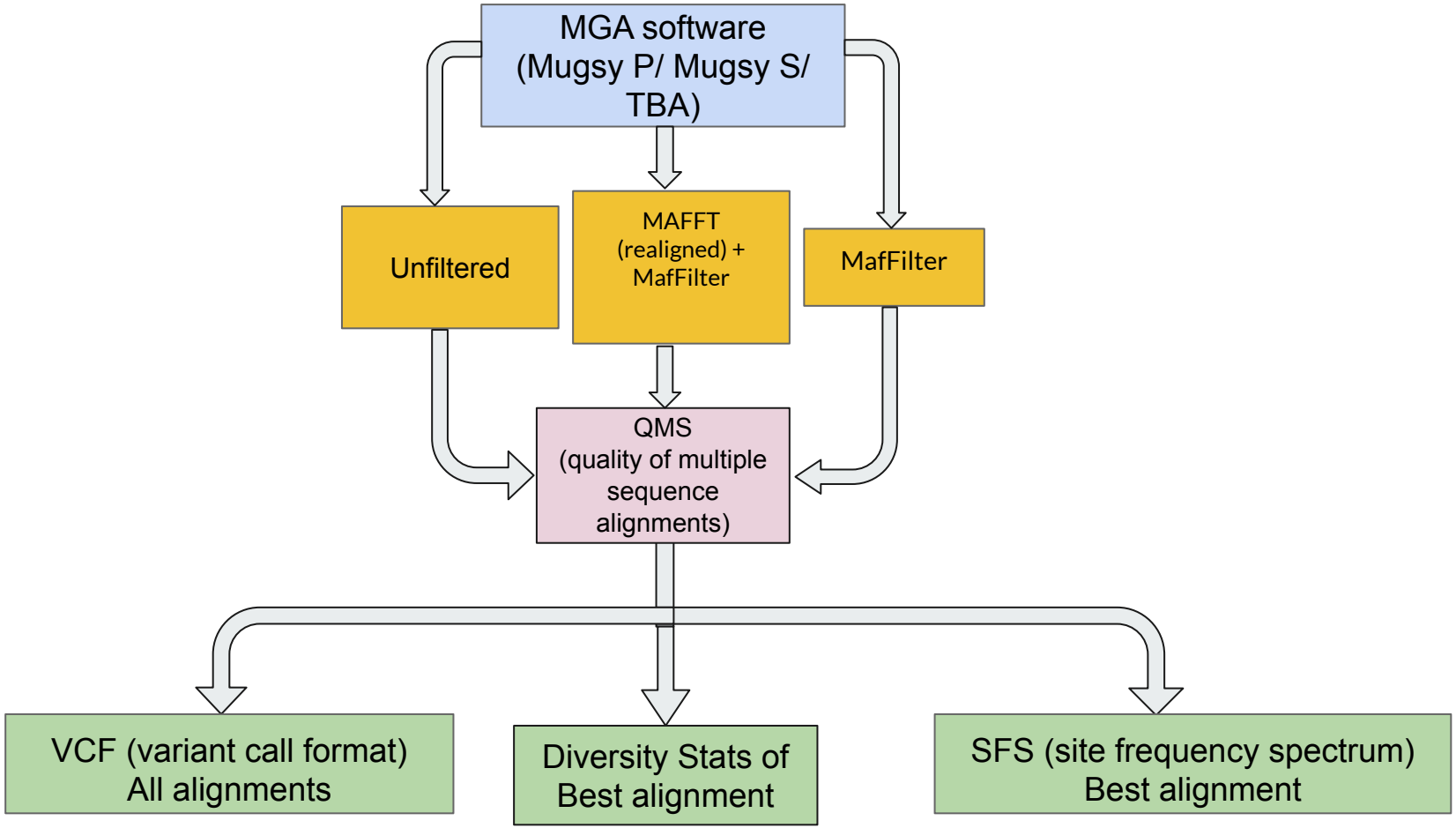
Strain	Region	Source	Country
Y8.5	European	*Quercus* spp.	UK
Z1.1	European	*Quercus* spp.	UK
Y9.6	European	*Quercus* spp.	UK
Z1	European	*Quercus* spp.	UK
Q59.1	European	*Quercus* spp.	UK
N-44	Far Eastern	*Quercus* spp.	Russia
YPS138	American	*Q. velutina*	USA
S36.7	European	*Quercus* spp.	UK
Y6.5	European	*Quercus* spp.	UK
Y7	European	*Quercus* spp.	UK
Q95.3	European	*Quercus* spp.	UK
T21.4	European	*Quercus* spp.	UK
W7	European	*Quercus* spp.	UK

Bergström et al. 2014

Reference: CBS432
Country: Russia
Exact location & source: unknown

Liti,Carter, et al. 2009

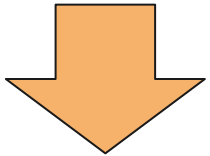
Pipeline for analysing the data



Generate pairwise alignments to generate MGAs

Basic procedure for all 3 programs:

Generate pairwise genome alignments



Generate MGA using the pairwise alignments

```
2 a score=1882
3 s N44.scaffold558_size1883 0 1883 + 1883
  agtcttcgcatcgacggattgctatcgttcccattatttttctcagaacct
  agttccttggattgtataaggttctcaacaatatgagaaggggaaaatacc
  ctagatcgggctcgttctgtagtattgtttgaactgtgtattttacttcaa
4 s Q59_1.scaffold100_size13350 1510 1885 + 13350
  agtctccgcatcgacggattgctatcgttcccattatttttctcagaacct
```

```
93 a score=3699
94 s Sp_CBS432.CP020251              35130 77 + 743843 TATTTTCATAGAA
95 s Sp_Y6_5.scaffold110_size30398    3439 77 -  30398 TATTTTCATAGAA
96 s Sp_Z1.scaffold6_size95555       5507 77 +  95555 TATTTTCATAGAA
97 s Sp_T21_4.scaffold100_size33679  3438 77 -  33679 TATTTTCATAGAA
98 s Sp_N44.scaffold281_size8884     5819 77 -   8884 TATTTTCATAGAA
```

