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

Ettore Zapparoli1,2, Gabriele Magris1,2, Michele Vidotto1,2, Davide Scaglione1, Sara Pinosio1,2, Fabio Marroni1,2, Michele Morgante1,2
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 to identify insertions of transposable elements (TEs) (internal algorithms) and deletions (Delly [6], GASV [7] algorithms).

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>A632</th>
<th>F7</th>
<th>HP301</th>
<th>W153R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletions</td>
<td>17,265</td>
<td>18,421</td>
<td>26,808</td>
<td>27,422</td>
</tr>
<tr>
<td>Insertions</td>
<td>20,386</td>
<td>23,843</td>
<td>23,202</td>
<td>25,091</td>
</tr>
<tr>
<td>Length (kb)</td>
<td>195,256</td>
<td>183,578</td>
<td>305,067</td>
<td>281,700</td>
</tr>
</tbody>
</table>

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 genetic analysis, 13517 genes on 29469 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].

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 genetic analysis, 13517 genes on 29469 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].

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 genetic component involved in SVs will help determine the potential phenotypic effect of the dispensable component of the pan-genome.

BIBLIOGRAPHY

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>A632</th>
<th>F7</th>
<th>HP301</th>
<th>W153R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletions</td>
<td>17,265</td>
<td>18,421</td>
<td>26,808</td>
<td>27,422</td>
</tr>
<tr>
<td>Insertions</td>
<td>20,386</td>
<td>23,843</td>
<td>23,202</td>
<td>25,091</td>
</tr>
<tr>
<td>Length (kb)</td>
<td>195,256</td>
<td>183,578</td>
<td>305,067</td>
<td>281,700</td>
</tr>
</tbody>
</table>

Figure 4. TE composition in deletions (a) and insertions (b).

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 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 (dark green) will map far from the insertion (light 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, significantly deviating from the library distribution.

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