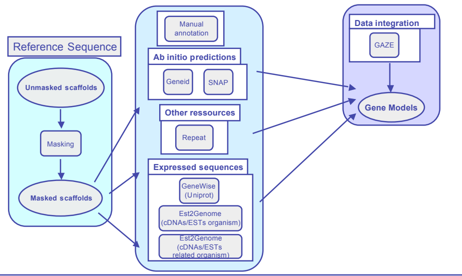
THE EUKARYOTE GENOME ANNOTATION PLATFORM AT GENOSCOPE

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

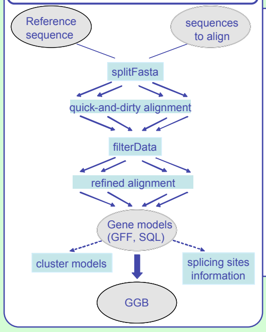
The number of annotation projects for the eukaryote genomes sequenced at Genoscope increases constantly. To answer the scaling-up problems, we have partially automated the genome annotation process, allowing today the annotation of 2 to 4 genomes per year.

Annotation workflow: use of data collections



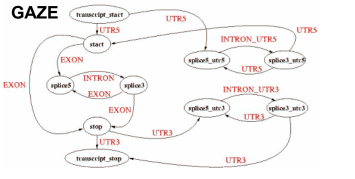
The annotation protocol is divided into three sequential steps: the first one consists on masking the genome sequence thanks to *ab initio* tools or repeat sequence libraries. During the second step, all collected resources are mapped on the genome: *ab initio* gene model predictions (already trained 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



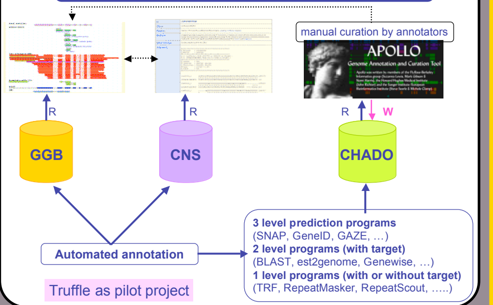
The mapping on the genome of the sequences used for the annotation has been automated to optimize the annotation workflow. After splitting both the reference genome and the data collections, a first raw alignement is performed to define the genomic region. After filtering the results to identify a single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

Data integration and annotation



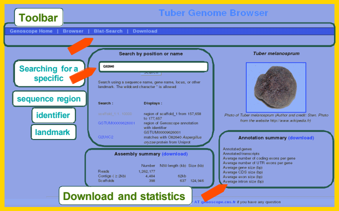
- Global method of gene prediction data integration based on dynamic programming (R. Durbin, K. Howe & T. Chothia)
- Use of two data structures : signals (start, stop,…) and segments (exons, conserved regions,…)
- Based on a graph which describes the gene structure : genomic region limited between two 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



3 level prediction programs
(SNAP, GeneID, GAZE, …)
2 level programs (with target)
(BLAST, est2genome, Genewise, …)
1 level programs (with or without target)
(TRF, RepeatMasker, RepeatScout, …..)

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.



Search by position or name

Q82640

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed

Search :	Displays :
scaffold_1:1..10000	region of scaffold_1 from 157,658 to 177,657
GSTUM00009626001	region of Genoscope annotation with identifier GSTUM00009626001
Q2U6C2	matches with O82640 *Aspergillus oryzae* protein from Uniprot

Tuber melanosprum

Photo of Tuber melanosprum (Author and credit: Sten. Photo from the website http://www.wikipedia.fr)

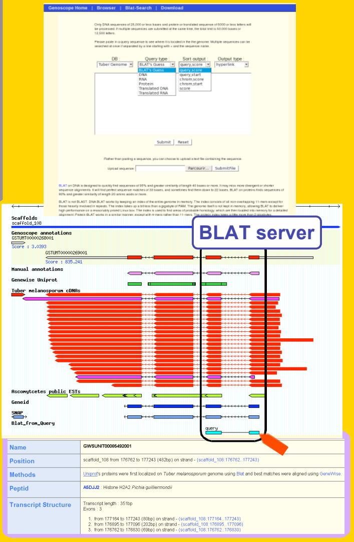
Assembly summary (download)

	Number	N50 length (kb)	Size (kb)
Reads	1,262,177		
Contigs (≥ 2kb)	4,484	62kb	
Scaffolds	398	637	124,946

Annotation summary (download)

Annotated genes
Annotated transcripts
Average number of coding exons per gene
Average number of UTR exons per gene
Average gene size (bp)
Average CDS size (bp)
Average exon size (bp)
Average intron size (bp)

AT genoscope.cns.fr if you have any question



Only DNA sequences of 25,000 or less bases and protein or translated sequence of 5000 or less letters will be processed. If multiple sequences are submitted at the same time, the total limit is 50,000 bases or 12,500 letters.

Please paste in a query sequence to see where it is located in the the genome. Multiple sequences can be searched at once if separated by a line starting with > and the sequence name.

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.

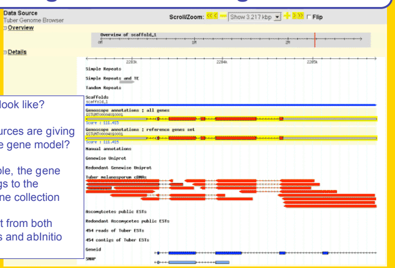
Upload sequence:

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more.