* Mugsy P runs multiple pairwise alignments simultaneously compared to Mugsy S

Filtering was accomplished with MafFilter

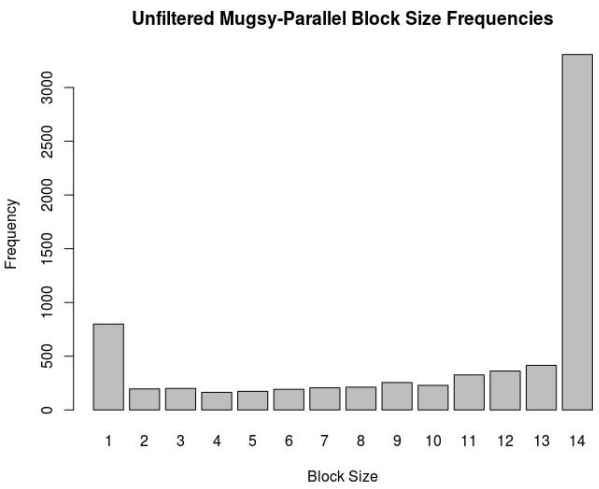
MafFilter options used to filter out unwanted alignment blocks:

1. MinBlockSize → only keeps blocks with all 14 individuals
2. XFull Gap → removes gap-only columns from blocks
3. AlnFilter2 → masks columns within a window that contain more than 5 gaps
4. MinBlockLength → only keeps blocks with more than 50 nt
5. MaskFilter → splits blocks by removing regions that contain too many masked columns (2)
6. AlnFilter → splits blocks by removing ambiguous regions

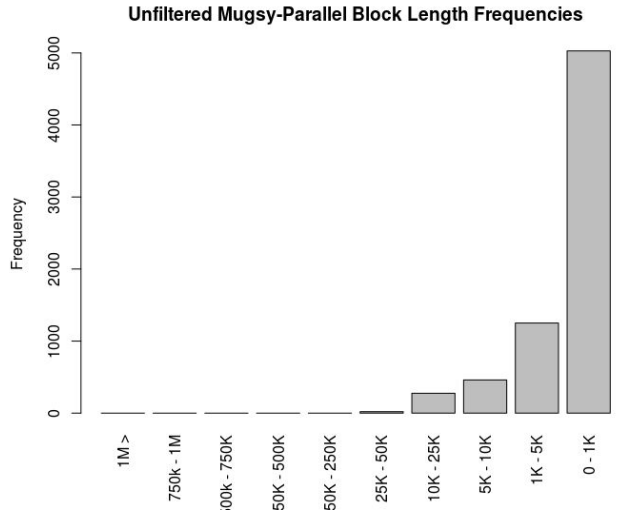
http://biopp.univ-montp2.fr/manual/html/maffilter/v1.2.1/maffilter.html#MaskFilter

QMS calculates alignment statistics including

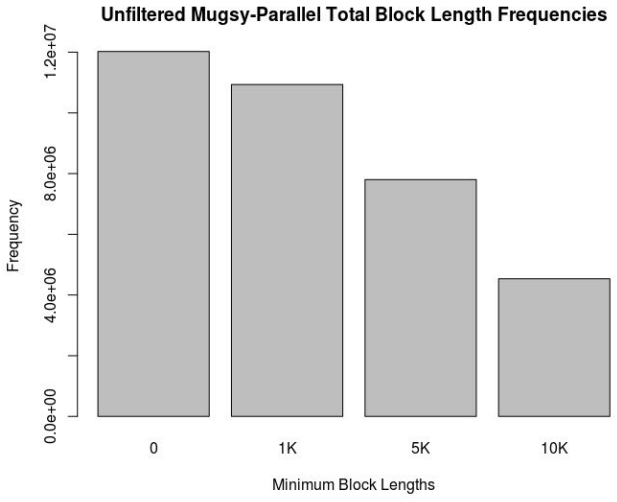
Block Size frequencies



Block Length frequencies

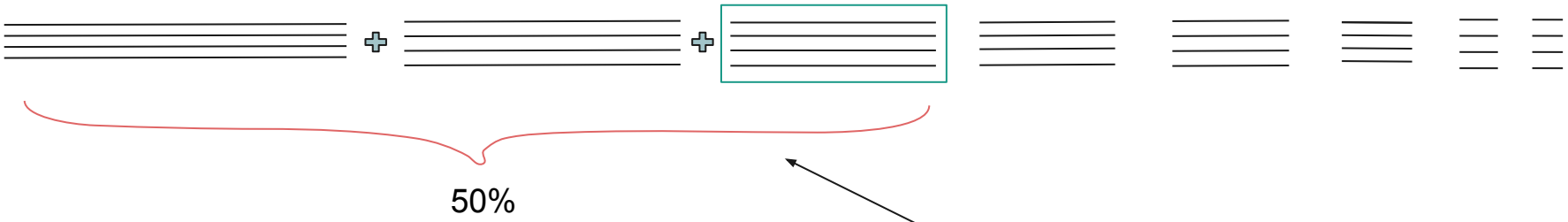


Total Block Length frequencies



Additional stats refer to the quality of the alignment

N50 & L50



N50 -> length of the smallest block included in the summed block lengths that make up 50% of total alignment

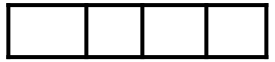
Larger N50 & L50 = higher quality

L50 refers to the number of summed blocks that give N50

L50 = 3

QMS alignment statistics scoring: filters

First, chose best type of filtering per software



	Total Length ↑	Blocks	Gap% ↓	N50 ↑	L50 ↑	Total Block Length ≥ 1kb ↑	Total Block Length: ≥ 5kb	Total Block Length: ≥10kb	Evaluation score
Tba UF	12229423	10061	1.28	3309	1101	11057280	7830332	4360009	NA
Tba F	8458355	40541	0.36	163	18831	236325	0	0	-7
Tba RF *	11028652	6397	0.3	1906	1940	9794899	3795678	946740	7
MugsyP UF	12023677	7036	2.59	3396	1049	10935200	7801887	4536046	NA
MugsyP F	11190803	6576	0.99	1792	2085	9818100	3536276	920109	-3
MugsyP RF	11176614	6301	0.77	1880	1982	9886060	3796347	1001818	3
MugsyS UF	12004021	6965	2.5	3424	1046	10931937	7785588	4501238	NA
MugsyS F	11187565	6529	0.98	1797	2080	9814145	3525269	920242	-3
MugsyS RF	11172450	6263	0.76	1886	1980	9883914	3773364	991599	3

* important result: when using TBA-- realigning is a must!

