# URGI genome annotation system:an integrated system for structural and functional genome annotation

Joelle J. Amselem, Michael Alaux, Nathalie Choisne, Nicolas Lapalu, Baptiste Brault, Aminah A. Keliet, Erik Kimmel, Françoise Alfama, Sandie Arnoux, Marc Bras, et al.

# URGI Genome Annotation System: an Integrated System for Structural and Functional Genome Annotation

Joelle AMSELEM[1,4], Michael ALAUX[1], Nathalie CHOISNE[1], Nicolas LAPALU[1,4], Baptiste BRAULT[1,4], Aminah KELIET[1], Erik KIMMEL[1], Françoise ALFAMA[1], Sandie ARNOUX[1], Marc BRAS[1], Laetitia BRIGITTE[1], Olivier INIZAN[1], Véronique JAMILLOUX[1], Jonathan KREPLAK[1], Fabrice LEGEAI[2], Isabelle LUYTEN[1], Cyril POMMIER[1], Sébastien REBOUX[1], Stéphanie SIDIBE-BOCS[3], Marc-Henri LEBRUN[4], Delphine STEINBACH[1] and Hadi QUESNEVILLE

[1] Unité de Recherche en Génomique-Info, UR1164 INRA, Route de Saint-Cyr, 78026, Versailles cedex, France
joelle.amselem@versailles.inra.fr

[2] UMR BIO3P INRA IRISA, Domaine de la Motte, BP35327, 35653, Le Rheu cedex , France

[3] UMR DAP CIRAD, TA A-96 / 03, Av Agropolis, 34398, Montpellier cedex 5, France

[4] Biologie et gestion des Risques en agriculture, UMR1290 INRA Agro-ParisTech, Av Lucien Brétignières, BP01, 78850, Thiverval-Grignon, France

**Abstract** *The URGI platform (http://urgi.versailles.inra.fr) develops a genome annotation system dedicated to plants and their pathogens. This Integrated System relies on: (i) pipelines for Transposable Elements annotation (REPET) and gene structural and functional predictions (ii) databases and user-friendly interfaces to browse and query the data (URGI Information System GnpIS, Genome Report System GRS), (iii) A distributed annotation system for curation of gene structure.*

**Keywords** genome annotation, pipelines, databases, interfaces, genes, transposable elements, plants, fungi.

## 1 Introduction

The INRA URGI (Unité de Recherche en Génomique-Info) develops and maintains an information system for plant and pathogens genomes. This system is used in number of national and international collaborative projects involving biologists and bioinformaticians. Nowadays, the recent development of new generations of sequencing tools leads to a spectacular increase of the number of sequenced genomes. But, genome annotation has difficulties to follows this pace, introducing a lack between the release of genome sequences and their annotations. To face this problem, the URGI develops and provids tools to annotate entirely sequenced genome (pipeline, databases, and interfaces).

## 2 The URGI Annotation System

The URGI annotation system relies on three components: pipelines, databases and interfaces.

### 2.1 Pipelines

- A transposable element detection and annotation package, called REPET [1,2] is composed of two pipelines: TEdenovo and TEannot. Thanks to their high level of automation and accuracy, they were used within many international genome projects concerning plants, fungi and insects.

- A gene prediction pipeline, based on *ab initio* and similarity gene finding softwares. It uses the EuGene program to integrate all sources of information [3].

- A functional annotation pipeline, based on (i) various methods of patterns matching and motifs recognition, (ii) intracellular targeting prediction methods, and (iii) comparative genomics with other fungal genomes.

### 2.2 Databases

Our database component relies on the well known schemas from the GMOD consortium (http://gmod.org). All annotation features and analysis results are primarily stored in the Chado or Bio ::SeqFeature ::Store schema according to the need (speed access or genericity). Data can then be searched through GnpIS QuickSearch (http://urgi.versailles.inra.fr/gnpis) and Biomart (GMOD).The GnpIS QuickSearch is based on the Apache Lucene™ full-featured text search engine library. Indexes are generated to query structural or functional data stored in same or separate DBs. Query results are returned according to significance with terms, and linked to GnpIS modules and/or Genome Report System (GRS). BioMart based datamarts were used as an advance search tool. Results of complex search criteria could be exported in different formats or directly send to Galaxy (http://main.g2.bx.psu.edu/) for further bioinformatic analysis.

### 2.3 Interfaces

We provide textual or graphical interfaces over the databases. We use GBrowse as graphical interface to display sequence annotations. Apollo or Artemis are used for gene structure curation shared by a community, as they are committed in the database using “pure JDBC” direct communication protocol between Apollo (or Artemis) and Chado. The Genome Report System GRS was developed (in Java) in the frame of the ANR GnpAnnot project. It provides comprehensive categories of reports through a user-friendly textual interface over structural and functional genomic data stored in Chado databases.GRS also proposes a Gene Ontology browser and an editing module (GRE) to allow manual functional curations.

### 2.4 Conclusions and Perspectives

The platform was chosen by the international grapevine consortium (IGGP) to manage grapevine genomic annotations and to help the community to perform the manual gene annotation. It also hosts wheat genomic and genetic data for the International wheat scientific community (IWGSC). It is used for the annotation of the first wheat chromosome (3B) sequence. The integrated genome annotation system was also successfully used for fungal genomes as *Botrytis cinerea* T4 (grey mould disease) and *Leptosphaeria maculans* (stem canker) [5] in the frame of their genome consortium for sequencing and annotation. Portal for the different plant and fungi species are available at http://urgi.versailles.inra.fr/index.php/Species.

Data integration of sequences from the next generation sequencing technologies is a new scientific challenge in bioinformatics. To face this challenge, evolution of GnpIS architecture is in progress: evolution of DB schemas and interfaces, new datamarts and galaxy workflow manager based pipelines to mine data.

## Acknowledgements

We ackowledge :
- URGI Information System, pipelines and data “Agile” development teams.
- ANR for the funding of GnpAnnot and GnpInteGr projects.

## References

[1] T. Flutre, E. Duprat, C. Feuillet and H. Quesneville, Considering transposable element diversification in de novo annotation approaches. *PLoS ONE* 6(1):e16526, 2011

[2] H. Quesneville, C.M. Bergman, O. Andrieu, D. Autard and D. Nouaud, Combined evidence annotation of transposable elements in genome sequences. *PLoS comput. Biol.,* 1(2):e22, 2005

[3] S. Foissac, J. Gouzy, S. Rombauts, C. Mathé, J. Amselem, Y. Van de Peer, P. Rouzé and T. Schiex, Genome Annotation in Plants and fungi : Eugene as a model platform. *Current Bioinformatics*, 3(2):87-97, 2008

[4] T. Rouxel, J. Grandaubert, J. Hane, C. Hoede, A. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally, S. Bourras, A. Cozijnsen, L. Ciuffetti, A. Degrave, A. Dilmaghani, L. Duret, I. Fudal, S. Goodwin, L. Gout, N. Glaser, J. Linglin, G.H. Kema, N. Lapalu, C. Lawrence, K. May, M. Meyer, B. Ollivier, J. Poulain, C. Schoch, A. Simon, J. Spatafora, A. Stachowiak, B.G. Turgeon, B. Tyler, D. Vincent, J. Weissenbach, J. Amselem, H. Quesneville, R. Oliver, P. Wincker, M.H. Balesdent and B. Howlett, Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat. Comms.*, 2:202, 2011