UNIL | Université de Lausanne
SIB Swiss Institute of Bioinformatics
Vital-IT High Performance Computing Center

Gap.Jumper:
a probabilistic approach for single nucleotide variant quality assessment from samples, replicates and software results

Pawel Rosikiewicz[1]*, Frédéric G Masclaux[1,2], Tania Wyss[1],
Frédéric Schütz[2], Marco Pagni[2], Ian R. Sanders[1]

[1] Department of Ecology and Evolution, University of Lausanne, Bâtiment Biophore, Lausanne, 1015, Switzerland
[2] Vital-IT, SiB, Swiss Institute of Bioinformatics, Bâtiment Génopode, Lausanne 1015, Switzerland,
* mail: prosikie@unil.ch

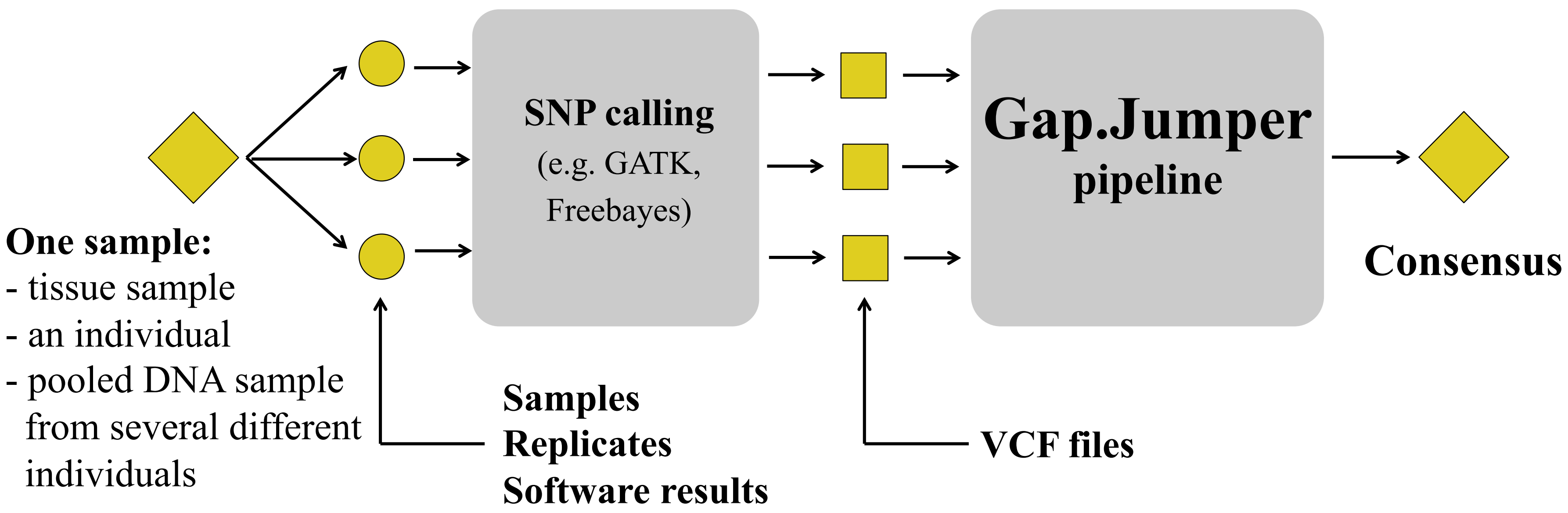
Next generation sequencing (NGS) allows screening of genetic polymorphisms in samples with a high genetic polymorphisms such as samples of carcinoma or from organisms with different ploidy (e.g. pathogenic fungi)

Problems are:

- limited coverage
- missing data
- sequencing errors
- different results obtained with different software's
- technical and biological differences between replicates

Consequently, researchers are faced with data containing a large number of apparently variable positions that need to be confirmed with independent experimental approach

