Estimation of Haplotype Frequencies from Pooled Targeted Sequencing in Maize Landraces

Minguella Raphaël, Aurélie Canaguier, Delphine Madur, Damien Hinsinger, Stephane Nicolas, Patricia Faivre Rampant

HAL Id: hal-04670321

https://hal.inrae.fr/hal-04670321v1

Submitted on 21 Aug 2024

Estimation of Haplotype Frequencies from Pooled Targeted Sequencing in Maize Landraces

Minguella Raphaël[1], Canaguier Aurélie[1], Madur Delphine[2], Hinsinger Damien D.[1], Nicolas Stéphane D.[2], Faivre-Rampant Patricia[1]

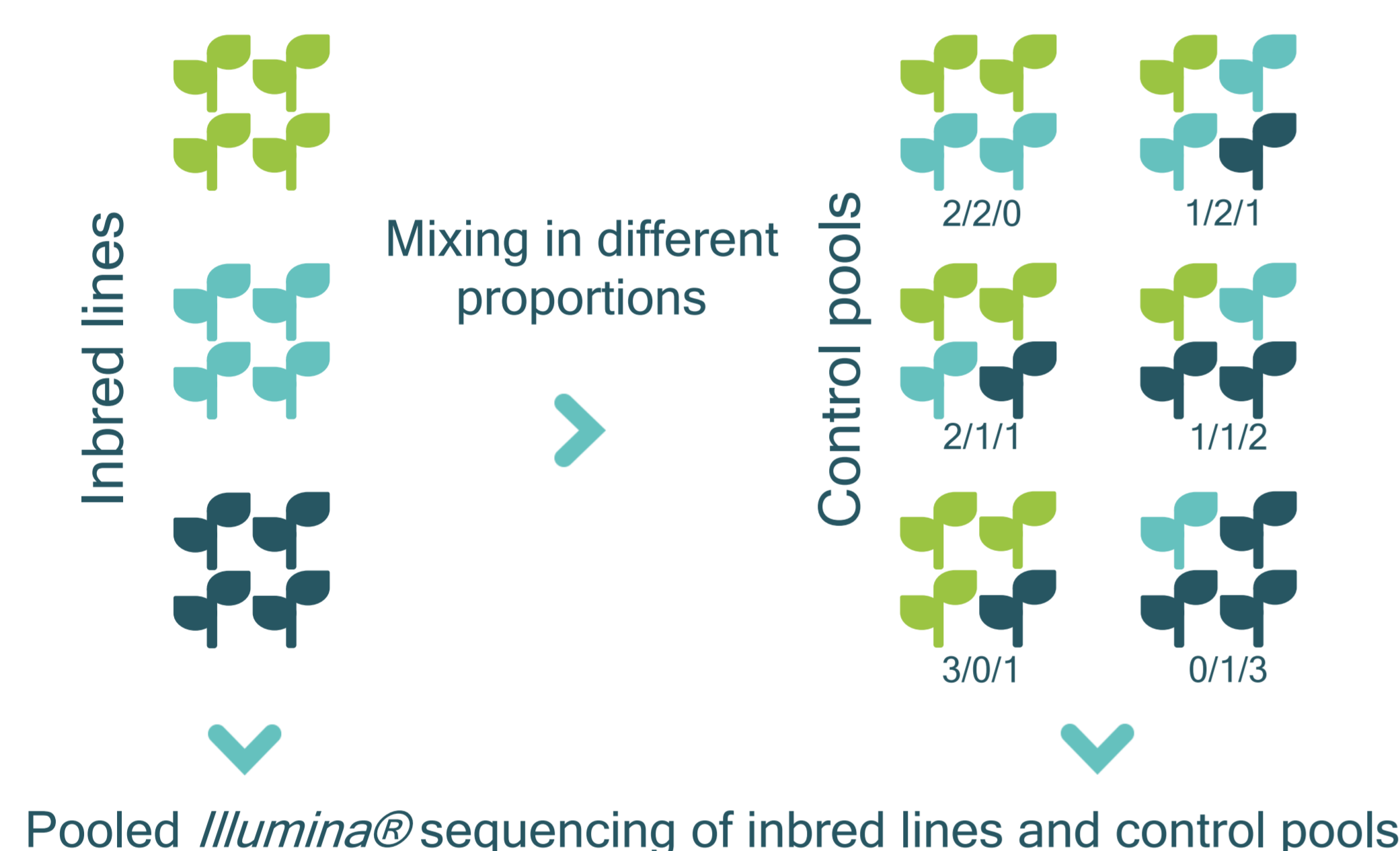[1] INRAE, EPGV, Evry, France   [2] INRAE, GQE, Gif-sur-Yvettes, France

