

The long hard road to a completed *Glomus* *intraradices* genome

***Glomus* Genome Consortium (GGC) Symposium, Nancy, France, September, 2008**

With the public release of the *Populus* genome (Tuskan *et al.*, 2006), the United States Department of Energy's Joint Genome Institute (JGI) embarked on an effort to create a community-wide genomics resource for bacterial and fungal associates of *Populus* (Martin *et al.*, 2004). Included in the list of species were several *Populus* endophytes (*Burkholderia cepacia*, *Pseudomonas putida*, *Enterobacter* spp., *Serratia proteamaculans*, *Stenotrophomonas maltophilia*; http://genome.jgi-psf.org/draft_microbes/), *Laccaria bicolor* (Martin *et al.*, 2008),

Melampsora laraci-populina (<http://www.jgi.doe.gov/sequencing/why/3088.html>) and *Glomus intraradices* (<http://www.jgi.doe.gov/sequencing/DOEMicrobes2004.html>). Ideally, the development of genomic tools for these organisms will facilitate the study of *Populus* and of its microbial associates in experimental and natural environments using whole-genome microarrays, models of predicted metabolite and protein interactions, cross-species promoter analyses and molecular surveys of community diversity. Together these resources will provide the possibility to take a holistic approach in understanding how symbionts and pathogens interact with the host tree in contrasting environments.

The production of a completely annotated and assembled *G. intraradices* genome has proven to be an especially arduous challenge and, after 4 yr of effort, it is not yet at hand. In this context, a workshop was held by the *Glomus* Genome Consortium (GGC) on September 16–17, 2008, in Nancy, France, to review the progress that has been made to date on the *Glomus* genome.

'... sequencing of G. intraradices will have a tremendous impact on the scientific community as it will give first access to so far intractable information about processes driving the biology and life cycle of AM fungal symbionts'

At the root of the Mycota Kingdom

The arbuscular mycorrhizal (AM) symbiosis between fungi in the Glomeromycota and plants involves over two-thirds of all terrestrial plant species, and is of great ecological significance (Fitter *et al.*, 2000; Rosendahl, 2008). There are around 150 described species in the Glomeromycota, and about 200 000 plant species are involved in the symbiosis. The key process in the symbiosis is the acquisition of poorly mobile orthophosphate ions from soil by the fungi, greatly enhancing plant phosphorus (P) uptake (Smith & Read, 2008). Colonization of plants by AM fungi results in a 5–20% net increase in photosynthesis (Smith & Read, 2008). Thus, AM fungi make a very large, if poorly understood, contribution to the global carbon cycling budget of ecosystems. From an evolutionary standpoint, AM fungi are unique obligate symbionts with multinuclear coenocytic hyphae that transport organelles and nutrients over long distances. The regulation of gene expression in such a system with multiple nuclei migrating long distances is completely unexplored. Furthermore, the

concept of an individual is unclear because nuclei within a single AM fungus appear to be genetically different in some species. Although this has not been directly shown in *G. intraradices*, it raises substantial questions about the natural selection and population genetics of these highly unusual organisms (Kuhn *et al.*, 2001; Hijri & Sanders, 2005). There is no known sexual cycle in AM fungi, although anastomosis and nuclear movement between hyphae has been described (Giovannetti *et al.*, 2001).

Several factors have led to *G. intraradices* being chosen for the first genome sequencing of an AM fungus. It is a widespread fungal species in that it is present in different ecosystems throughout the world, including temperate and tropical locations (Smith & Read, 2008). As a symbiont, *G. intraradices* is highly effective in mobilizing, taking up and transferring mineral nutrients other than inorganic P, including nitrogen, from soils to plants (Govindarajulu *et al.*, 2005), and it readily colonizes many plant species, including agriculturally important species such as wheat, alfalfa, rice, as well as model plants such as *Medicago truncatula*, *Lotus japonicum* and *Populus trichocarpa*. *Glomus intraradices* is one of the most commonly studied AM fungi as it colonizes host plants rapidly. It is the prime ingredient in several commercial inocula. Moreover, it can be grown *in vitro* in dual culture with transformed carrot roots and it is the only species whose spores are available commercially in pure form in large quantities. Genome sequencing of *G. intraradices* will have a tremendous impact on the scientific community as it will give the first access to so far intractable information about processes driving the biology and life cycle of AM fungal symbionts.

