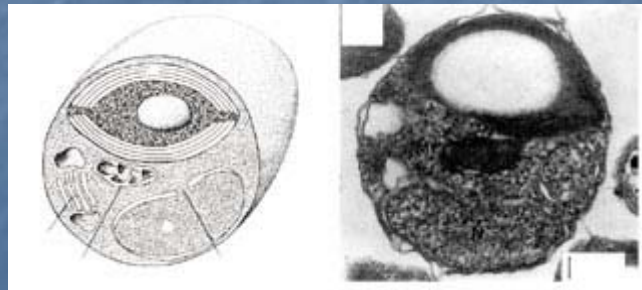


# Annotation of (new) genomes

Stephane Rombauts

Case: *Ostreococcus tauri*



# The genomes

- *Arabidopsis thaliana*
- *Ostreococcus tauri* (green algae)
- *Populus balsamifera* ssp. *trichocarpa*
- *Medicago truncatula*
- *Laccaria bicolor* (fungi)... starting soon

# Arabidopsis

**TIGR : 29 993** total genes / **26 207** predicted coding genes  
(without pseudogenes)

**EUGENE : 27 979** predicted protein encoding genes

- 49% of EuGene predictions have identical CDS compared to TiGR (13727 genes)
- 22880 genes located on same locus (but are different for at least 1 position)
- 2051 genes from TiGR not predicted by EuGene
- 1392 genes from EuGene not predicted by TiGR
- Pilote-project at TiGR to sequence those genes TiGR 'missed'

# Ostreococcus...

## what did we know

- Little is known...
  - Smallest eukaryote, green algae
- not much data available
  - No genetic map
  - < 3000 ESTs (contaminated...)
- We thought it would be easy...
  - expected to be like yeast due to the genome compactness, with strong signals

# Ostreococcus...

## what have we learned already

- 3 different kinds of genes
  - single exon genes → great majority
  - “small”-intron genes
    - very small with a typical size range of 36-70bp
    - Splicing signals are not consensual, as in higher plants
    - no clear branch-point motif
    - resemble higher plant in being more AT-rich than the nearby exons
    - Restricted to 1 chromosome
  - “large”-intron genes
    - typical size range of 80-500bp
    - very canonical splicing signals (donor – branch-point – acceptor)
    - contrary to Ath introns, they are not AT-rich
    - Few introns

# Ostreococcus...

## what have we learned already

Artemis Entry Edit: Chr1\_emb1

File Entries Select View Goto Edit Create Write Graph Display

Selected feature: bases 51 intron ()

Entry:  Chr1\_emb1

intron intron intron intron

1096800 1096900 1097000 1097100 1097200 1097300 1097400 1097500 1097600

intron intron

I Y Q N M L V P I E V I R \* N \* A P F Q P W R T  
S I K T C + Y R S K L L D E I E H L F N L G E L  
L S K H V S T D R S Y + M K L S T F S T L E N S  
L S S P TCTATCAAAACATGTTAGTACCGATCGAAGTTATTAGATGAAATTGAGCACCTTTTCAACCTTGGAGAACTCR V  
Y Q A Q  
S I K P 1097280 1097300 1097320 1097340 V #  
STATCAAGCCCAAGATAGTTTTGTACAATCATGGCTAGCTTCAATAATCTACTTTAACTCGTGGAAAAGTTGGAACCTCTTGAGPTGTGA  
:40 GATAGTTCGGGT + \* F M N T G I S T I L H F Q A G K \* G Q L V G ACAT  
R D L G D I L V H # Y R D F N N S S I S C R K L R P S S T  
+ \* A W  
I L G L R D F C T L V S R L # # I F N L V K E V K S F E H L

# Poplar

- International Poplar genome consortium (IPGC)
- Sequencing done by JGI  
(whole genome shotgun sequencing)
- Gene structure annotation:
  - In parallel whole genome annotation aiming at the best possible annotation from the very first public release on
    - ORNL → Grail-exp
    - JGI → FGenesH + Genewise
    - Ghent → EuGene
  - JGI's Rule-system to elect for each locus the 'best' gene-model

