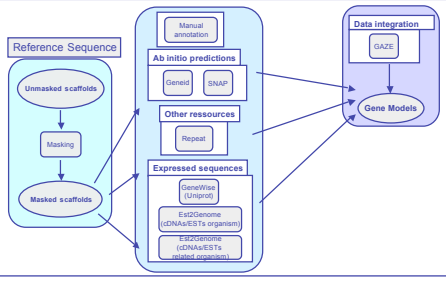
THE EUKARYOTE GENOME ANNOTATION PLATFORM AT GENOSCOPE

Betina M. Porcel, Franck Aniere, Sylvain Bonneval, Benjamin Noel, Jean-Marc Aury, Corinne Da Silva, Olivier Jaillon, France Denoeud, Claude Scarpelli, Jean Weissenbach, Patrick Wincker and François Artiguenave.

Genoscope, Institut de Génomique, CEA, 2 Rue Gaston Crémieux, CP 5706, 91057 EVRY cedex

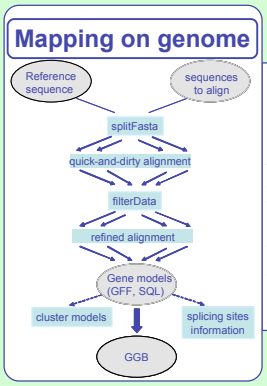
The number of annotation projects for the eukaryote genomes sequenced at Genoscope increases constantly. To answer the scaling-up problems, we have partially automated the genome annotation process, allowing today the annotation of 2 to 4 genomes per year.

Annotation workflow: use of data collections



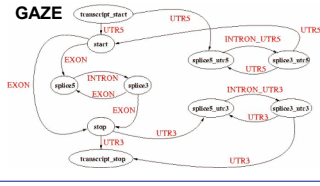
The annotation protocol is divided into three sequential steps: the first one consists on masking the genome sequence thanks to *ab initio* tools or repeat sequence libraries. During the second step, all collected resources are mapped on the genome: *ab initio* gene model predictions (already trained on manually annotated genes) and homology searches, using collections of expressed sequences - full length cDNAs, ESTs or massive-scale mRNA sequences from the same or closely related organisms – proteins or other genomic sequences. After a final integration of all gene evidence using GAZE (1), the final proteome is delivered with computed annotation data, such as ortholog and paralog associations, functional domains and ontology.

Mapping on genome



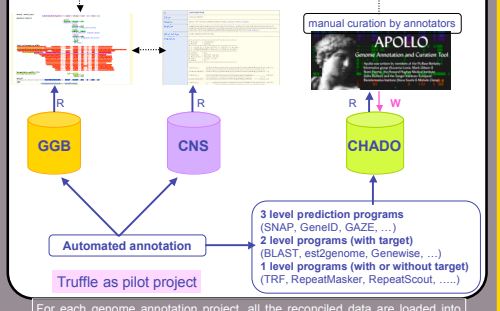
The mapping on the genome of the sequences used for the annotation has been automated to optimize the annotation workflow. After splitting both the reference genome and the data collections, a first raw alignement is performed to define the genomic region. After filtering the results to identify a single locus for each sequence, the alignement is refined for the identification of exon-intron boundaries.

Data integration and annotation

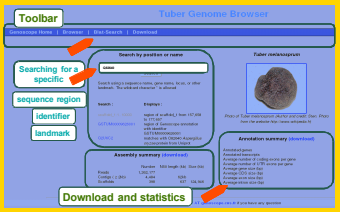


- Global method of gene prediction data integration based on dynamic programming (R. Durbin, K. Howe & T. Chothia)
- Use of two data structures : signals (start, stop,…) and segments (exons, conserved regions,…)
- Based on a graph which describes the gene structure : genomic region limited between two signals (i.e. between a stop and a start, it is an intergenic region)
- A ratio could be associated at each segment and signals to choose the best structure along of a genomic sequence

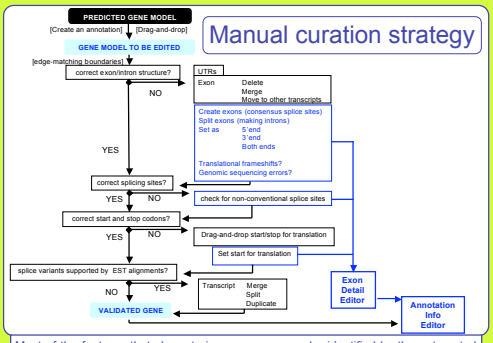
Eukaryotic Annotation Platform



For each genome annotation project, all the reconciled data are loaded into dedicated databases which are connected to a genome browser accessible by the web. We now provide collaborators carrying sequencing projects with a distributed annotation platform allowing expert evaluation of the annotation, in addition to our automated gene prediction pipeline.

