Maize Pan-genome estimation from next-generation sequencing data analysis

Ettore Zapparoli[1,2], Gabriele Magris[1,2], Michele Vidotto[1,2], Davide Scaglione[3], Sara Pinosio[1,2], Fabio Marroni[1,2] , Michele Morgante[1,2]

[1] Dipartimento di Scienze agroalimentari, ambientali e animali, Università di Udine, Udine, Italy - [2] Istituto di Genomica Applicata, Udine, Italy - [3] IGA Technology Services, Udine, Italy

e-mail: *ezapparoli@appliedgenomics.org*

1. INTRODUCTION

The pan-genome concept describes all the genomic sequences present in a population of different varieties of the same species and it is commonly applied in bacterial genomics [1].
The sequences present in all strains/varieties represent the "core" genome, while sequences present in some and absent in others are attributed to the "dispensable" genome and represent the variable portion of the pan-genome.
The pan-genome concept has been shown to be applicable to plants in recent years.
Structural variants (SVs) are an important source of genetic variation in plants, mostly due to large (>1000bp) insertions and deletions of transposable elements (TEs), and maize is one of the most involved species in this phenomenon. The identification of structural variants (SVs) is a strategy to characterize the dispensable genome of plants [2,3,4].
Here, we apply this strategy to characterize the maize pan-genome using 6 varieties selected from the parental lines of the MAGIC [5] maize population (A632, H99, HP301, F7, Mo17, W153R) and the reference variety B73.

2. METHODS

We Integrated paired-end mapping (PEM) and split-read mapping (SR) approaches in order to identify insertions of transposable elements (TEs) (internal algorithms) and deletions (Delly [6], GASV [7] algorithms).

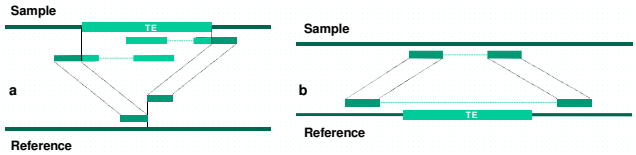


Figure 1. Schematic representation of read mapping in presence of insertions (a) and deletions (b).
(a) Insertion of a TE in the sample, which is absent in the reference. Reads from the TE (pale green) will map far from the insertion breakpoint identified by discordant reads (dark green). (b) Deletion of a TE from the sample, which is present in the reference. Reads across the TE borders will map with a larger insert size, significatively deviating from the library distribution.

3. IDENTIFICATION OF STRUCTURAL VARIANTS

Integrating paired-end mapping (PEM) and split-read mapping (SR) approaches, we obtained a wide collection of high-confidence deletions (present in the reference and absent in at least one of our varieties) and insertions (absent in the reference and present in other varieties) within varieties. Despite their technical definition, insertions and deletions may not necessarily refer to their corresponding biological meaning, as it is possible that an SV detected as a deletion may actually be an insertion in the B73 reference genome. Moreover, those results correlate well (r > 0.7) with genetic distances between each of our varieties and B73, as calculated in the previous study [5].

Table 1

Sample	A632		F7		H99		HP301		Mo17		W153R	
	Deletions	Insertions	Deletions	Insertions	Deletions	Insertions	Deletions	Insertions	Deletions	Insertions	Deletions	Insertions
Count hom.	17,265	18,421	26,808	27,422	21,386	23,843	23,202	25,091	23,833	25,954	20,212	18,850
Count het.	2,030	2,273	2,021	3,715	1,683	3,141	1,764	2,411	1,947	2,500	4,109	4,598
Length hom. (kb)	195,256	183,578	305,067	281,700	242,497	238,568	261,844	254,264	272,609	265,050	227,353	183,974
Length het. (kb)	24,627	23,803	24,066	49,204	19,985	33,525	20,987	30,091	23,223	31,835	47,994	51,667

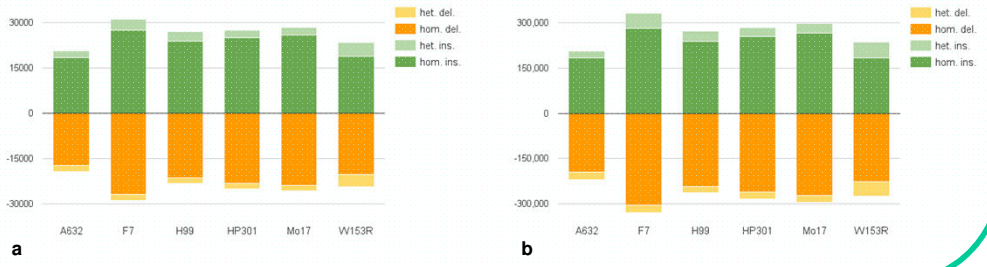
Table 1. *Structural variants distribution within varieties. Homozygous and heterozygous counts and lengths are shown.*

Figure 2.

