The GEMO project: Analysis and comparison of genomes of the fungal pathogen Magnaporthe oryzae

To cite this version:

HAL Id: hal-01000676
https://hal.science/hal-01000676
Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
The GEMO project: Analysis and comparison of genomes of the fungal pathogen *Magnaporthe oryzae*.

Guérin C.1, Mallet L.1, Kreplak J.2, Amselem J.2, Fournier E.3, A. Couloux4, C. Cruaud4 and Chiapello H.1

1 INRA, UR MIG, 78352 Jouy-en-Josas, France; 2 INRA, URGI, 78026 Versailles, France; 3 INRA, UMR BGPI, TA 54K, 34398 Montpellier, France; 4 Genoscope, Centre National de Sequençage, 2 rue Gaston Crémioux, 91 507 Evry, France.

Abstract

The GEMO project (granted by the French National Research Agency) aims at identifying genetic determinants of pathogenicity in a model fungal pathogen of rice, *Magnaporthe oryzae*. The genomes of the 8 strains of the species *M. oryzae* representing different genetic groups pathogenic of different species of plants and one strain of the sister species *M. grisea* have been sequenced using NGS technologies. We are currently performing gene annotation and comparative analyses of genomes and gene families known or expected to be involved in pathogenicity (synthesis of secondary metabolism, cell wall degrading enzymes, small secreted peptides). A dedicated database is currently being developed on the model of existing databases for eukaryotic genomes using standard open source tools. This database has been designed to facilitate further evolutionary analysis and other genomic data integration.

Genome sequence assembly results

- Strains/species: 5 rice infecting isolates of *M. oryzae* (FR13, GV11, PH14, TH12, TH16); 3 *M. oryzae* strains infecting other plants (US71 from *Setaria viridis*, BR32 from wheat, CD156 from *Elsinoe indica*); 1 *M. grisea* strain (BR29, pathogenic on Digitaria plants);
- The 9 genomes have been sequenced using 454 and Illumina/Solexa technologies and reads have been assembled by Genoscope. Statistics are presented in table 1.

<table>
<thead>
<tr>
<th>Strain</th>
<th>BR29</th>
<th>BR32</th>
<th>CD156</th>
<th>FR13</th>
<th>GV11</th>
<th>PH14</th>
<th>TH12</th>
<th>TH16</th>
<th>US71</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb contigs</td>
<td>9566</td>
<td>6094</td>
<td>26355</td>
<td>79160</td>
<td>11588</td>
<td>9908</td>
<td>11772</td>
<td>4134</td>
<td>7398</td>
</tr>
<tr>
<td>Nb scaffolds</td>
<td>169</td>
<td>111</td>
<td>237</td>
<td>2051</td>
<td>1964</td>
<td>940</td>
<td>721</td>
<td>271</td>
<td>220</td>
</tr>
<tr>
<td>Size (Mb)</td>
<td>40.9</td>
<td>41.9</td>
<td>42.7</td>
<td>43.1</td>
<td>46.1</td>
<td>48.5</td>
<td>49.8</td>
<td>39.1</td>
<td>41.2</td>
</tr>
<tr>
<td>SN</td>
<td>4.13%</td>
<td>4.9%</td>
<td>4.6%</td>
<td>22.5%</td>
<td>7.9%</td>
<td>5.8%</td>
<td>10.2%</td>
<td>5.8%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Sca. max size (Mb)</td>
<td>242.4</td>
<td>377.1</td>
<td>180.1</td>
<td>20.9</td>
<td>23.6</td>
<td>51.6</td>
<td>70.1</td>
<td>228.8</td>
<td>187.3</td>
</tr>
<tr>
<td>Sca. size (Mb)</td>
<td>2523.4</td>
<td>4783.3</td>
<td>2843.9</td>
<td>557.7</td>
<td>1102.7</td>
<td>2302.1</td>
<td>2490.7</td>
<td>3133.6</td>
<td>3111.3</td>
</tr>
<tr>
<td>Sca NOG (kb)</td>
<td>955.4</td>
<td>1760.5</td>
<td>1006.4</td>
<td>101.6</td>
<td>187.3</td>
<td>590.5</td>
<td>607.1</td>
<td>938.5</td>
<td>813.9</td>
</tr>
</tbody>
</table>

Table 1. assembly statistics of 9 Magnaporthe genomes (454 and Illumina/Solexa data assembled by Newbler 2.6)

Gene prediction was performed using the EuGene software [1] and pipelines developed at URGI ([http://urgy.versailles.inra.fr](http://urgy.versailles.inra.fr)).