UF= unfiltered, F = filtered, RF = realigned & filtered

QMS alignment statistics scoring: software

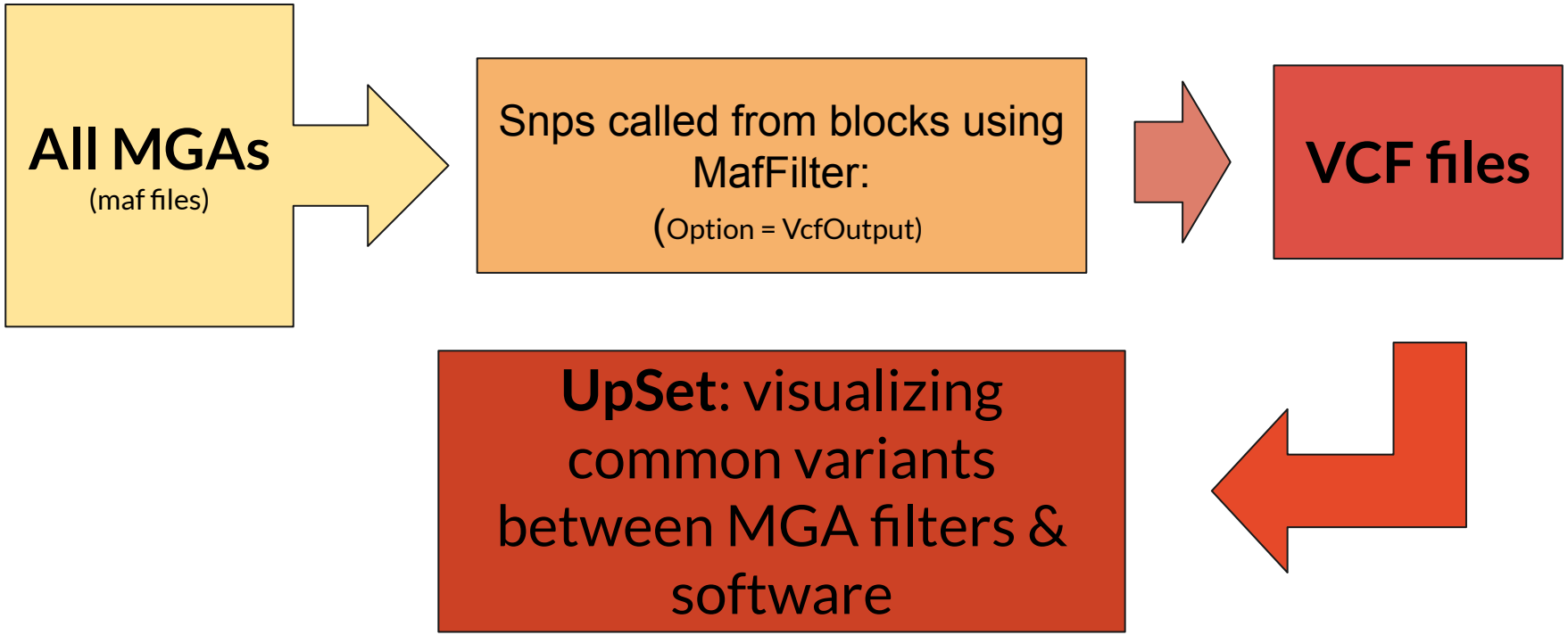
The best software-filtering combination: **MugsyP - realigned & filtered**

Key	-1	0	1	NA

	Total Length ↑	Blocks	Gap% ↓	N50 ↑	L50 ↑	Total Block Length: 1kb ↑	Total Block Length: 5kb	Total Block Length: 10kb	Evaluation score
Tba RF	11028652	6397	0.3	1906	1940	9794899	3795678	946740	-2
MugsyP RF	11176614	6301	0.77	1880	1982	9886060	3796347	1001818	3
MugsyP F	11190803	6576	0.99	1792	2085	9818100	3536276	920109	-1
MugsyS RF	11172450	6263	0.76	1886	1980	9883914	3773364	991599	0

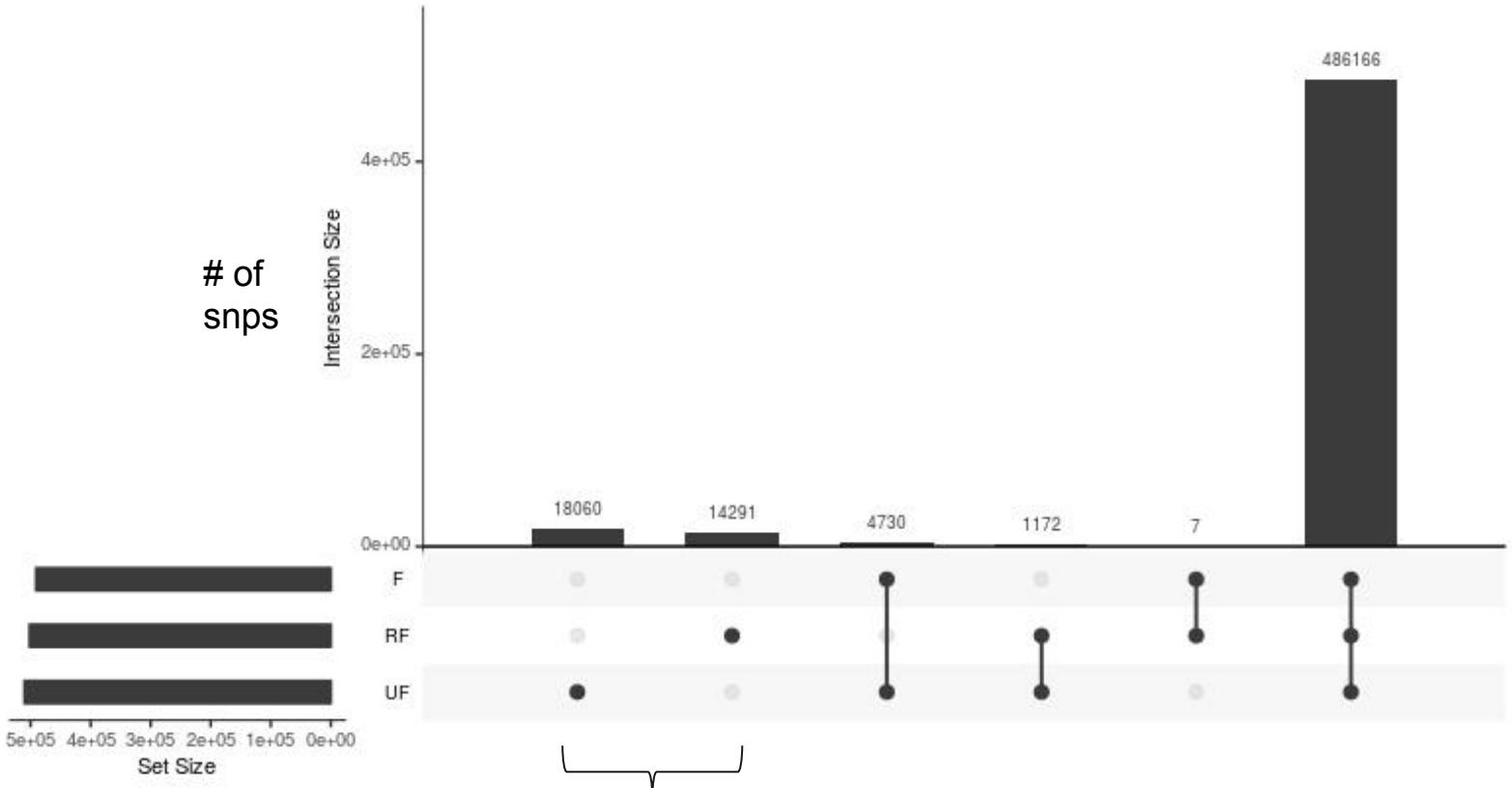
* Mugsy P-RF and Mugsy S-RF are very comparable= no significant differences