Gap.jumper allows integration of variant calling data obtained from different samples, replicates and software results



Gap.Jumper allows estimating uncertainty associated with each nucleotide based on available empirical data

Three replicates that were sequenced separately

	Replicate 1	Replicate 2	Replicate 3
j =1	A	A	A
j =2	A/T	A/T	A/T
j =3	G	G	na
j =4	A/C	A/T	A/T
j =5	A	na	na
j =6	A	A/T	A/T
j =7	na	A	A
j =8	A/G	A/G	A
j =9	C	na	C
j =10	T	na	na

j1 – j10 denotes positions with alleles different than in the reference genome

Consensus

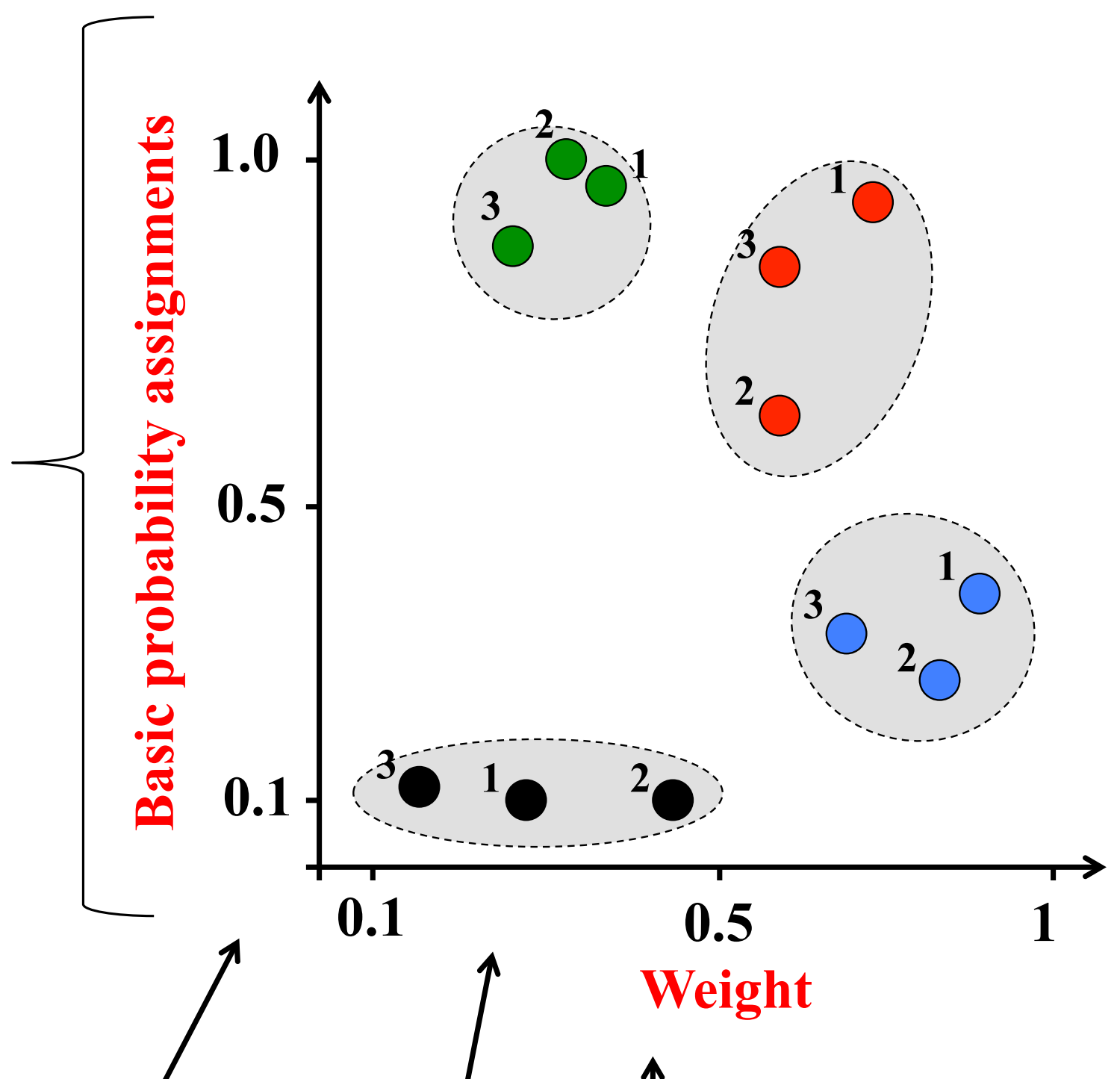
Evidence scores

A	C	G	T
1.0	.	.	.
1.0	.	.	1.0
.	.	1.0	.
1.0	0.2	.	0.9
0.3	.	.	.
1.0	.	.	0.8
0.9	.	.	.
1.0	.	0.7	.
.	0.8	.	.
.	.	.	0.5

Based on:
- Allele frequency
- Number of reads

Evidence scores inform how reproducible are the results among the replicates

Calculated as weighted probability

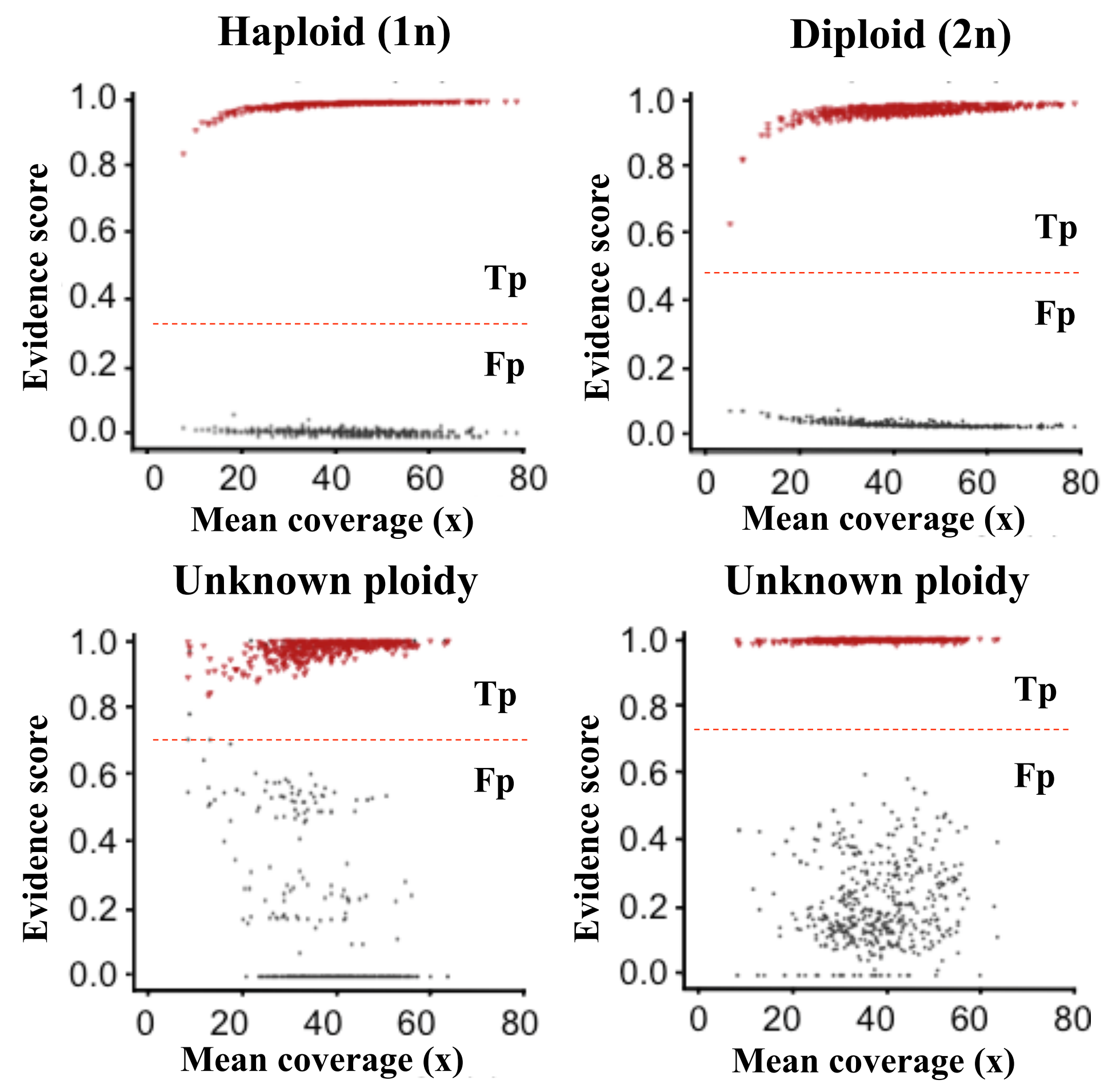


Based on:
- Coverage
- Error rate
- Replicates similarity
- Amount of missing data