1. Background

Maize **landraces** harbor a wider **diversity** than inbred lines and hybrids that can be harnessed for plant breeding. Although this diversity has been largely studied in the past, much more remains to be done. However, analyzing it requires to genotype many individuals per populations. A solution to reduce the cost of the study is to genotype **pools** : several plants of a given accession are mixed, then the DNA extraction is done in pool. In addition, using short **haplotypes** usually improve population structure inference but requires sequencing. For this purpose, we hijacked a **targeted** Genotyping By Sequencing (tGBS) technology to sequence haplotypes from targeted genomics regions chosen among the many QTLs being known in maize.
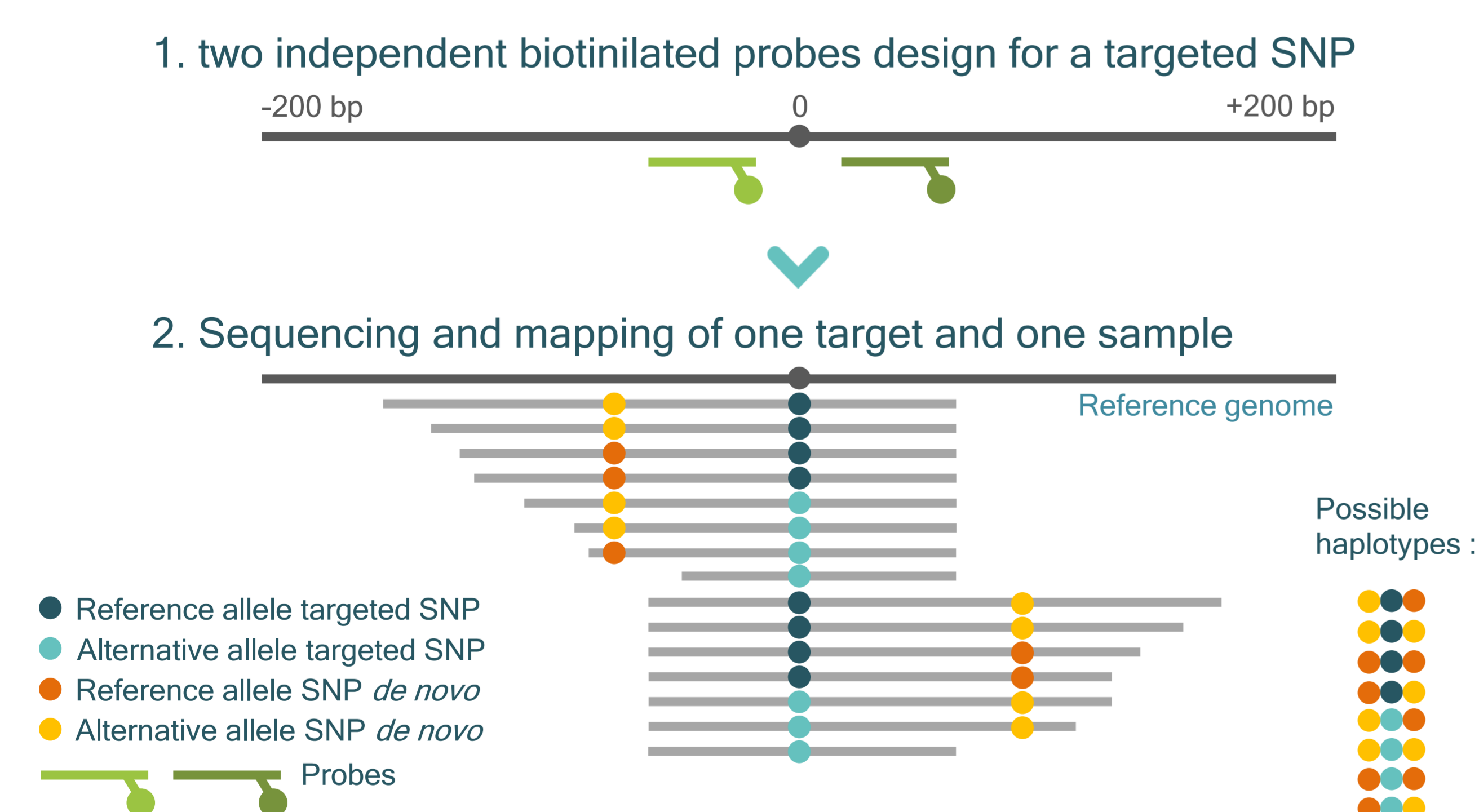
Herein, we **validate** haplotypes frequencies **estimation** quality for pool genotyping using control pools.
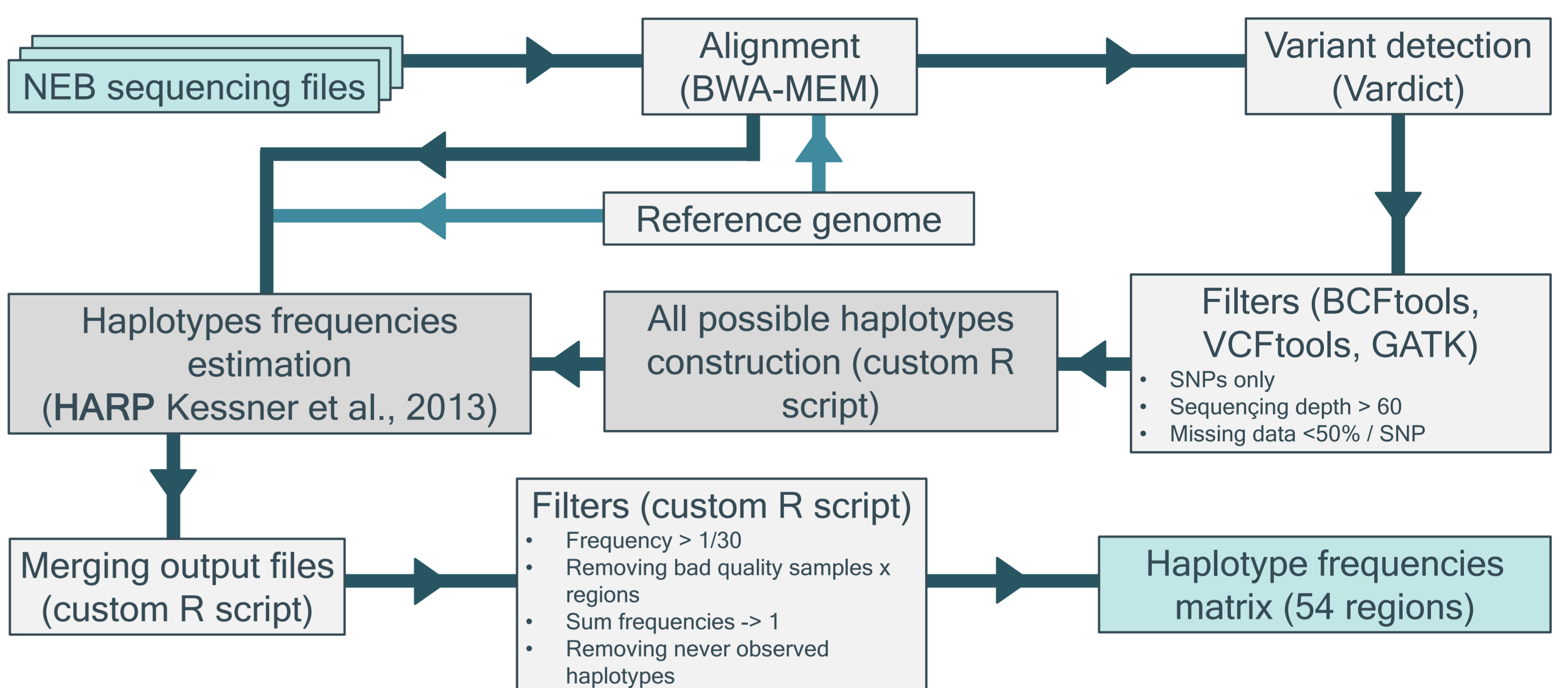
2. Control pools design