- Three ab initio prediction software were used: EugeneNetMM (probabilistic models discriminating coding from non-coding sequences), SpliceMachine [2] (prediction of CDS start sites ad intron splicing sites) and FGeneSh ([http://linux1.sohberry.com/berry.phtml](http://linux1.sohberry.com/berry.phtml)) (ab initio gene predictor).
- BlastX and BlastN software were used for comparisons against Uniprot, fungal protein and ESTs databases.
- EuGene training was performed using a set of 300 genomic/coding *M. oryzae* cDNA pairs: 1/3 used for training ab initio softwares, 1/3 used for EuGene parameters weight optimization, 1/3 used for EuGene accuracy evaluation.
- Eugene accuracy (3rd dataset): Sensitivity: Gene: 81.8%, Exon: 89.4%; Specificity Gene: 81.8 %, Exon: 92.9 %.
- EuGene results obtained on a previous assembly of 3 GEMO genomes are resumed in table 2.

<table>
<thead>
<tr>
<th>Strains</th>
<th>Predicted genes</th>
<th>Unannotated DNA</th>
<th>mRNA number</th>
<th>ncRNA number</th>
<th>Short genes (CDS&lt;300 nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR29</td>
<td>12 550</td>
<td>54.1 %</td>
<td>12 454</td>
<td>96</td>
<td>1 571</td>
</tr>
<tr>
<td>BR32</td>
<td>17 281</td>
<td>63.5 %</td>
<td>17 140</td>
<td>141</td>
<td>4 419</td>
</tr>
<tr>
<td>CD156</td>
<td>15 695</td>
<td>47.5 %</td>
<td>15 660</td>
<td>35</td>
<td>3 120</td>
</tr>
</tbody>
</table>

Table 2. Eugene predictions obtained on scaffolds of BR29 *M. grisea* strain and BR32 and CD156 *M. oryzae* strains

Genome comparison

A preliminary study was performed to evaluate genomic distances between scaffolds of the 9 GEMO genomes and 3 public complete genomes.

- **Dataset**: 12 genomes = 9 GEMO genomes + 3 public genomes: *Magnaporthe oryzae* strain 7015 (V8); *Goeaumannomyces graminis* var. tritici R3-111a and *Magnaporthe poae* strain ATCC 64411 (all 3 obtained from Broad Institute, [http://www.broadinstitute.org](http://www.broadinstitute.org)).
- **Method**: (i) computations of MUM (Maximal Unique Matches) using Mummer software [3] between all pair of genomes, (ii) calculation of a matrix of MUM-based distance (1-totalMUMLength/meanGenomelength) and (iii) application of the Neighbor-Joining algorithm using the Phylip package [4] to build an unrooted tree.
- **Results**: evaluation of genomic distances between the 12 analyzed genomes. Figure 1 presents the NJ tree obtained on the 12 genomes.

Conclusion and Perspectives

- Sequence quality and assembly results are correct for the 9 GEMO genomes;
- Eugene predictions are relevant thanks to a good training set of manually curated 300 full length CDSAs/genomic pairs;
- Pairwise genome comparisons of 11 *Magnaporthe* and 1 *Gaeumannomyces* genomes reveal consistent results with the expected phylogeny of the sequenced GEMO strains.
- A dedicated resource (GEMObase) is under construction and will be used to perform genome comparisons and to detect molecular signatures of positive or purifying selection in coding and non-coding sequences.

References