Minguella Raphaël[1], Canaguier Aurélie[1], Madur Delphine[2], Berard Aurélie[1] , Le Clainche Isabelle[1], Galaretto Agustin-Oscar[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

Pipeline for Haplotype Frequencies Estimation from Pooled Targeted Sequencing in Maize

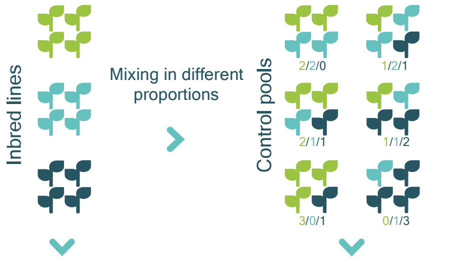
1. Background

Haplotypes are useful markers in population genetics due to a tighter link to populations history than SNP and are therefore considered more informative for populations structuration analyses. However, capturing populations **diversity** can be challenging because it requires to genotype many individuals which can be very expensive. An usual solution is to genotype populations in **pool**, meaning that several individuals of a same population are mixed and their DNA extraction is done in pool. Unfortunately, information about haplotypes is lost during the process because the DNA is fragmented and mixed. Several approaches have been proposed to rebuild haplotypes with reads overlap. Here implemented one of these algorithm in a pipeline to estimates short haplotypes frequencies from targeted **genotyping by sequencing** (tGBS) technology, which **reduces costs** by sequencing only desired genomic regions.

We assessed the accuracy of our pipeline using control pools with known haplotypes frequencies and we show that our pipeline gives correct haplotypic frequencies estimations when frequencies are higher than 5%.

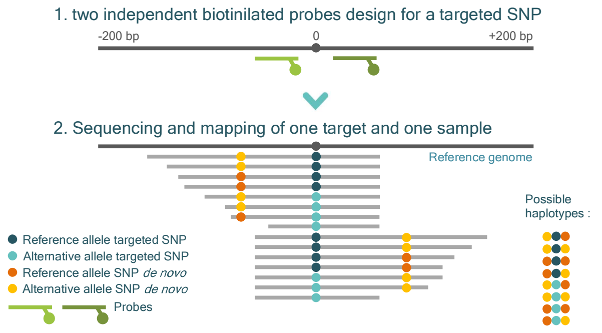
2. Control pools design

To **evaluate** haplotypic frequencies **estimation quality** we designed **control pools** that are **mixes** of DNA from homozugous **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 genetic groups.

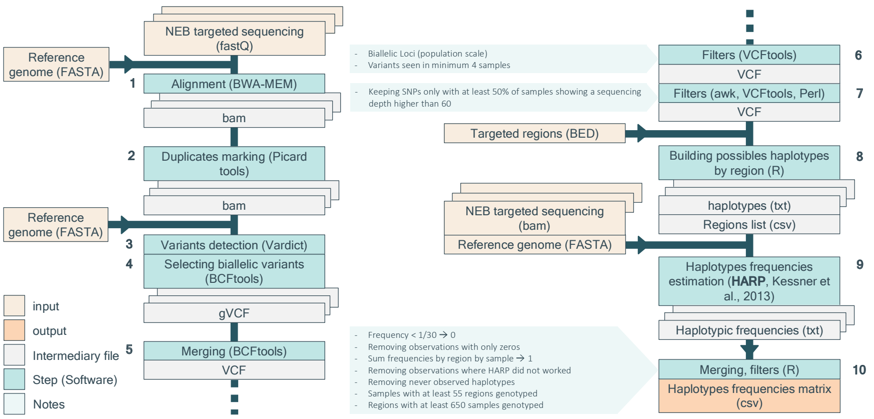


Pooled *Illumina®* sequencing of inbred lines and control pools

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
Possible haplotypes :
Reference allele targeted SNP
Alternative allele targeted SNP
Reference allele SNP de novo
Alternative allele SNP de novo
Probes