To **evaluate** haplotypic frequencies **estimation quality** we designed **control pools**. These pools are **mixes** of DNA from **inbred lines** in **known proportions** or known F1 hybrids. Therefore, we can calculate **expected** haplotype **frequencies** for each control pool from the haplotyping of inbred lines that is assumed to be correct, and then, compare it to the **observed frequencies** in each control pools. We used 30 control pools : 5 F1 hybrids and 25 mixes of 3 inbred lines including different heterotic groups.
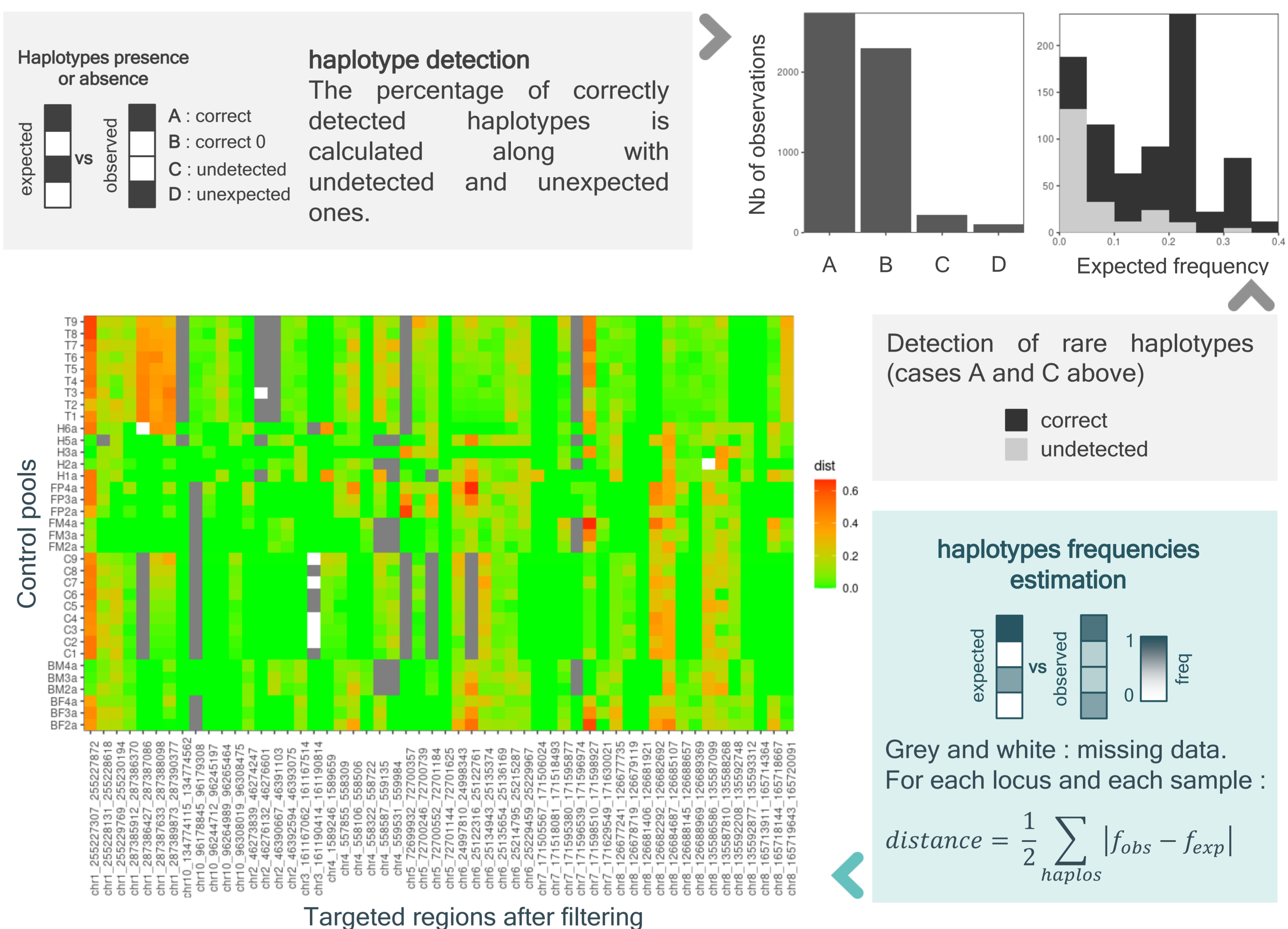