Snp calling procedure



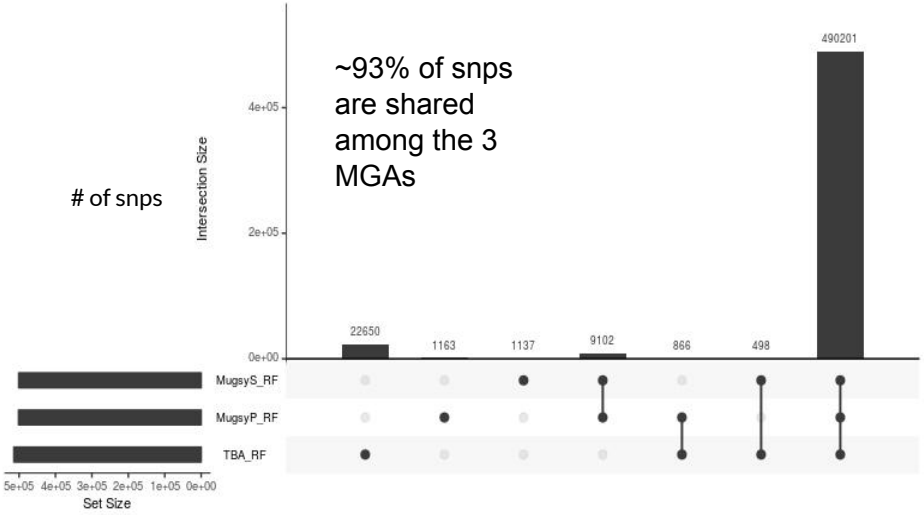
http://biopp.univ-montp2.fr/manual/html/maffilter/v1.2.1/maffilter.html#MaskFilter

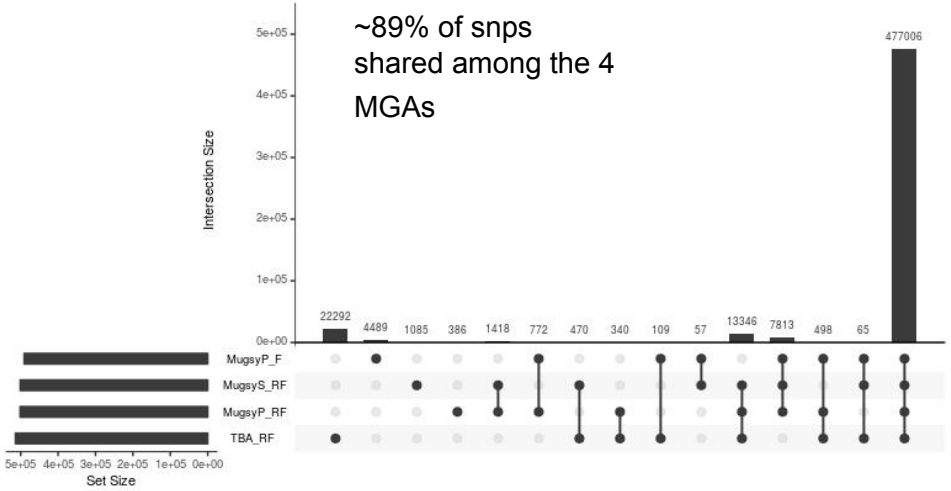
Large # of variants (~92%) are shared between all filter types