4. Pipeline

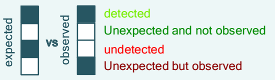

NEB targeted sequencing (fastQ)
Reference genome (FASTA)
1
Alignment (BWA-MEM)
bam
2
Duplicates marking (Picard tools)
bam
Reference genome (FASTA)
3
Variants detection (Vardict)
4
Selecting biallelic variants (BCFtools)
gVCF
5
Merging (BCFtools)
VCF
input
output
Intermediary file
Step (Software)
Notes
- Biallelic Loci (population scale)
- Variants seen in minimum 4 samples
- Keeping SNPs only with at least 50% of samples showing a sequencing depth higher than 60
Filters (VCFtools)
6
VCF
Filters (awk, VCFtools, Perl)
7
VCF
Targeted regions (BED)
Building possibles haplotypes by region (R)
8
haplotypes (txt)
Regions list (csv)
NEB targeted sequencing (bam)
Reference genome (FASTA)
Haplotypes frequencies estimation (HARP, Kessner et al., 2013)
9
Haplotypic frequencies (txt)
- Frequency < 1/30 → 0
- Removing observations with only zeros
- Sum frequencies by region by sample → 1
- Removing observations where HARP did not worked
- Removing never observed haplotypes
- Samples with at least 55 regions genotyped
- Regions with at least 650 samples genotyped
Merging, filters (R)
10
Haplotypes frequencies matrix (csv)

5. Performances

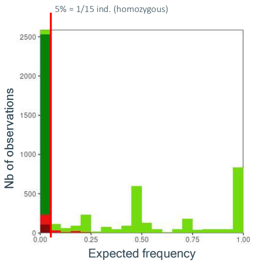
To estimate the **quality of haplotypes detection**, we assessed the qualitative detection of this haplotype (right) for each haplotypes of each region and each pool. Then, we calculated for each expected frequency the proportion of each category that we plotted (right).

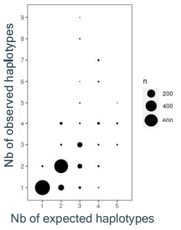
When haplotype frequency >5% :
- Very few false positives (=unexpected but observed haplotypes)
- Very few false negatives (=undetected haplotypes)

Haplotypes presence or absence



Good haplotypes detection if found in at least 1/15 ind. in the pool





To estimate the ability of our pipeline to **properly detect all haplotypes in a pool**, we calculated the number of observed and expected haplotypes for each region of each control pool :

- Correct number of detected haplotypes (1-2 haplotypes in a pool)
- When >2 haplotypes in the pool : a few undetected haplotypes
- >3 haplotypes observed due to residual heterozygosity in inbred lines
- Few cases with detected hap. > expected hap. (HARP errors)

The pipeline retrieves the expected number of haplotypes

Haplotypes frequencies

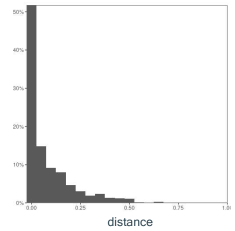
expected
vs
observed
1
0
freq

To estimate the **quality of frequencies estimation**, we calculated for each region and each pool the sum of absolute differences between expected and observed haplotypic frequencies :

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

- >50% of freq. correctly estimated (distance <0.05)
- >80% of freq. estimated with a distance <0.2

Observed frequencies close from expected ones, no very distant observation



6. Conclusions & perspectives

- Haplotypes are detected if there frequency is more than 5% (=1/15 individual in the pool) and frequencies are correctly estimated in pools
- Cheap short haplotypes sequencing approach, but improvement could be useful during probes design
- First results on actual data show that maize landraces are more diversified than maize inbred lines
- Maize landraces harbor haplotypes that are not in inbred lines

Center
Île-de-France – Versailles-Saclay

Acknowlegment

MineLandDiv
Mining allelic diversity of maize landraces for tolerance to abiotic and biotic stresses
GENOSCOPE
Amaizing

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

INRAE - Unité EPGV US1279
CEA - Institut de biologie François Jacob
Site d'Evry - Bat G1
2 rue Gaston Crémieux
91057 - Evry Cedex
France
support-epgv@inrae.fr

https://www6.versailles-grignon.inrae.fr/epgv