Additional weights are assigned to missing nucleotides in order to asses the uncertainty associated with nucleotides in the other replicates

FNSNF
FONDS NATIONAL SUISSE
DE LA RECHERCHE SCIENTIFIQUE

Evidence scores can be used to remove potential errors or to rank positions based on their quality

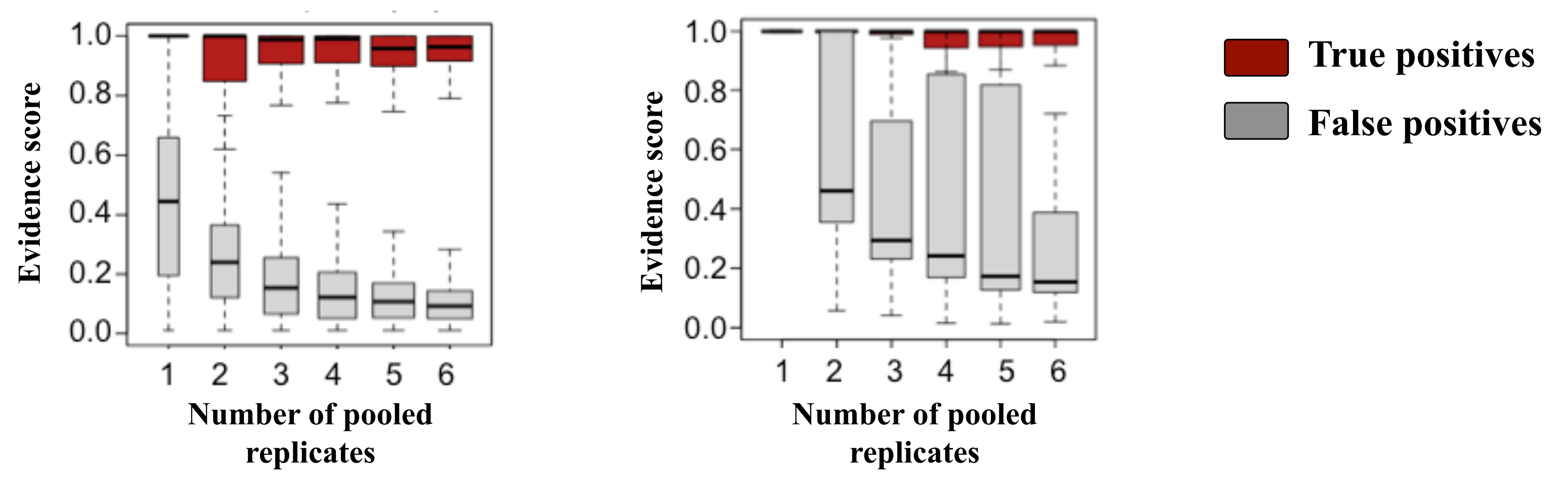


Optimal thresholds can be calculated individually allowing identification of true positives (Tp) and false positives (Fp) in each dataset by using small number of validated positions.

