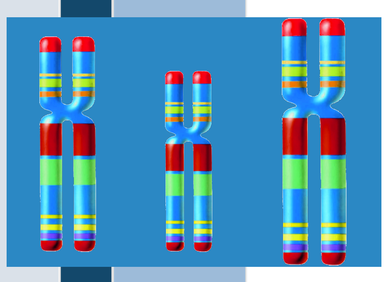
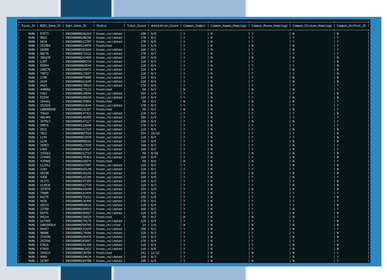
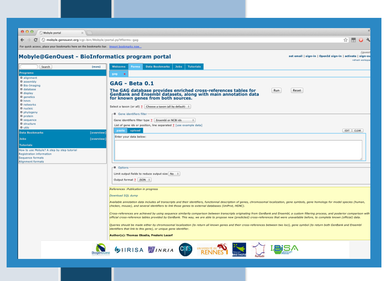
The GAG database: A new resource to gather genomic annotation cross-references

Thomas Obadia[1,2,4], Olivier Sallou[3], Marion Ouédraogo[1,2], Grégory Guernec[1,2,5], and Frédéric Lecerf[1,2]

[1] INRA, UMR1348 PEGASE, F-35000 Rennes, France
[2] Agrocampus OUEST, UMR1348 PEGASE, F-35000 Rennes, France
[3] GenOuest Platform, INRIA/Irisa – Campus de Beaulieu, F-35042 Rennes Cedex, France
[4] Present address: INSERM, UMR S 707, Paris, France
[5] Present address: INSERM, UMR 1027, F-31000, Toulouse, France