Genome sequencing

The first insight into the global organization of the *G. intraradices* genome was in 2004 with the publication of an estimated genome size (Hijri & Sanders, 2004), based on flow cytometry and re-association kinetics, of 14–16.5 Mb per nucleus. In addition, 600 kbp of random genomic survey sequencing (GSS), spread among 680 sequences, were made available in GenBank. Subsequently, nearly 3000 expressed sequence tags (ESTs) from *G. intraradices* have been deposited in GenBank, along with nearly 1500 ESTs from other *Glomus* species. Over the last 2 yr, the JGI has achieved much more extensive whole-shotgun sequencing (WGS) of the *G. intraradices* DAOM 197198 isolate. About 80 Mb of WGS (i.e. five times more than the expected genome size) have been sequenced using the Sanger and 454 GS-20 pyrosequencing technologies. In addition, the sequences of 29 fosmids (~1.0 Mb) from the DAOM 197198 isolate have been finished by Jane Grimwood and her colleagues at the Stanford Human Genome Sequence Center (CA, USA). As expected for AM fungi (Hosny *et al.*, 1997), the genome of *G. intraradices* has a very low G+C percentage (~30%). At the GGC workshop, Harris Shapiro (JGI, CA, USA) presented

the assemblies obtained using these WGS data sets. He used three types of analysis to estimate the effective sequence depth of the existing WGS data set: (1) coverage of the available subcloned fosmid sequences by the WGS data, (2) the fraction of ESTs not covered by the WGS data, and (3) depths of contigs from Newbler assemblies of the WGS data. Poisson curve-fits of the depth of coverage of each contig base in the assembly ranged from 1.73 ± 0.12 to 2.38 ± 0.22 . However, as the assemblies are still at a fairly low depth, it is not possible to derive a precise depth estimate from these data alone. Possible technical, informatic and biological reasons that may underlie this were discussed at length at the GGC workshop, without a clear answer emerging. One favoured possibility is that WGS assembly is hindered by the occurrence of multiple copies of many (perhaps most or all) nuclear genes, somewhat diverged in sequence – for which genomic and EST sequences provide evidence. A high level of polymorphism is evident within *G. intraradices*, so that some regions of the genome appear as different assemblies ('haplotypes') within the set of scaffolds. The presence of alternative haplotypes leads to a larger effective genome space. The number of gene/genome copies is still unknown, but this raises interesting questions about functional vs pseudogene copies, and about possibly different expression patterns for different alleles. Most software packages that are used to assemble WGS data cannot effectively segregate reads between two (even less between several) alleles. To complicate matters further, the genome is rich in short repeated sequences that increase the chance of misassemblies and are easily confused with alleles. Sequence contigs that end in repeated sequences can usually be joined with the help of higher-order information from the genome's organization. Unfortunately, the lack of a sexual cycle in *G. intraradices* has precluded the production of a meiotic genetic map.

Dr Mohamed Hijri's group (IRBV, Université de Montréal, Canada), in collaboration with Génome Québec, has carried out two 454 Genome Sequencer runs (containing a total of 113 Mb in 503 697 reads) on genomic DNA from the DAOM 197198 isolate. Finally, the University of York, UK (Peter Young *et al.*), has carried out three half 454 Genome Sequencer runs on another *G. intraradices* isolate (494) and reported on their findings. They generated 152 Mb of sequence in 687 751 reads that have been assembled using the most recent Roche DeNovo assembler, generating an assembly of 25 Mb; the largest contig is 32 kbp (a mitochondrial contig).