Inbred lines → Mixing in different proportions → Control pools
2/2/0   1/2/1
2/1/1   1/1/2
3/0/1   0/1/3

Pooled *Illumina®* sequencing of inbred lines and control pools

3. tGBS : *NEBnext® direct genotyping solution*

1. two independent biotinilated probes design for a targeted SNP
-200 bp   0   +200 bp

2. Sequencing and mapping of one target and one sample
Reference genome

● Reference allele targeted SNP
● Alternative allele targeted SNP
● Reference allele SNP *de novo*
● Alternative allele SNP *de novo*
— Probes

Possible haplotypes :

4. Pipeline

NEB sequencing files → Alignment (BWA-MEM) → Variant detection (Vardict)

Reference genome

Haplotypes frequencies estimation (**HARP** Kessner et al., 2013) ← All possible haplotypes construction (custom R script) ← Filters (BCFtools, VCFtools, GATK)
• SNPs only
• Sequencing depth > 60
• Missing data <50% / SNP

Merging output files (custom R script) → Filters (custom R script)
• Frequency > 1/30
• Removing bad quality samples x regions
• Sum frequencies -> 1
• Removing never observed haplotypes
→ Haplotype frequencies matrix (54 regions)

5. Performances

Haplotypes presence or absence
expected vs observed
A : correct
B : correct 0
C : undetected
D : unexpected

haplotype detection
The percentage of correctly detected haplotypes is calculated along with undetected and unexpected ones.

Nb of observations
A B C D
Expected frequency

Control pools (y-axis) / Targeted regions after filtering (x-axis)
dist 0.6 0.4 0.2 0.0

Detection of rare haplotypes (cases A and C above)
■ correct
■ undetected

haplotypes frequencies estimation
expected vs observed   freq

Grey and white : missing data.
For each locus and each sample :

$$distance = \frac{1}{2} \sum_{haplos} |f_{obs} - f_{exp}|$$

6. Conclusions

• Haplotypes are qualitatively detected and their frequencies are correctly estimated in pools
• The rarest haplotypes (frequency < 0.05) are hardly detected but otherwise haplotypes are detected
• Landraces can be studied with this pipeline
• Further analysis can be performed on this data including Genetic-Environment Association

Bibliography

NEBnext® direct genotyping solution:
Emerman AB, Bowman SK, Barry A, Henig N, Patel KM, Gardner AF, Hendrickson CLJCPiMB: NEBNext Direct: A Novel, Rapid, Hybridization-Based Approach for the Capture and Library Conversion of Genomic Regions of Interest. 2017;119(1):7.30. 31-37.30. 24.

HARP :
Darren Kessner, Thomas L. Turner, John Novembre, Maximum Likelihood Estimation of Frequencies of Known Haplotypes from Pooled Sequence Data, Molecular Biology and Evolution, Volume 30, Issue 5, May 2013, Pages 1145–1158, https://doi.org/10.1093/molbev/mst016