# Poplar

(first preliminary results)

- 58036 gene-models selected
  - 39.8% EuGene,
  - 39.0% FGenesH,
  - 14.3% Grail-exp,
  - 6.9% Genewise
- Press release 21-09-2004
  - Results will be accessible through the JGI genome browser

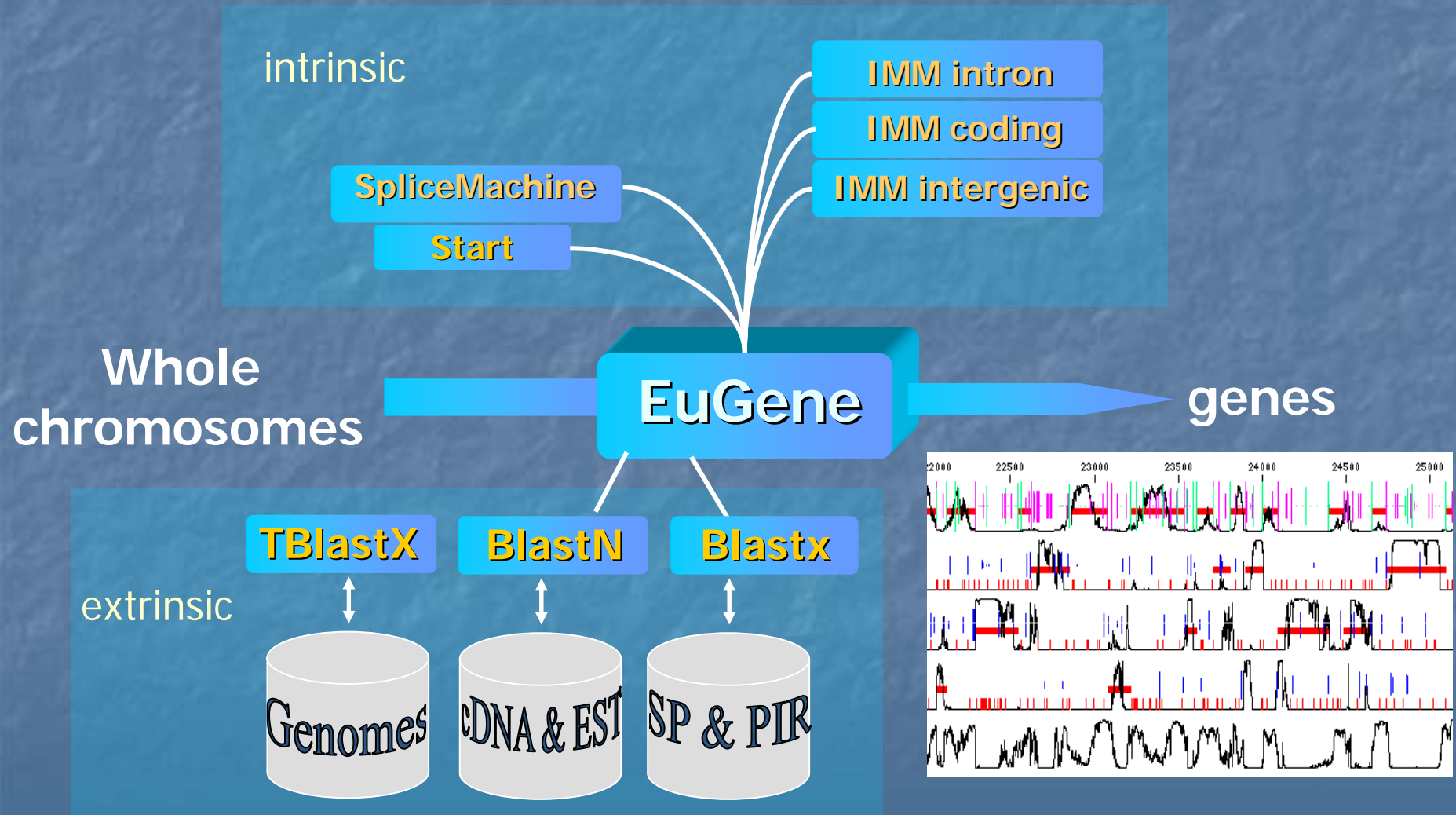
# Medicago truncatula

- US-project and European project (FP6)
  - 67 members for the European side only
  - Annotation (in Europe) shared between INRA-Toulouse (FR), Ghent (B) and MiPS (GER)
  - BAC-sequencing
    - Generating training sets and building models is less performing due to a gradual accessibility of data
    - Models ready, >350 BAC annotated

Structural

**Annotation**

# Gene prediction platform: **EuGene**




# Needs for structural annotation

To build models for the intrinsic part of EuGene

- Curated gene-models, for:
  - intron borders,
  - ATG-modelling
  - Coding-IMM,
  - Intron-IMM
- Genes in their genomic context
  - Non-coding IMM (intergenic)
- Need for as many as possible transcripts mapped on genomic sequence
  - Transcripts = full length or ESTs, no assembled ESTs

# Structural Annotation: the models

- 
- ATG-prediction
    - As reliable as possible starts of genes
  - IMM
    - as many as possible complete genes!
    - + introns, ++ intergenics
  - Training of EuGene
    - = tuning the  $\alpha$  &  $\beta$  ( $\alpha \cdot \text{score}^\beta$ )

as many as possible **complete genes**  
in **genomic configuration**

# Proposal for Physcomitrella

- We host someone at our institution to learn and use all our tools (2 months min).
  - Build training set
  - Train software
  - Run structural gene prediction
- Share similarly as with poplar all data and results with in consortium
- Regular contact through conference calls.