As we write, the novel 454 pyrosequencing data have been merged to the JGI sequences (i.e. 345 Mbp of WGS in total) and the most recent assembly is 52.5 Mb in 163 968 contigs. The depth has edged up from two times, to just more than three times the genome coverage. The lack of effective sequence depth is probably the result of a genome space that is much larger than the amount of DNA per nucleus. This space is now estimated to be > 150 Mb. Single nucleotide polymorphism variation among the genomic reads has been observed in many contigs, confirming the high polymorphism of the

G. intraradices genome. The JGI is planning to sequence 96 more fosmids by shearing, end tagging and pooling subclone DNA for a 454 FLX Titanium run. The INRA GlomusDB database (<http://mycor.nancy.inra.fr/IMGC/GlomusGenome/>) allows the GGC members access to the current sequence data.

Mitochondrial genome

The JGI, Université de Montréal and the University of York 454 data sets have allowed good assembly of the *G. intraradices* mitochondrial genome, which shows much lower levels of sequence polymorphisms than either EST or genomic sequences. Annotation of the mitochondrial genome (70 608 bp) of the *G. intraradices* 494 isolate was presented at the GGC workshop by P. Young. This genome contains a standard set of fungal mitochondrial proteins and RNA genes, introns and LAGLIDADG homing endonuclease genes.

EST sequencing and clustering

Diederik Van Tuinen (INRA-Dijon, France), Philipp Franken (Institute of Vegetable and Ornamental Crops-Grossbeeren), Helge Küster (Leibniz Universität Hannover, Germany), Yair Shachar-Hill (Michigan State University, USA) and Francis Martin (INRA-Nancy, France) updated the GGC on the current status of EST sequencing and EST databases. *Glomus intraradices* ESTs are derived from polymerase chain reaction-amplified or normalized cDNA libraries of spores germinated in the absence or presence of strigolactones/flavonoids and extraradical symbiotic mycelium (ERM). The EST data set contains 83 539 clusters (or tentative consensus sequences) assembled from 318 945 Sanger and 454 ESTs generated by JGI, New Mexico State University, Michigan State University, INRA/Toulouse University and the German MolMyk programme. The high number of EST clusters results from high EST sequence polymorphism. While functions of 7750 transcript clusters may be inferred from sequence homology to genes in other organisms, most of the predicted genes still have no known function and genes so far unique to *G. intraradices* represent a very large proportion of the total predicted transcript set. This is a surprisingly high percentage that probably reflects both the symbiotic lifestyle and the evolutionary distance between *G. intraradices* and other fungal genomes sequenced to date.

Sequences of raw ESTs and EST clusters, and their automated annotation are stored at both the INRA GlomusDB database and the MolMyk SAMS database (<http://www.cebitec.uni-bielefeld.de/legume/molmyk.html>). Additional EST sequencing on laser-dissected intraradical mycelium (IRM) and isolated arbusculated cells from colonized rice and *Medicago* roots using the Illumina-Solexa and 454 GS FLX Titanium is underway. A NimbleGen array is under construction using transcript sequences.

Unwrapping the *Glomus* genome: a formidable task ahead

Several key questions remain unanswered. Is polymorphism overall low enough to eventually get a good WGS assembly? Are certain regions of the genome more subject to polymorphism? Can we make an assembly with the rest? When looking at expression, are we seeing the real picture or are there many other variants of a given gene that are not being expressed? By mid-2009, the mycorrhizal research community may have access to more genomic information that will prove valuable for investigations of the plasticity and evolution of *G. intraradices*.

The JGI and the GGC are pleased to announce that the initial genome assembly and EST sequencing data (as of December 2008), and their automated and manual annotations, will be made available to the scientific community in January 2009 through the INRA GlomusDB database.

Acknowledgements

The *Glomus* Genome Consortium Workshop was sponsored by the Network of Excellence EVOLTREE. We would like to thank the sequencing staff at the United States Department of Energy Joint Genome Institute and Stanford Human Genome Sequencing Center for their outstanding efforts towards sequencing the *Glomus intraradices* genome. We would also like to acknowledge the members of the *Glomus* Genome Consortium for their work aiming to decipher the challenging *G. intraradices* genome. The genome sequencing and analysis of *G. intraradices* is funded by grants from the United States Department of Energy. Additional genome and EST sequencing was funded by Génome Québec, the University of York, INRA, McGill University, New Mexico University and Michigan State University.

