SymbAphidBase: a new database dedicated to aphid symbionts to store novel sequenced genomes and standardize their annotations

Mathieu Labernardiere, Patrice Baa-Puyoulet, Jean-Pierre Gauthier, Gérard Febvay, Federica Calevro, Yvan Rahbé, Hubert Charles, Jean-Christophe Simon, Stefano Colella

SymbAphidBase: a new database dedicated to aphid symbionts to store novel sequenced genomes and standardize their annotations.

Mathieu Labernardiere[1], Patrice Baa-Puyoulet[1], Jean-Pierre Gauthier[2], Gérard Febvay[1], Federica Calevro[1], Yvan Rahbé[1,3], Hubert Charles[1,3], Jean-Christophe Simon[2] and Stefano Colella[1,3]

[1] UMR203 BF2I, Biologie Fonctionnelle Insectes et Interaction, INRA, INSA-Lyon, Université de Lyon, F-69621 Villeurbanne, France

{Mathieu.Labernardiere, Patrice.Baa-Puyoulet, Gerard.Febvay, Yvan.Rahbe, Stefano.Colella}@lyon.inra.fr; {Federica Calevro, Hubert.Charles}@insa-lyon.fr

[2] Institute of Genetics, Environment and Plant Protection (IGEPP), INRA, UMR 1349, Domaine de la Motte, BP 35327, 35653, Le Rheu Cedex, France

{Jean-Pierre.Gauthier, Jean-Christope.Simon}@rennes.inra.fr

[3] Bamboo, INRIA Rhône-Alpes, France

Complete sequences of bacterial genomes are accumulating with an unprecedented speed due to the democratization of NGS technologies. This is also true for symbiotic bacteria and genomic comparisons are key to understand their contribution to host biology. We developed SymbAphidBase [1]: an *ad hoc* genome database to store and analyse aphid symbionts' genome sequences. Aphids harbour an obligate primary endosymbiont, *Buchnera aphidicola,* and several facultative secondary symbionts. SymbAphidBase is designed to integrate data from all these bacteria. At present it includes the sequenced genomes of 17 strains of *B. aphidicola* from 8 different aphid species available in GenBank. To implement this database we used the GMOD's tools: the chado database to store the genomic data and annotations, coupled with the JBrowse genome browser. SymbAphidBase includes an interface that gives access to data in different formats: a genome browser, a Blast server, comparative genes/proteins statistics and downloadable files. From the beginning of the project, the need to generate a unified gene annotation and identification scheme was apparent. In fact, if we were to use the original gene functional annotations and names, often a small fraction of genes would be found to be common in the different *B. aphidicola* genomes when performing pairwise comparisons (as low as 10%). In light of these results, we decided to re-annotate the genomes using EuGene-P, a prokaryotic gene finder tool. The genes are later re-annotated (or annotated for the new genome sequences) using a Blastx analysis against the HAMAP [2] protein database that includes 10 highly curated *B. aphidicola* genomes. The final assignment of gene names is prioritized in a filtered pipeline to include the SwissProt or TrEmbl IDs when available with variable homology criteria (that are registered in the new gene ID). With this approach we are able to increase the number of common genes when performing pairwise comparisons among *B. aphidicola* genomes (40-99% with our method depending on the chosen parameters). For genes that do not get a name and functional annotation using this automated method, we are working on other approaches that would use phylogeny and/or expert manual annotation. Beyond this novel unified gene annotation, to facilitate the direct comparison of different genomes, we implemented a double browser interface to facilitate the contemporary visualization of two genomes at the same time. All these database generation steps are automated with specific pipelines developed using mainly Perl, PHP, and jQuery languages. In conclusion, SymbAphidBase is a companion database to AphidBase [3] (the aphid genome database) to facilitate genomic data storage and analysis to study symbiosis in the aphid model.

[1] http://symbaphidbase.cycadsys.org/; [2] http://hamap.expasy.org/;
[3] http://www.aphidbase.com/aphidbase/

Keywords: genome database; genome annotation; aphid symbionts; SymbAphidBase