Fungal species resources: an integrated system for structural and functional genome annotation

Nicolas Lapalu[12], Joelle Amselem[12], Baptiste Brault[1], Laetitia Brigitte[12], Jonathan Kreplak[12], Françoise Alfama[1], Aminah Keliet[1], Erik Kimmel[1], Isabelle Luyten[1], Sébastien Reboux[1], Delphine Steinbach[1], Marc-Henri Lebrun[2] and Hadi Quesneville[1]

[1]INRA-URGI, Versailles, France, [2]INRA-BIOGER, Thiverval grignon, France;

Contact : nicolas.lapalu@versailles.inra.fr, joelle.amselem@versailles.inra.fr

Abstract

With the recent development of new generation sequencing technologies, a large number of genomes are sequenced leading to an increase in derived genomic data. This huge amount of data has to be stored, analyzed and searched (gene families, SNPs …). To face this challenge, the URGI platform (http://urgi.versailles.inra.fr) provides tools to store and annotate entirely sequenced genome such as annotation pipelines, databases and user-friendly interfaces to browse and query the data.
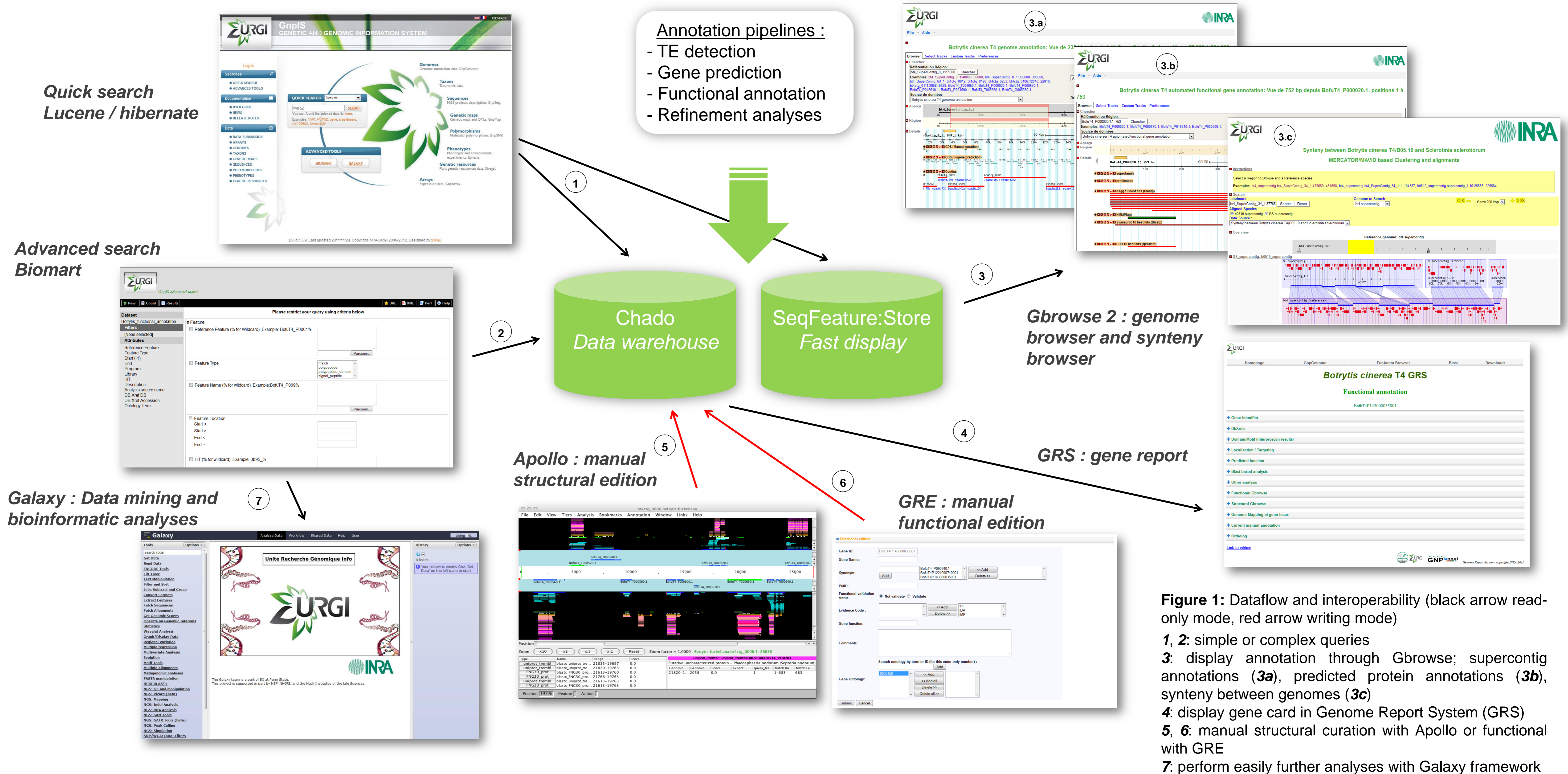
Annotation pipelines :
- TE detection
- Gene prediction
- Functional annotation
- Refinement analyses