Structural variants distribution within varieties.

***a**. Counts of insertions and deletions. Homozygous (green) and heterozygous (pale green) insertions counts are represented as positive counts; homozygous (orange) and heterozygous (pale orange) are represented as negative counts.*

***b**. Sizes of insertions and deletions, in kb. Insertions (green) are represented as positive value while deletions (orange) are represented as negative value.*



4. PAN-GENOME ESTIMATION

While the maize B73 reference genome size is around 2,5 Gb, we identified 606 Mb present in the reference and absent in at least one of our varieties (deletions), and more than 1 Gb for sequences absent in the reference and present in other varieties (insertions).

From a first genic analysis, 13517 genes on 39469 annotated genes have undergone a deletion and/or an insertion in at least one variety. In 9142 of them, SVs involved at least one exon, probably affecting gene function.

Lastly, an homology-based annotation of deletions and insertions reveals that the sequences present in the set of deletions are similar in composition to those involved in the generation of insertions. As expected, a large amount of LTR retrotransposons (RLC, RLC, RLX) was found, in line with previous evidences [8].

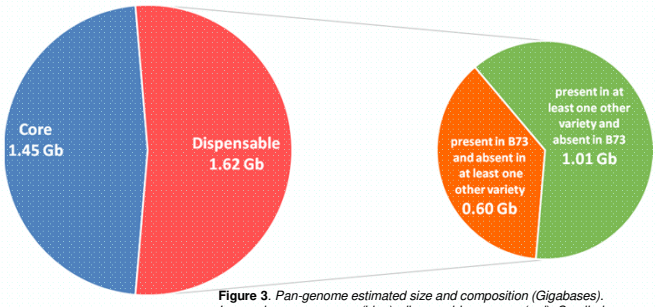


Figure 3. *Pan-genome estimated size and composition (Gigabases). Large pie: core genome (blue); dispensable genome (red). Small pie: sequences present in B73 and absent in at least one other variety (orange); sequences present in at least one other variety and absent in B73 (green).*

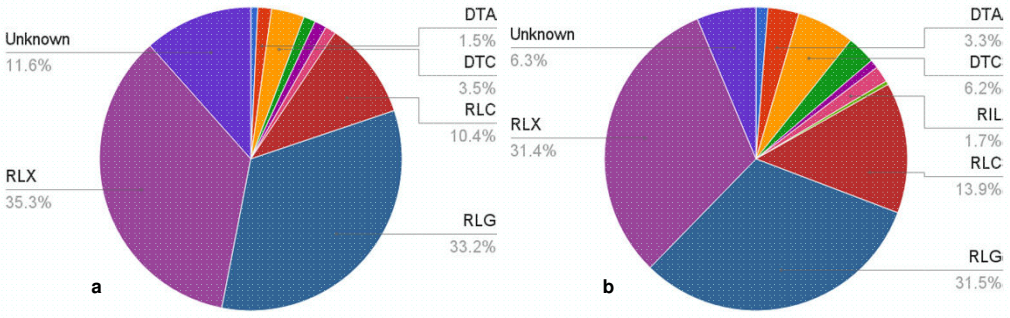


Figure 4.

*TE composition in deletions (**a**) and insertions (**b**).*

5. CONCLUSIONS AND OUTLOOKS

We identified more than 20K deletions and insertions in each of the six varieties investigated. This confirms the high structural variability of the maize genome and confirms that one reference genome is not enough for the description of the species genome.

Further efforts are underway to improve characterization of the pan-genome leveraging information obtained with de-novo assembly and with the use of longer reads. Further investigation of the genic component involved in SVs will help determine the potential phenotypic effect of the dispensable component of the pan-genome.

BIBLIOGRAPHY

[1] Tettelin et al. Genome analysis of multiple pathogenic isolates ofStreptococcus agalactiae: Implications for the microbial "pan-genome". Proc Natl Acad Sci U S A, 2005
[2] Brunner et al. Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell, 2005
[3] Morgante et al. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol, 2007
[4] Marroni et al. Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol, 2014
[5] Dell'Acqua et al. Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. Genome Biol, 2015
[6] Rausch et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics, 2012
[7] Sindi et al. A geometric approach for classification and comparison of structural variants. Bioinformatics, 2009
[8] Baucom et al. Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. PLoS Genet, 2009

Supported by: "Novel variation in plant breeding and the plant pan-genomes NOVABREED", Grant agreement no.: 294780, **European Research Council** (**ERC**)