*Generated using UpSet

Large # of variants shared between software





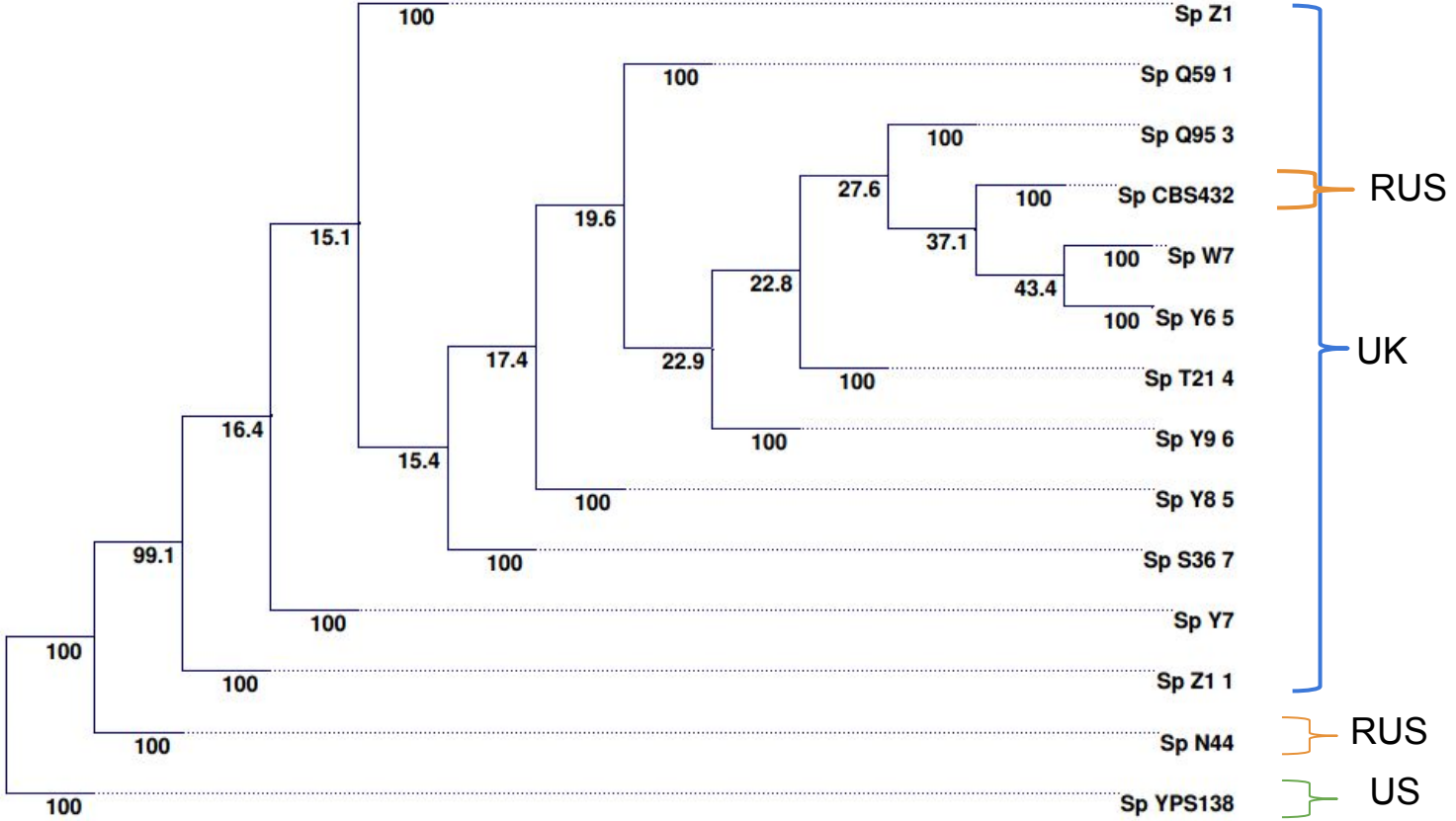
We chose to analyze the **Mugsy P-filtered** alignment

Next steps include:

- Generating sequence diversity statistics (Tajima's D, Tajima's Pi, Watterson Theta) (pop gen!) (generated from maf files)

- Analyzing population structure using PCA

Consensus Tree shows hypothesized relationships



*Consensus tree generated using genome-wide SNPs. Sub-trees constructed using UPGMA

Tajima's D provides insight into evolutionary mechanisms

Mean pairwise difference (Taj Pi) — # of segregating sites (Wat Θ) = Taj D

(2 measurements of nucleotide diversity)

What do we find?

When **Taj D > 0**

Low # of rare alleles in population indicating:

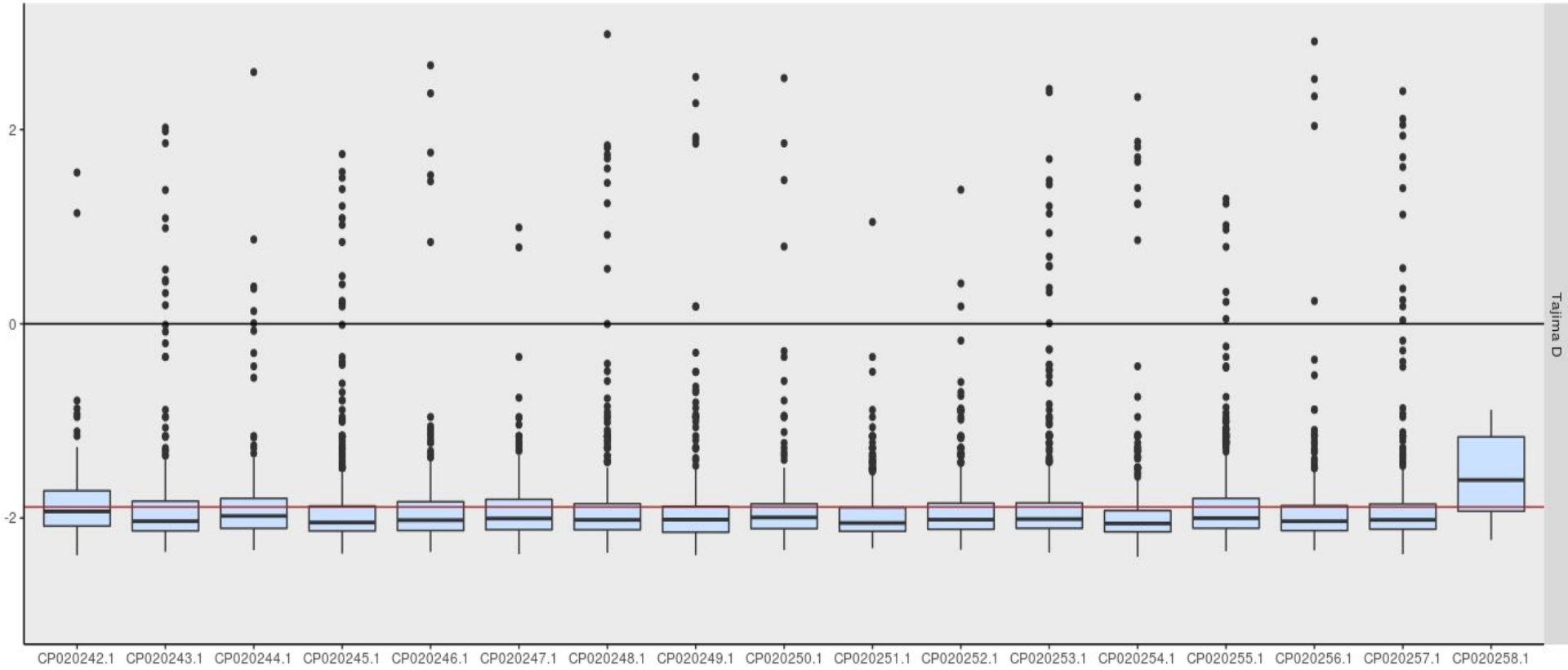
- Population shrinkage
- Balancing selection