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers except for those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to deliver high performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed alignment. Protein BLAT works in a similar manner, except with 4-mers rather than 11-mers. The protein index takes a little more than 2 gigabytes.

Name	GWSUNIT00085492001
Position	scaffold_108 from 176762 to 177243 (482bp) on strand - (scaffold_108:176762..177243)
Methods	Uniprot's proteins were first localized on *Tuber melanosporum* genome using Blat and best matches were aligned using GeneWise.
Peptid	A5DJJ2 : Histone H2A2 *Pichia guilliermondii*
Transcript Structure	Transcript length : 351bp Exons : 3 1. from 177164 to 177243 (80bp) on strand - (scaffold_108:177164..177243) 2. from 176895 to 177096 (202bp) on strand - (scaffold_108:176895..177096) 3. from 176762 to 176830 (69bp) on strand - (scaffold_108:176762..176830)

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



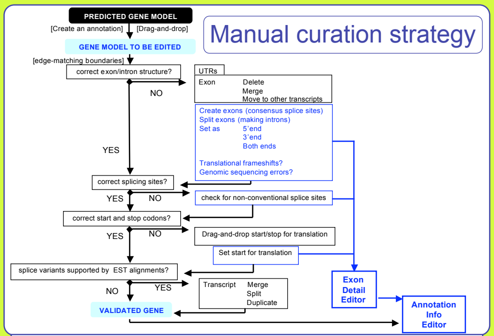
How does it look like?

Which ressources are giving support to the gene model?

In our example, the gene model belongs to the reference gene collection

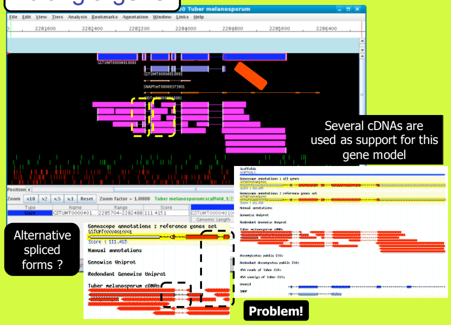
It has support from both *Tuber* cDNAs and abInitio predictions

Manual curation strategy


PREDICTED GENE MODEL
[Create an annotation]
[Drag-and-drop]
GENE MODEL TO BE EDITED
[edge-matching boundaries]
correct exon/intron structure?
NO
YES
UTRs
Exon
Delete
Merge
Move to other transcripts
Create exons (consensus splice sites)
Split exons (making introns)
Set as
5'end
3'end
Both ends
Translational frameshifts?
Genomic sequencing errors?
correct splicing sites?
YES
NO
check for non-conventional splice sites
correct start and stop codons?
YES
NO
Drag-and-drop start/stop for translation
Set start for translation
splice variants supported by EST alignments?
NO
YES
Transcript
Merge
Split
Duplicate
VALIDATED GENE
Exon Detail Editor
Annotation Info Editor

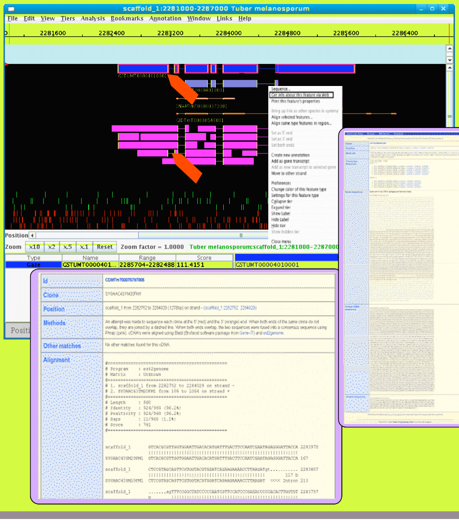
Most of the features that characterize a genome can be identified by the automated procedure. However, gene models and annotation can still be improved by human manual annotation to find possible errors or to resolve incongruous evidence on the automatic annotation of the genome.

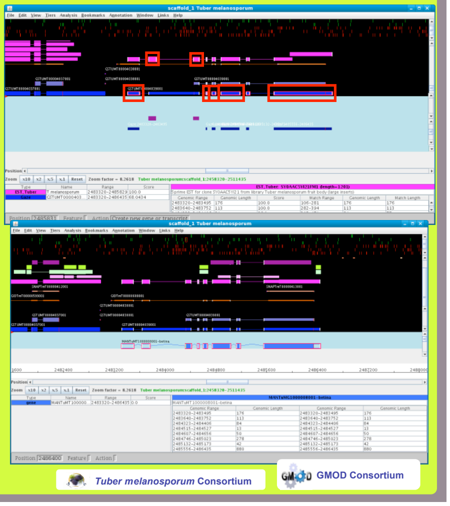
Editing a gene



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.





Type	Name	Range	Score
gene	MANTuMT100000...	2483320-2486435	0.0

MANTuMG1000008001-betina
MANTuMT1000008001-betina

Genomic Range	Genomic Length	Genomic Range	Genomic Length
2483320-2483495	176	2483320-2483495	176
2483640-2483752	113	2483640-2483752	113
2484323-2484406	84	2484323-2484406	84
2484515-2484527	13	2484515-2484527	13
2484607-2484656	50	2484607-2484656	50
2484746-2485023	278	2484746-2485023	278
2485132-2485173	42	2485132-2485173	42
2485556-2486435	880	2485556-2486435	880

Position 2486400

Tuber melanosporum Consortium

GMOD Consortium