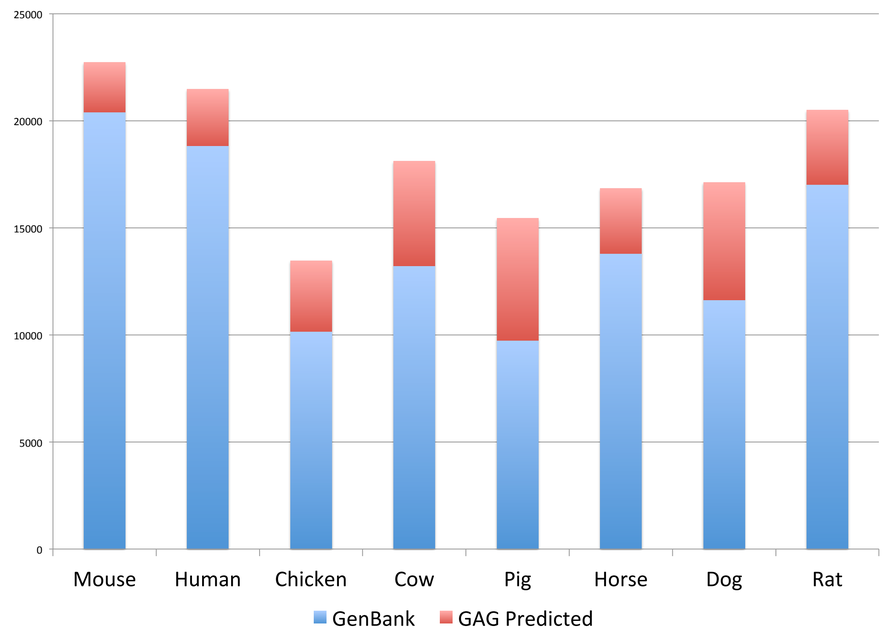


Background

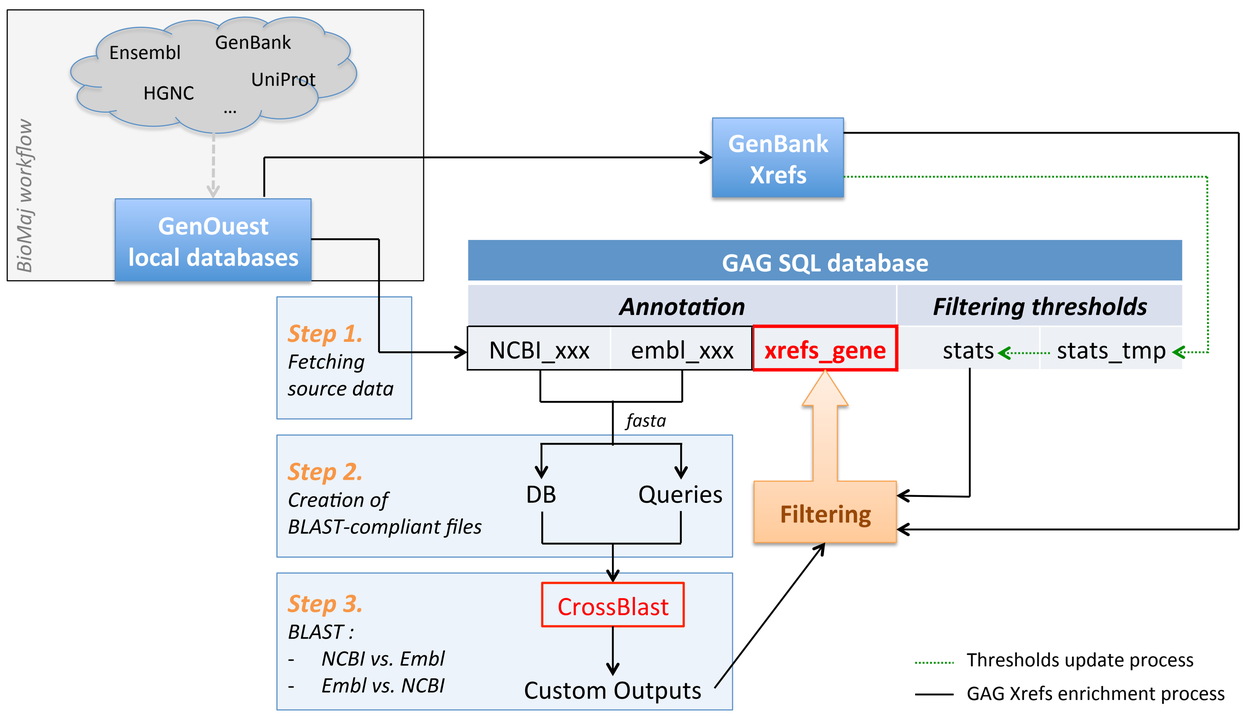
High-throughput sequencing technologies now allow access to hundreds of gigabytes of raw and processed genomics data. A key point in linking sequences to biological function is the use of annotation data from reference genomes. NCBI GenBank and EBI Ensembl databases are two of the main sources to provide researchers with genomic annotation and links to external resources, such as proteomics (*via* UniProt/SwissProt) and unified gene names (*via* HGNC). Although cross-references are generated on a regular basis by these foundations, we have discovered that available data differ slightly and are fragmented, sometimes being either complementary or redundant.

Objectives

Since data fragmentation quickly becomes a time-consuming matter when one needs to gather annotation for a large set of genes, we have developed an automated process that generates enriched lists of cross-references for the 8 included species. The process is conducted on the set of coding RNA. Functionnal annotation is thus quickly available by simple SQL queries, ran by a user-friendly interface using the Mobyle framework.



The GAG process



Applications

The GAG database allows for an average increase of 28.4% when comparing our filtered results with official GenBank cross-references tables. Results range from 5.5% (human) to 57.7% (pig).

To illustrate a concrete example of enriched cross-references, we present below the case of NCBI genes for which no official cross-reference is provided, and the newly-discovered « GAG » cross-references.

GenBank ID	Ensembl ID	Common description ?	Common homologs ?	Common symbol ?	Common outgoing references ?
4888 *NPY6R*	ENSG00000226306 *NPY6R*	Y	N	Y	Y
643332 *ECRP*	ENSG00000136315 *RP11-84C10.2*	N	N	N	Y
10140 *TOB1*	ENSG00000141232 *TOB1, TOB, TROB, TROB1*	Y	Y	Y	Y

Conclusion

GAG is a complete process for generating enriched cross-references tables based upon sequence similarity comparison and filtering of results. Although exhaustive cross-references are impossible to achieve, we provide a new database that can be used as a solid new basis for further bioinformatics development.

The GAG database has been deployed on the GenOuest platform and allows for querying through a webservice available at **http://gag.genouest.org**, and the content is kept up to date by an automated process.

13e Journées Ouvertes en Biologie, Informatique et Mathématiques ■ Rennes, 3 - 6 juillet 2012

INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE
CENTRE de Rennes • PEGASE
65, rue de Saint-Brieuc • CS 84215 35042 Rennes cedex • Tél : +33(0) 223 485 464 •
Courriel : lecerf@agrocampus-ouest.fr • http://www.rennes.inra.fr • http://www.agrocampus-ouest.fr