# Additional analyses

- i-ADHoRe for genome structure and organisation
- MIRfinder for micro-RNAs
- Other ncRNAs
- Transposable elements
  
- Also possible:
  - Functional annotation
  - Promoter analysis

# Data display

Show data through Gbrowse  
(= Genome annotation browser)  
and share it through LDAS or bioMoby

Restricted  
access for  
consortium  
during  
project,

World wide  
if OK with  
consortium  
after  
project  
completion

Showing 100 kbp from LG\_I, positions 100,001 to 200,000

**Instructions:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

**Examples:** [LG\\_I](#)

[\[Hide banner\]](#) [\[Hide instructions\]](#) [\[Bookmark this view\]](#) [\[Link to an image of this view\]](#) [\[Publication quality image\]](#) [\[Help\]](#)

**SEARCH**

Landmark or Region     Flip

Scroll/Zoom:     Show 100 kbp

**VIEW**

Overview of LG\_I

**Exon**

S\_00010010.exon S\_00010014.exon S\_00010017.exon S\_00010019.exon S\_00010020.exon S\_00010022.exon S\_00010025.exon  
S\_00010011.exon S\_00010015.exon S\_00010018.exon S\_00010021.exon S\_00010023.exon S\_00010026.exon  
S\_00010012.exon S\_00010016.exon S\_00010013.exon S\_00010024.exon

**CDS**

S\_00010010.CDS S\_00010014.CDS S\_00010016.CDS S\_00010017.CDS S\_00010019.CDS S\_00010020.CDS S\_00010022.CDS S\_00010025.CDS  
S\_00010011.CDS S\_00010015.CDS S\_00010018.CDS S\_00010021.CDS S\_00010023.CDS S\_00010026.CD  
S\_00010012.CDS S\_00010013.CDS S\_00010024.CDS

**DATA SOURCE SELECTION**

Data Source

# Acknowledgments

Sven Degroeve (gene prediction software)

Steven Robbens (ostreococcus)

Lieven Sterck (poplar)

Pierre Rouze

Yves Van de Peer

Thomas Schiex' group  
at INRA-Toulouse

Contact: [strom@psb.ugent.be](mailto:strom@psb.ugent.be)