When **Taj D < 0**

High # of rare alleles in population indicating:

- Population expansion after a bottleneck
  - Selective sweep (adaptation)
- Purifying selection (deleterious)

A low genome-wide Tajima's D



Chromosome 1 - 16 (including mtDNA*)

*

Potential reason for a low genome-wide D value

Technical

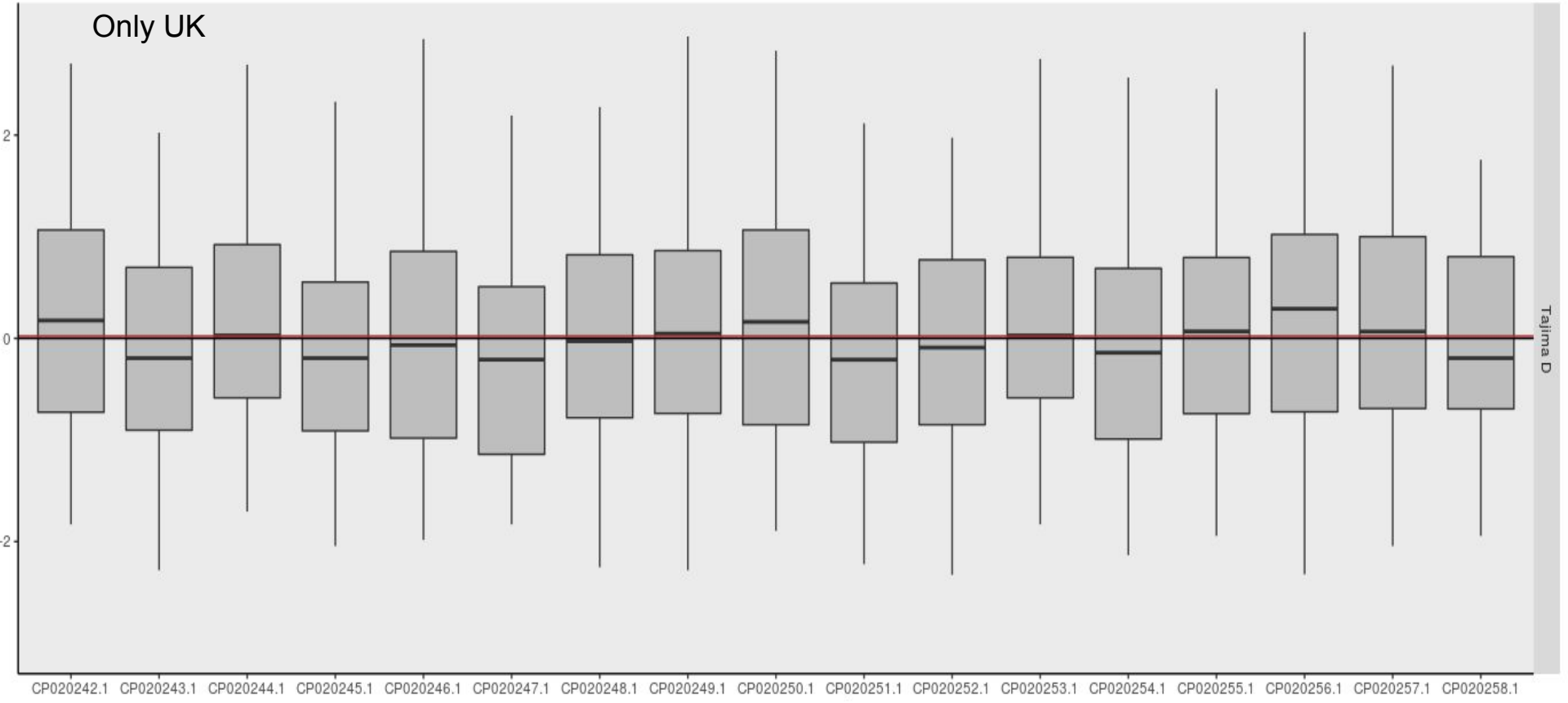
- Sampling bias: due to lack of equal population sample sizes

➔ Perhaps removal of more distantly related isolates could remove excess rare alleles

Biological (how to check)