**F. Martin^{1*}, V. Gianinazzi-Pearson², M. Hijri³,
P. Lammers⁴, N. Requena⁵, I. R. Sanders⁶,
Y. Shachar-Hill⁷, H. Shapiro⁸, G. A. Tuskan⁹
and J. P. W. Young¹⁰**

¹UMR1136 INRA-Nancy Université Interactions Arbres/Micro-organismes, IFR110, Centre de Nancy, 54280 Champenoux, France; ²UMR 1088 INRA/5184 CNRS/Burgundy University Plante-Microbe-Environnement, INRA-CMSE, BP 86510, 21065 Dijon, France; ³Institut de la Recherche en Biologie Végétale, Département de sciences biologiques, Université de Montréal, 4101 Rue Sherbrooke est, Montréal, Que., Canada H1X 2B2; ⁴Department of Chemistry and Biochemistry, New Mexico State University,

Department 3MLS, PO Box 3001, Las Cruces, NM, 88003–8001, USA; ⁵University of Karlsruhe, Institute for Applied Biosciences, Heisenberg-Group 'Plant-Fungal-Interactions', Hertzstrasse 16, D-76187 Karlsruhe, Germany; ⁶University of Lausanne, Institute of Ecology, Biology Building, Lausanne 1015, Switzerland; ⁷Michigan State University, Department of Plant Biology, East Lansing, MI 48824-1312, USA; ⁸Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA; ⁹Oak Ridge National Lab, Environmental Sciences, PO Box 2008, Oak Ridge, TN 37831, USA; ¹⁰University of York, Department of Biology, PO Box 373, York, YO10 5YW, UK
(*Author for correspondence: email fmartin@nancy.inra.fr)

References

- Fitter AH, Heinemeyer A, Staddon PL. 2000. The impact of elevated CO₂ and global climate change on arbuscular mycorrhizas: a myco-centric approach. *New Phytologist* 147: 179–187.
- Giovannetti M, Fortuna P, Citernesi AS, Morini S, Nuti MP. 2001. The occurrence of anastomosis formation and nuclear exchange in intact arbuscular mycorrhizal networks. *New Phytologist* 151: 717–724.
- Govindarajulu M, Pfeffer PE, Jin H, Abubaker J, Douds DD, Allen JW, Bucking H, Lammers PJ, Shachar-Hill Y. 2005. Nitrogen transfer in the arbuscular mycorrhizal symbiosis. *Nature* 435: 819–823.
- Hijri M, Sanders IR. 2004. The arbuscular mycorrhizal fungus *Glomus intraradices* is haploid and has a small genome size in the lower limit of eukaryotes. *Fungal Genetics and Biology* 41: 253–261.
- Hijri M, Sanders IR. 2005. Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* 433: 160–163.
- Hosny M, País de Barros J-P, Gianinazzi-Pearson V, Dulieu H. 1997. Base composition of DNA from glomalean fungi: high amounts of methylated cytosine. *Fungal Genetics and Biology* 22: 103–111.
- Kuhn G, Hijri M, Sanders IR. 2001. Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi. *Nature* 414: 745–748.
- Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V *et al.* 2008. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452: 88–92.
- Martin F, Tuskan GA, Difazio SP, Lammers P, Newcombe G, Podila GK. 2004. Symbiotic sequencing for the *Populus* mesocosm: DOE tackles the genomes of endomycorrhizal *Glomus intraradices* and ectomycorrhizal *Laccaria bicolor*. *New Phytologist* 161: 330–335.
- Rosendahl S. 2008. Communities, populations and individuals of arbuscular mycorrhizal fungi. *New Phytologist* 178: 253–266.
- Smith SE, Read DJ. 2008. *Mycorrhizal symbiosis*, 3rd edn. London: Academic Press.
- Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa*. *Science* 313: 1596–1604.

Key words: arbuscular mycorrhizal fungi, clustering, expressed sequence tag (EST), genome, *Glomus intraradices*, Joint Genome Institute (JGI), sequencing, symbiosis.