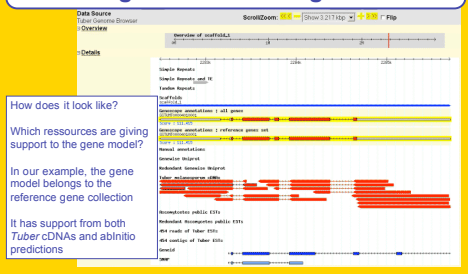


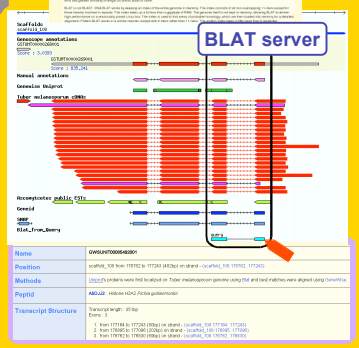
Manual curation strategy



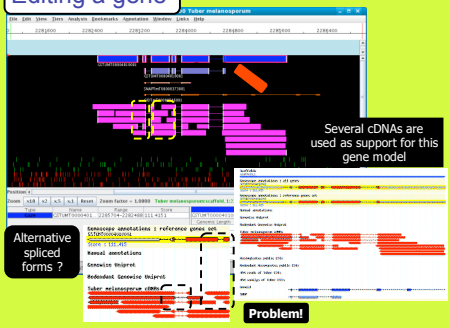
Most of the features that characterize a genome can be identified by the automated procedure. However, gene models and annotation can still be improved by human manual annotation to find possible errors or to resolve incongruous evidence on the automatic annotation of the genome.

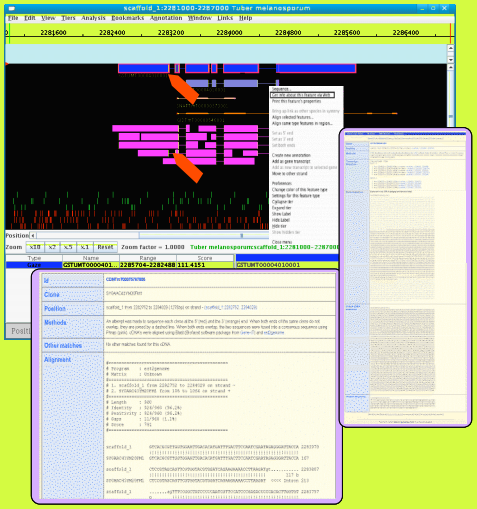
The Genoscope Gbrowse portal: searching for the cuttest gene ever …

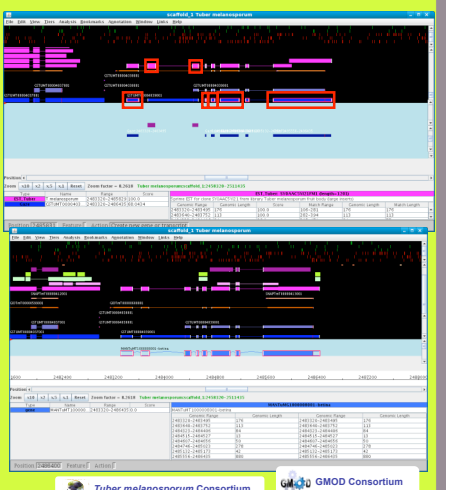




Editing a gene







Standard cases	predicted gene models that are mostly correct.
Tricky cases	predicted gene models that should be split or merged
	genes that are not in the gene collection
	genes that cross scaffolds

To ensure at most the participation of the scientific community, an annotation tool for revising annotations has been set up using components of the Generic Model Organism Database (GMOD) toolkit, which provides tools for managing organism databases. A CHADO database, linked to an Apollo graphical interface, permit users to correct gene structures and store them in a dedicated organism database, as we show on a few examples using the truffle genome as a pilot project.

Tuber melanosporum Consortium

GMOD Consortium