Quick search
Lucene / hibernate

Advanced search
Biomart

Chado
Data warehouse

SeqFeature:Store
Fast display

Gbrowse 2 : genome browser and synteny browser

GRS : gene report

Apollo : manual structural edition

GRE : manual functional edition

Galaxy : Data mining and bioinformatic analyses

Figure 1: Dataflow and interoperability (black arrow read-only mode, red arrow writing mode)

1, 2: simple or complex queries
3: display annotation through Gbrowse; supercontig annotations (*3a*), predicted protein annotations (*3b*), synteny between genomes (*3c*)
4: display gene card in Genome Report System (GRS)
5, 6: manual structural curation with Apollo or functional with GRE
7: perform easily further analyses with Galaxy framework

Query, display and analyze

All the analysis and annotation data are stored in two different database schemas to insure performance in display (SeqFeature:Store schema) and data persistence (Chado schema).

Users could request the system in two ways. First, a simple "google-like" form (fig 1, 1), which explores database indexes and returns results according to significance with term criteria. Second, a Biomart instance, rich and complex web forms, used as an advanced search on user-fitted datamarts (fig 1, 2). Both provide links to Genome browsers (Gbrowse) or Genome Report System (GRS).

Results of BioMart could be exported in different formats or directly send to a Galaxy framework for further bioinformatic analysis (fig 1, 7).

Gene structure & gene function curation

The system allows curation of gene models with the GMOD tool, "Apollo" (fig 1, 5).

We also developed a functional edition form, Genome Report Edition (GRE), as a module of the GRS (fig 1, 6). This allows addition of supplementary data as Gene Ontology terms or gene function, that add value to gene annotation before submission to public sequence databases.

Curated data are shared by the consortium community as soon as they are committed in the database.

The Genome Report System (GRS): an interface to display annotation focused on gene

The GRS was developed to produce various and user-friendly web interfaces (fig 2). GRS uses structural and functional data stored in Chado databases to provide users with comprehensive categories of reports (functional domains, orthologous genes, blast hits). The GRS implements a Gene Ontology tree showing genes sharing same GO terms. It is also useful with the GRE module to associate GO terms with gene model (fig 1, 6).

Figure 2: Genome Report System screenshot
1: gene locus and sequence
2: cellular localization (SignalP, TargetP)
3: Domains/motifs
4: orthology analyses, link to other species (ortholog genes)
5: gene ontology tree and cross references with all genes sharing this GO.

Conclusion and Perspectives

The integrated genome annotation system was successfully set up for fungal genome annotation *Botrytis cinerea T4* (grey mould disease) **[1]** and *Leptosphaeria maculans* (stem canker) **[2]** in the frame of their whole genome annotation and in progress for *Venturia inaequalis* and *Microbotryum violaceum*.

[1] Amselem J et al, *Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea*. PLoS Genet. 2011 Aug;7(8)
[2] Rouxel T et al, *Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations*. Nat Commun. 2011 Feb 15