Annotating Genomes from Eukaryotes besides Human

Pierre Rouzé[1,2], Stephane Rombauts[1], Lieven Sterck[1], Yao-Cheng Lin[1], Jeffrey Fawcett[1], Steven Robbens[1], Cindy Martens[1] and Yves Van de Peer[1]

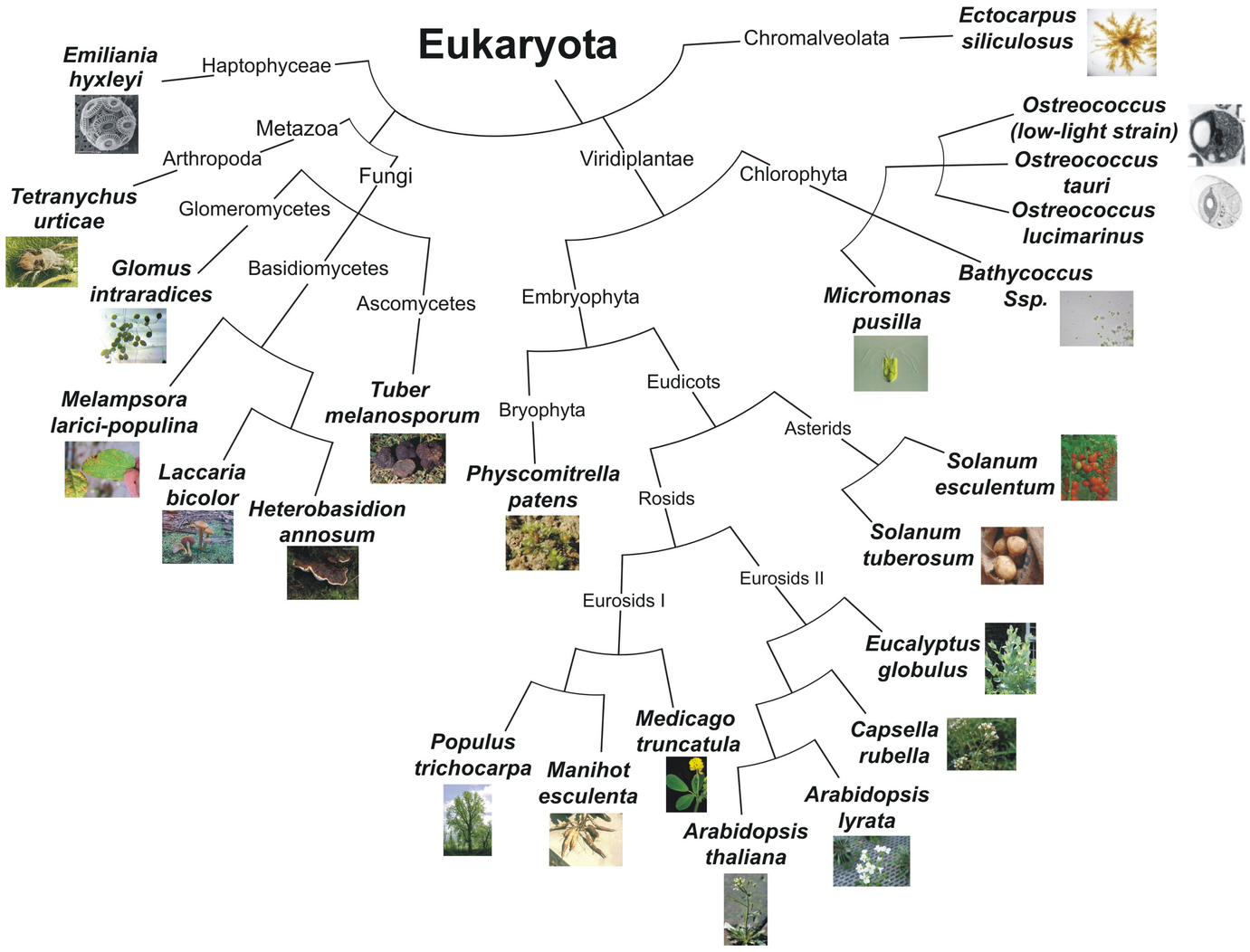
1 BioInformatics & Evolutionary Biology, Department of Plant Systems Biology, Gent University, VIB Gent University, Technologiepark 927, B-9052 GENT, Belgium
2 Laboratoire associé de l'INRA (France)
E-mail: (pierre.rouze, yvpee)@psb.ugent.be