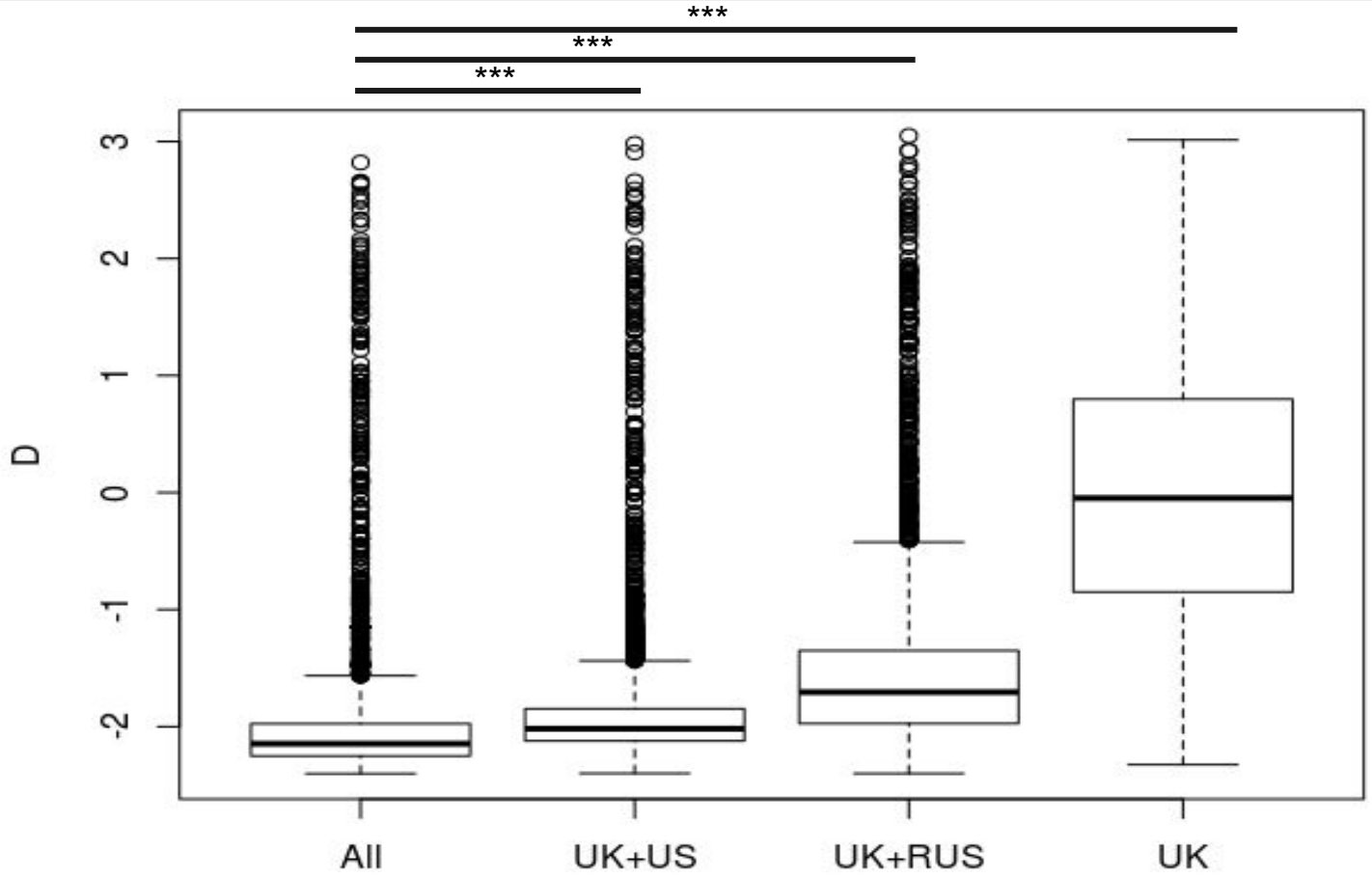
- Clonal propagation (linkage disequilibrium)
- Inbreeding (F_{ST} & F_{IS} inbreeding coefficients)
- Recent bottleneck event -> pop expansion (effective population sizes)

Removing RUS & US isolates from analysis increases D



Chromosome (including mtDNA)

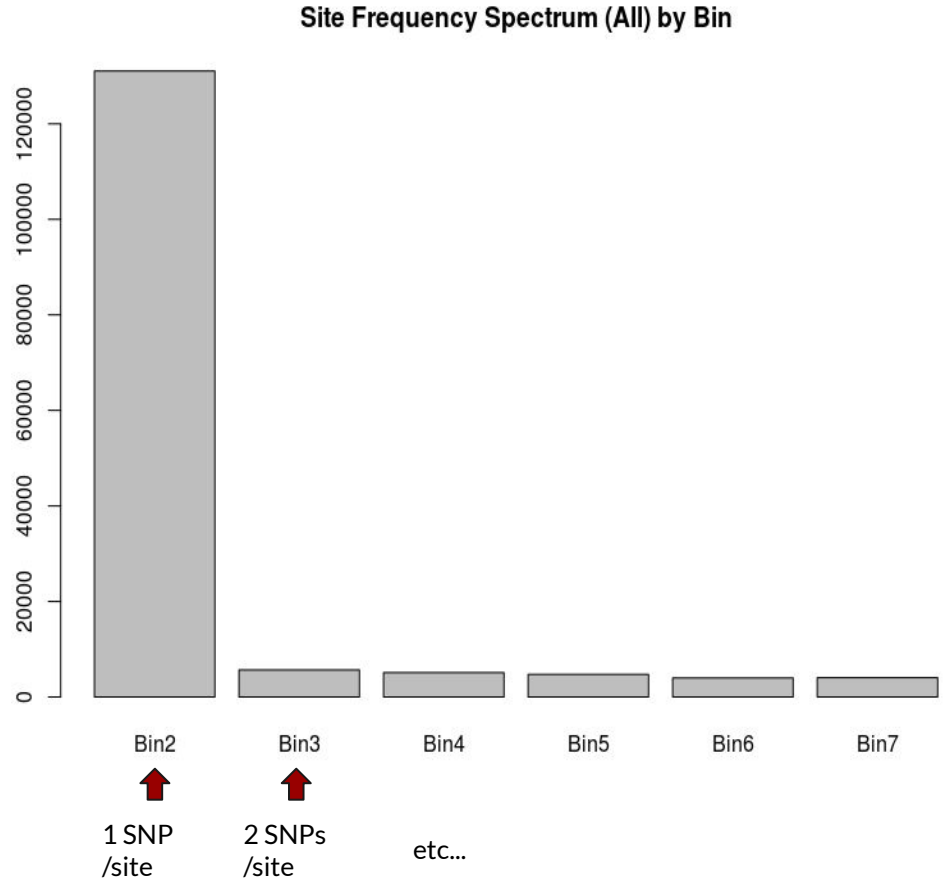
The US and RUS add a large amount of rare alleles to the studied sample.