Accuracy improves with increasing number of pooled replicates

Sequence data with high coverage and low amount of missing data

Sequence data with low coverage and high amount of missing data



Evidence scores can be used to estimate uncertain of polymorphisms detected between different samples

Application example:

Identification of the fungal clonal lines

Goal:

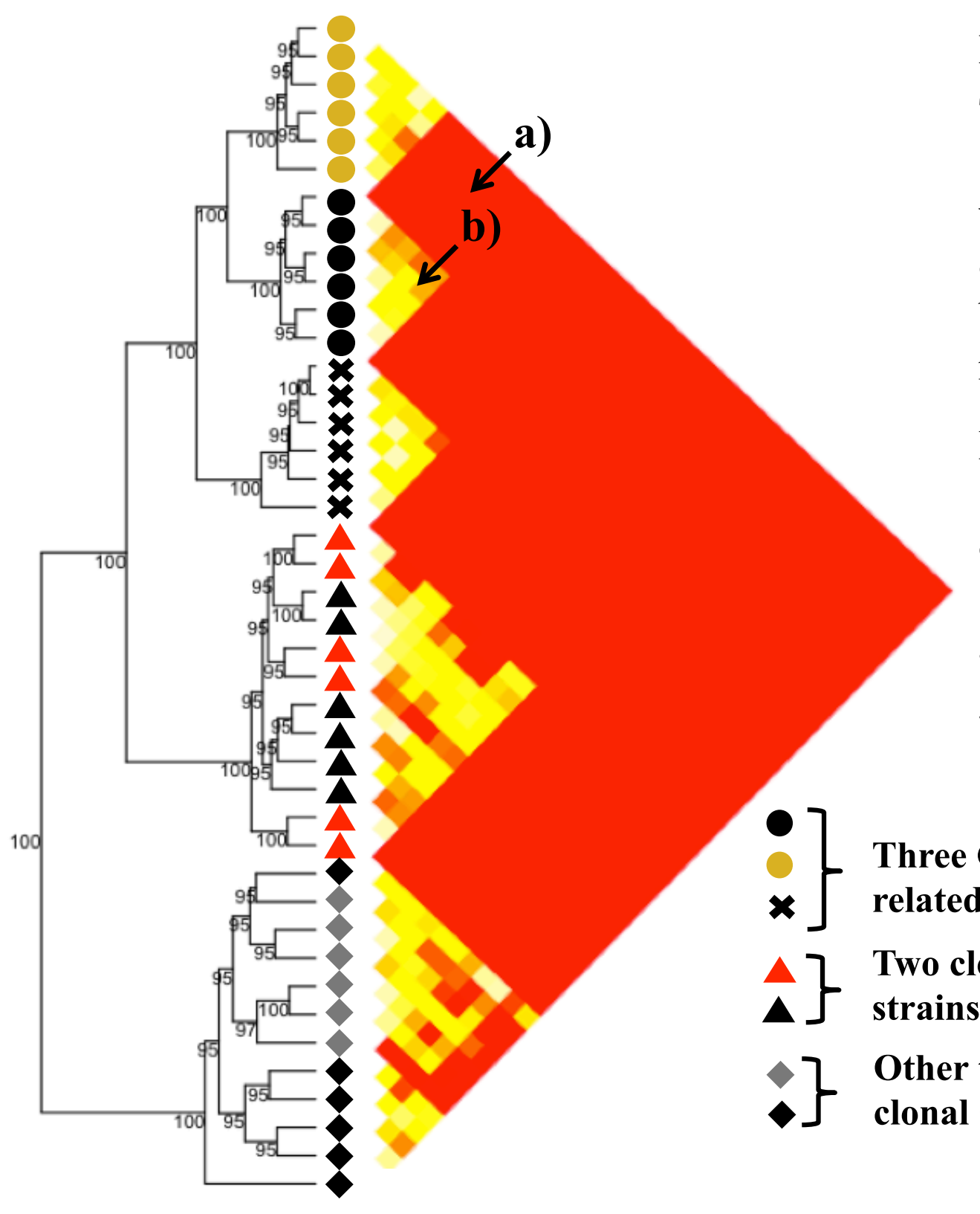
To identify which fungal strains are clonal offspring produced from a common parent (benchmarking studies with known classes)

