Development of a new high throughput 105K Presence/Absence Variation genotyping array for quantitative genetic studies

Clément Mabire, Jorge Duarte, Aude Darracq, Ali Pirani, Hélène Rimbert, Delphine Madur, Clémentine Vitte, Nathalie Rivière, Valerie Combes, Johann Joets, et al.

Development of a new high throughput 105K Presence/Absence Variation genotyping array for quantitative genetic studies

Mabire Clément [1*], Duarte Jorge [2*], Aude Darracq [1], Ali Pirani [3], Hélène Rimbert [2], Delphine Madur [1], Clémentine Vitte[1], Nathalie Rivière[2], Valérie Combes [1], Johann Joets [1], Jean-Philippe Pichon [2], Stéphane D. Nicolas [1]

*These two authors contributed equally to the work, *1 Génétique Quantitative et Evolution - Le Moulon, INRA - Université Paris-Sud - CNRS - AgroParisTech, Ferme du Moulon, F-91190, Gif-sur-Yvette, France*, *2 Biogemma - Upstream Genomics Group, route d'Ennezat, CS 90126, Chappes 63720, France, 3 Affymetrix - 3420 Central Expy, Santa Clara, CA, 95051, USA*

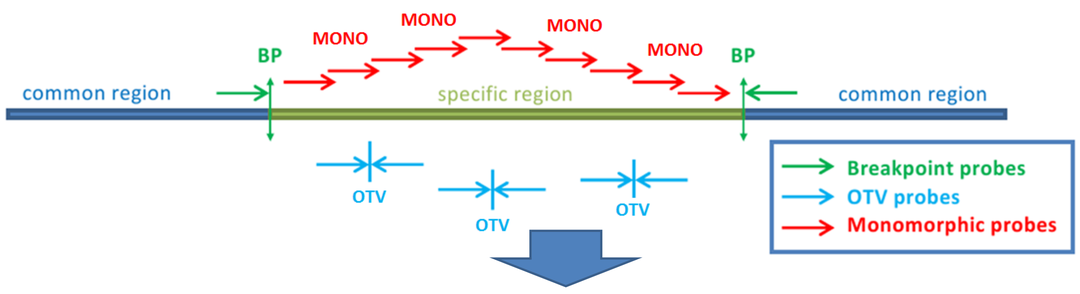
Context

In the last decades, there has been a growing interest for structural variations in plant genomes. By affecting gene content and order between individuals of the same species, copy number variations, including presence-absence variations (PAVs), can indeed have strong functional impacts. In maize, thousand of PAVs have been discovered using Comparative Genomic Hybridization technology (Springer et *al.* 2010) or by resequencing inbred lines (Hirsch et *al.*, 2016, Darracq et *al.*, 2018). Currently, there are indirect proofs that PAVs could have strong impact on trait variations (Lu et *al.*, 2015, Chia et *al.*, 2012) but no formal demonstration because of the lack of a high-throughput approach to accurately genotype PAVs on a large set of individuals. To address this issue, we developed a high throughput array allowing to genotype 105K PAVs on a large number of inbred lines, thus opening the way to PAV-based quantitative and population genetic analyses.

A two steps approach to genotype 105K Presence-Absence Variations (PAV)

1) **Discovery of 117K PAV sequences** and breakpoints by sequencing and assembling F2, C103 and PH207 lines. 79,969 insertions and 25,958 deletions regarding B73 reference genome were used for designing the genotyping array.

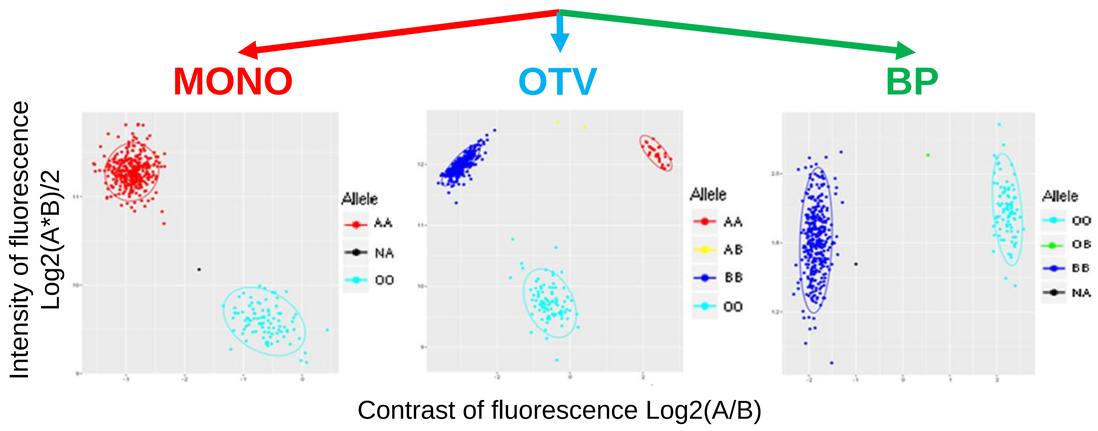
2) **Design of a high throughput genotyping Affymetrix axiom array** to genotype ~105k PAVs using ~600,000 probes (~6 PAVs / probe). 3 types of probe were designed: breakpoints (BP), SNP within PAV sequence (OTV) and monomorphic sequences in PAV (MONO):



Conversion of probes fluorescence into genotyping

By using existing Affymetrix algorithms to call BP and OTV.

➔Development of a **new algorithm** to call presence/absence for MONO probes.



Presence/absence genotyping was based on the variation of fluorescence contrast (BP) or intensity (OTV & MONO). Allele O represents the absence of the sequence while A and B the presence of the sequence.