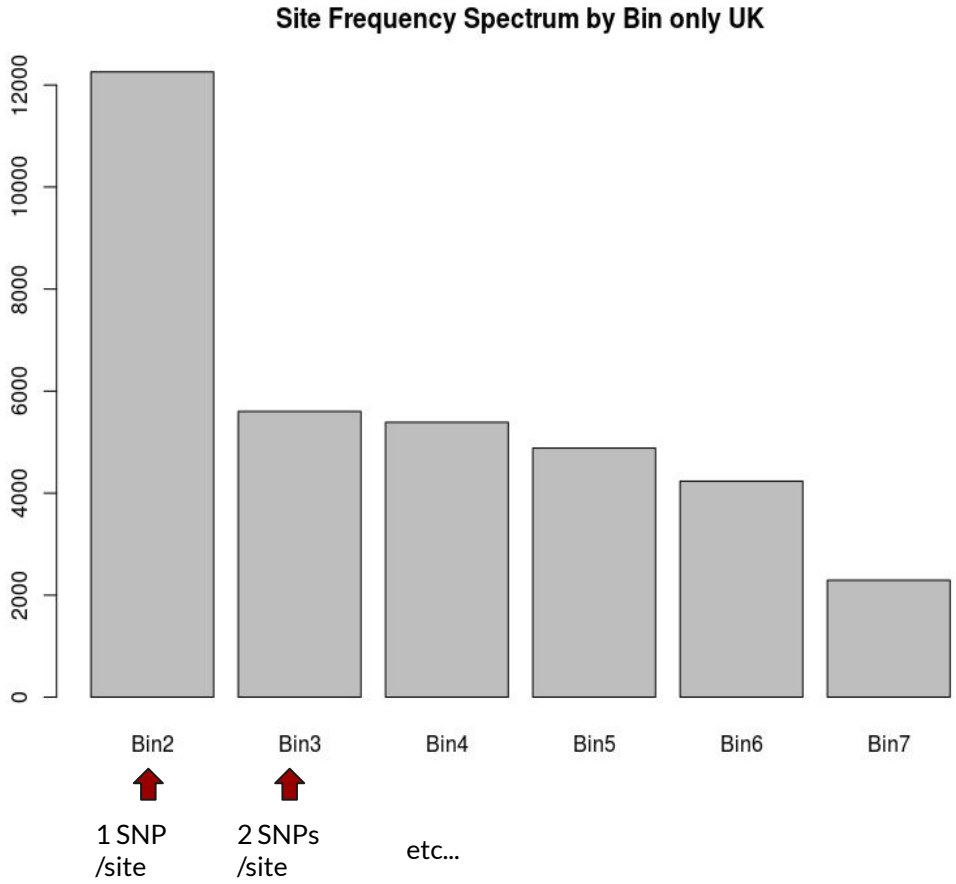


*** p-value < 2.2e-16

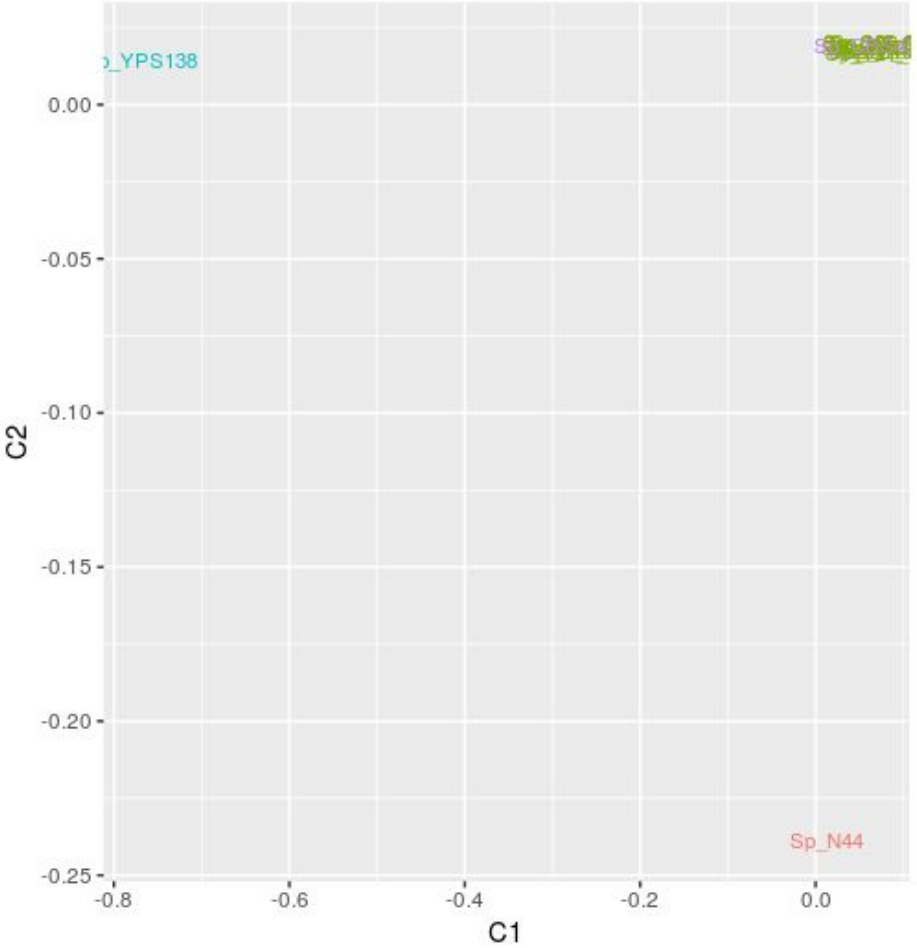
Wilcoxon rank sum test

Rare alleles are more frequent among whole dataset





PCA concludes: sub-“populations” vary drastically



Source

a	Exudate of Q. mongolica	Far East (Russia)
a	Quercus spp bark	UK
a	Soil beneath Q. velutina	US
a	Unknown	Moscow, Russia

- Variation comes from the US and Russia isolates

Outlook: investigating temperature adaptation

- Improve dataset: sample and population size, regional distribution
- Investigate potential adaptation to temperature
    - Identify potential genes that correlate to snps and regions of interest
    - GWAS and QTL

Thank you!

Supervisors: Christoph Eschenbrenner & Prof. Dr. Eva H. Stukenbrock

References

1. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 2011;863–74

2. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al. Population genomics of domestic and wild yeasts. Nature [Internet]. 2009;458(7236):337–41. Available from: http://dx.doi.org/10.1038/nature07743

3. Dutheil JY. MafFilter Manual 1.2.1 [Internet]. 2017. Available from: http://biopp.univ-montp2.fr/manual/html/maffilter/v1.2.1/maffilter.html#VcfOutput

4. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. (3):617–24.