Methods:

We genotyped 42 fungal strains (RAD-seq) - three replicates of each strain were sequenced.

The replicates of each strain were used to built a consensus (DST-based approach), which was compared to consensus built for other strains

Two group of strains which were previously identified as clonal lines were used as control (black and red triangles and black and grey squares).

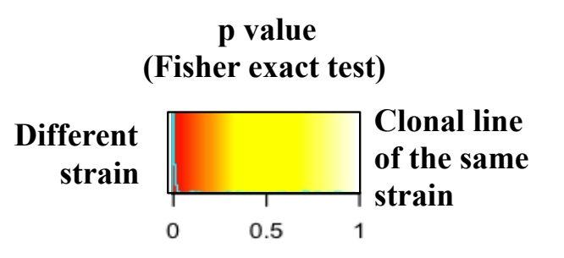


● ● × } Three Closely related strains

▲ ▲ } Two clonal strains

◆ ◆ } Other two clonal strains

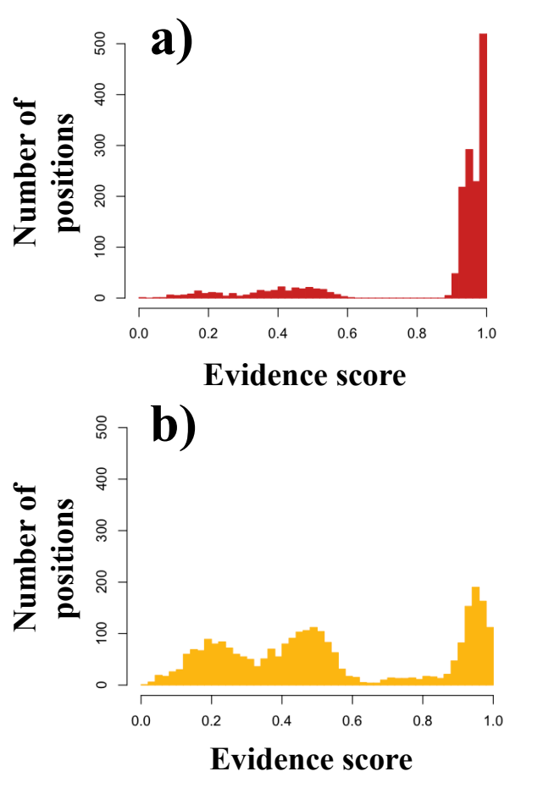
To classify two strains as clones we compared the number of polymorphisms to the expected number of errors in each comparison



Results:

The comparison performed between all strains showed large number of positions with different nucleotide composition, even in the comparisons performed between strains that were used as negative control (see legend).

However, polymorphisms found between strains classified to different groups of clones obtained higher evidence scores (a) than the polymorphisms found among strains classified into the same group of clones (b)



CONCLUSIONS:

- validates SNPs accurately in sets with a relatively small number of replicates (2-6)
- handles missing information easily
- handles different ploidy levels

APPLICATIONS:

- in screening studies
- to identify rare alleles and mutations
- to identify novel genetic markers
- to evaluate results obtained with other software's