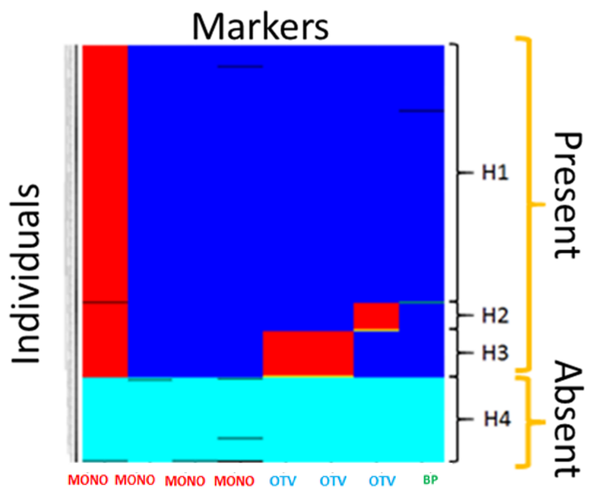
Accurate genotyping of 105k PAVs on 464 maize inbred lines

479,027 probes passed Affymetrix quality controls and were converted into robust markers by Affymetrix algorithms.

89,393 PAVs (84%) were called with an average of 5.4 probes by PAV. 38,134 PAV were called only with MONO probes.

90% of PAV genotyping were consistent between resequencing and array genotyping of 4 lines used to discover PAVs. Gap in genome assembly of 4 inbred lines could lead to technical error in expected genotyping.

Complementary information between three probe types



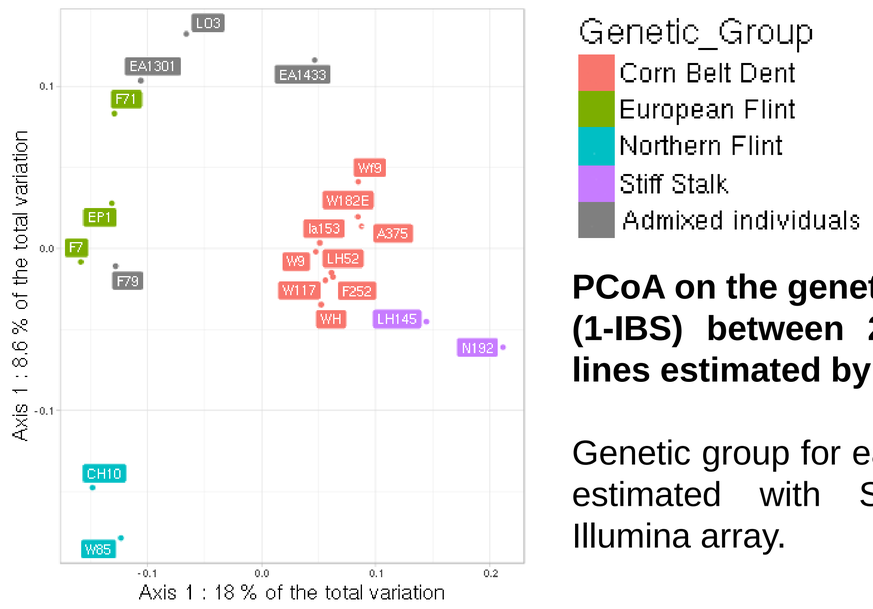
Example of one PAV genotyping for 445 inbred lines (rows) with four MONO probes, three OTV probes and one BP probe (columns).

This PAV sequence was genotyped absent for 91 individuals (O allele).

MONO and BP probes could exhibit two alleles (Absent = O, Present) and OTV probes three alleles (A, B and O). **Genotyping of these eight probes was totally consistent for presence and absence of the sequence.**

Four different haplotypes were revealed by combining the genotyping of different probes: three when the sequence is present (H1, H2 and H3) and one when the sequence is absent (H4).

Genetic structuration of a core collection of 20 temperate lines based on PAVs



PCoA on the genetic distance (1-IBS) between 20 maize inbred lines estimated by 67,479 PAVs.

Genetic group for each individual was estimated with SNPs from 50K Illumina array.

1st axis well discriminated European Flint (EF) lines from Corn Belt Dent and Stiff Stalk lines, while the 2nd axis discriminated EF and Northern Flint.

Clusters of individuals in PCoA based on **genetic distance estimated with PAVs (1-IBS) provided consistent results** with the genetic structuration estimated with SNPs.

Conclusions / Prospects

By coupling high depth sequencing with high throughput array genotyping, we efficiently and accurately genotyped **89,393 PAVs** along the maize genome. Combining different types of probe was highly valuable: (i) MONO probes allowed to genotype PAV sequences without polymorphism and without known breakpoints, (ii) BP probes allowed to genotype PAVs carrying repeats, as transposable elements, (iii) OTV probes allowed to investigate the genetic diversity within PAV sequences.

Our array will be a useful tool to better understand the role of PAVs in trait variation as well as in adaptation and genetic structuration of populations. It will also allow to test the complementation hypothesis to explain heterosis.

References

1. Lu, F. *et al.* High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications* **6,** (2015).
2. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics* **44,** 803–807 (2012).
3. Springer, N. M. *et al.* Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genetics* **5,** e1000734 (2009).
4. Hirsch, C. N. *et al.* Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize. *The Plant Cell* **28,** 2700–2714 (2016).
5. Darracq, A. *et al.* Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of PAV. *BMC Genomics* **19,** (2018).
6. Camus-Kulandaivelu, L. Maize Adaptation to Temperate Climate: Relationship Between Population Structure and Polymorphism in the Dwarf8 Gene. *Genetics* **172,** 2449–2463 (2005).

Acknowledgments

This work was supported by the ANR CNV-Maize and Amaizing programs (www.amaizing.fr). C. Mabire PhD were funded by CNV4Sel and INRA BAP department.