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

Cyprien Guerin Guérin, Ludovic Mallet Mallet, Jonathan J. Kreplak, Joelle J. Amselem, Elisabeth E. Fournier, Arnaud Couloux, Corinne Cruaud, Helene H. Chiapello

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. Couloux[4], C. Cruaud[4] 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 Crémieux, 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, GY11, PH14, TH12, TH16) ;** 3 *M. oryzae* strains infecting other plants **(US71** from *Setaria viridis***, BR32** from wheat**, CD156** from *Eleusine indica***)** ; 1 *M. grisae* strain **(BR29,** pathogenic on Digitaria plants**);**

• **The 9 genomes** have been sequenced using 454 and Ilumina/Solexa technologies and reads have been assembled by Genoscope. Statistics are presented in table 1.

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

Strains	BR29	BR32	CD156	FR13	GY11	PH14	TH12	TH16	US71
Nb contigs	9644	6044	26535	79619	13188	9908	11772	4114	7398
Nb scaffolds	169	111	237	2051	1964	940	711	171	220
Size (Mb)	40.9	41.9	42.7	43.1	46.3	48.5	49.8	39.1	41.2
%N	4.1 %	4.9 %	6.6 %	22.5%	7.9%	5.8%	10.2%	5.8%	5.4%
Sca. avg size (kb)	242.4	377.1	180.1	20.9	23.6	51.6	70.1	228.8	187.3
Sca. max size (kb)	2523.4	4783.3	2814.9	557.7	1 102.7	2 302.1	2495.7	3133.6	3113.9
Sca N50 (kb)	955.4	1760.5	1066.4	101.6	187.3	590.5	697.1	938.5	813.9

Genome annotation

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

• **Three *ab initio* prediction softwares were used:** EugeneIMM (probabilistic models discriminating coding from non-coding sequences), SpliceMachine [2] (prediction of CDS start sites ad intron splicing sites) and FGENESH (http://linux1.softberry.com/berry.phtml) (*ab initio* gene predictor).

• **BlastX and BlastN** softwares were used for comparisons against Uniprot, fungal protein and ESTs databases.

• **EuGene training was performed using a set of 300 genomic/full-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 2. Eugene predictions obtained on scaffolds of BR29 *M. grisae* strain and BR32 and CDC156 *M. oryzae* strains

Strains	Predicted genes	Unanotated DNA	mRNA number	ncRNA number	Short genes (CDS<300 nt)
BR29	12 550	54.1 %	12 454	96	1 571
BR32	17 281	63.5 %	17 140	141	4 419
CD156	15 695	47.5 %	15 660	35	3 120

Genome comparison

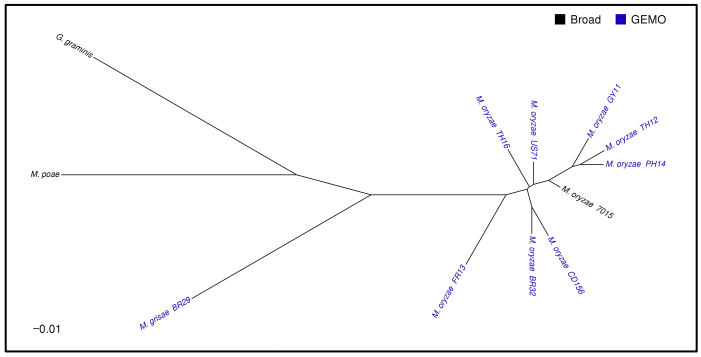
A preliminary study was performed to evaluate genomic distances between scaffolds of the 9 GEMO genomes and 3 public complete genomes.

• **Dataset : 12 genomes =** 9 GEMO genomes + 3 public genomes: *Magnaporthe oryzae* strain 7015 (v8) ; *Gaeumannomyces graminis*, var. tritici R3-111a and *Magnaporthe poae* strain ATCC 64411 (all 3 obtained from Broad Institute, http://www.broadinsitutue.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 unrooteed tree.

• **Results:** evaluation of genomic distances between the 12 analyzed genomes. Figure 1 presents the NJ tree obtained on the 12 genomes.

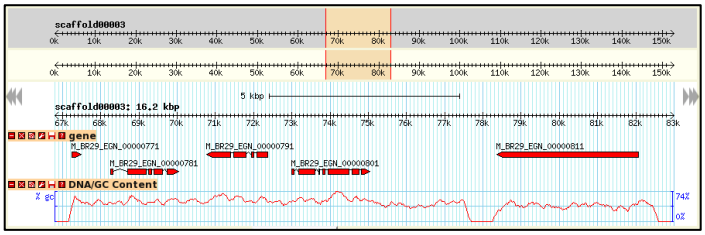
Figure 1. NJ tree obtained by comparison of MUM-based distances between the 12 pairs of genomes



The GEMObase Genome browser

The GEMObase prototype was developed using Gbrowse version 2 [5] and GMOD SeqFeature databases (one per genome). Figure 2 shows an example of Gbrowse gene visualization on a scaffold of BR29.

Figure 2. Example of GEMObase genome browser displaying Eugene annotations



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 cDNAs/genomic pairs.
• Pairwise genome comparisons of 11 *Magnaporthe* and 1 *Gaeuannomyces* 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

[1] Schiex *et al.*, 2001, LNCS 2066, Springer.
[2] Foissac *et al.*, 2008, Current Bioinform.
[3] Kurtz *et al.*, 2004, Genome Biol.
[4] Felsenstein, 1993, Phylogeny Inference Package.
[5] Stein *et al.*, 2002, Genome Res.