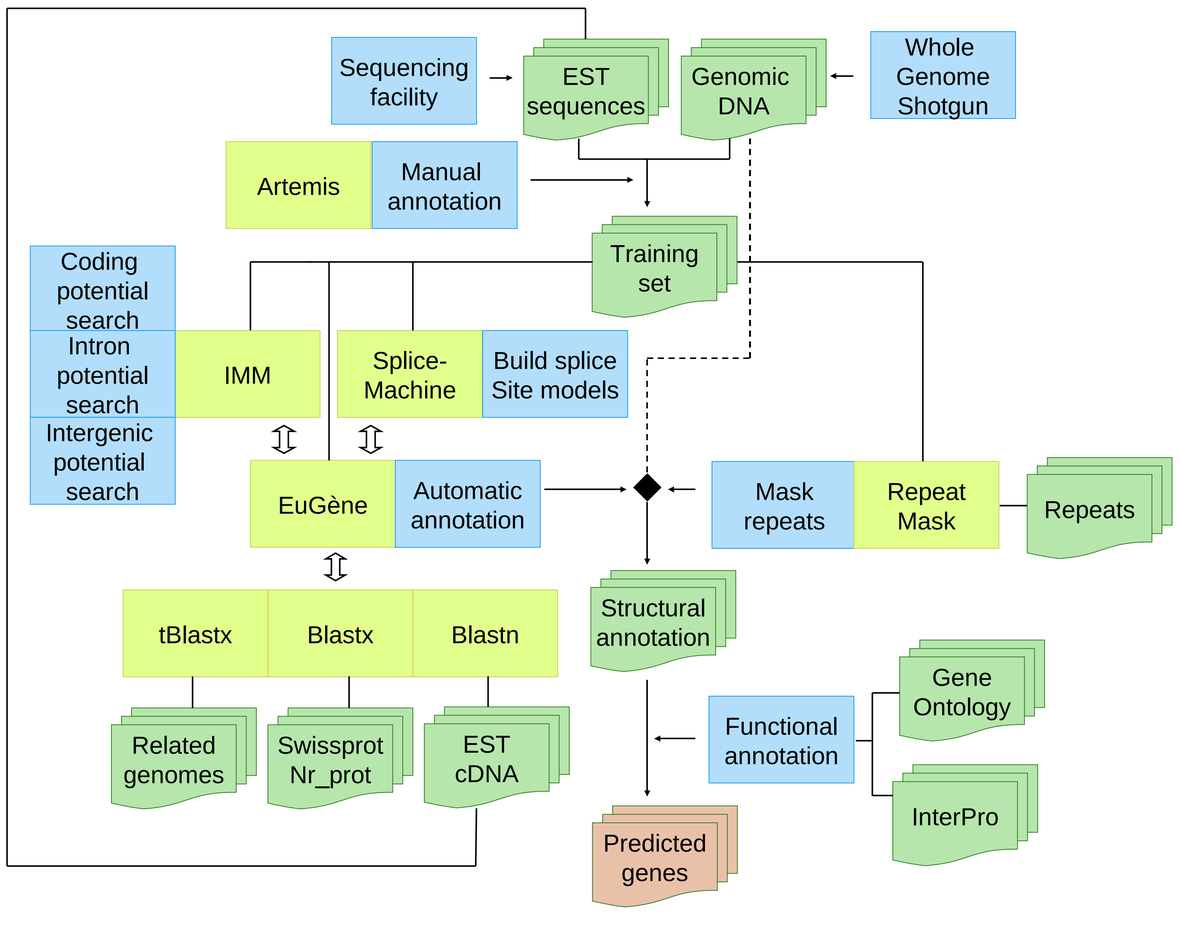
Genome annotation @ beg.psb.gent

If the annotation of the human genome is seen as an important issue for humans, the annotation of genomes from a variety of eukaryotic species which matter for humans, either as pathogens, or as providers of safe food and environment and other human needs, or simply as building stones of their life history is probably as important.

Our team has been developing tools for genome modeling and has been involved in many large scale annotation projects since 2000. Besides *Arabidopsis thaliana*, these annotations have been targeted at species that were not model organisms, ranging from higher plants to moss and green algae in the green lineage, symbiotic and pathogen fungi associated with plants, animal plant-grazing pest, as well as several photosynthetic marine unicellular and multicellular organisms.

Because of the wide distribution of target species in term of phylogeny, in term of resources available and in term of genome style led us to encounter and solve issues that are often ignored or neglected in genome annotation projects





So what ?

Genome annotation is increasingly an automatic process done by a few dedicated teams around the world. It is then important that this process is not stuck to a dominant model gained from the study of a few model organisms (and evaluations only based on it, as NGASP) but is thought having in mind the diversity in genome structures and styles and the possible occurrence of clade-speficic features.

Foissac S et al. (2008) Genome Annotation in Plants and Fungi: EuGene as a model platform. Current Bioinformatics (in press)
Martin F et al. (2008) The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature 452, 88-92
Rensing SA et al. (2008) The genome of the moss Physcomitrella patens reveals evolutionary insights into the conquest of land by plants. Science 319, 64-9
Velasco R et al. (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLOS One 2, e1326.
Palenik B et al. (2007) The Tiny Eukaryote Ostreococcus Provides Genomic Insights Into The Paradox Of Plankton Speciation. Proc. Natl. Acad. Sci. USA 104, 7705-10.
Cannon SB et al. (2006) Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proc. Natl. Acad. Sci. USA 103, 14959-64
Tuskan G et al. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray ex Brayshaw). Science 313, 1596-604.
Derelle E et al. (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc. Natl. Acad. Sci. USA 103, 11647-52

The last genome is not like the previous one

For gene modeling we make use of Eugène, an integrative gene prediction and modeling platform allowing and weighting contribution from ab initio predictions and homology-based searches using independently protein, EST and genomic sequence data.

If the general scheme is the same, in practice the modules we use differ as well as the confidence weight we can put on each. This is due to well known data availability but also to less valued variation in genome style

Data availability

Depending on practical tractability and resources available for the project the amount of ESTs vary from one project to the next, which sometimes ESTs coming from different strains or related species. For an organism from a less explored clade, protein similarity often end up in many orphan genes and no genome will be close enough for an useful comparison.

Genome style

Each genome has specific features that should first be delineated and properly be taken into account to decide on the choice of modules to be integrated, on their training and sometimes on specific developments (e.g. size and content of intergenic regions, existence and nature of transposable elements, intron number, size distribution and sites). Introns are typical examples, being biased in composition in plants and many other organisms, but not in vertebrates. Splice prediction tools developed having animals as main targets are often not able to capture the features of plant genes, and will have low performance whatever the training.
Genome modeling also assumes that the style is uniform for the whole genome. We had cases where this does not stand true, such as for prasinophytes with 2 chromosomes differing from the all the others.