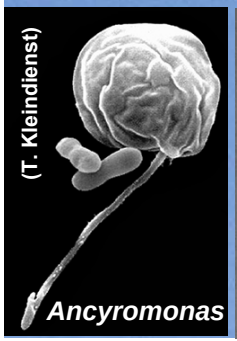
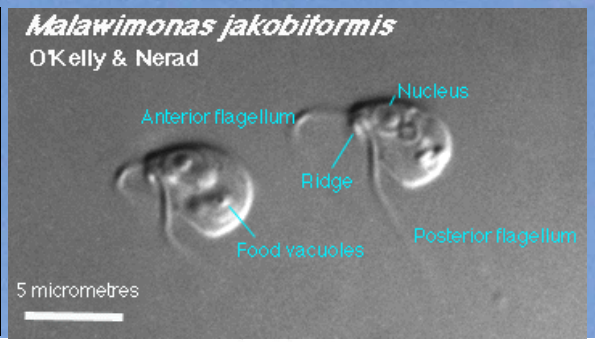
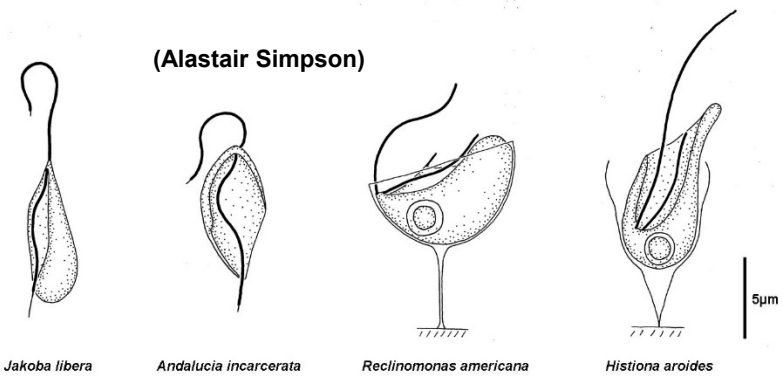
# Spliceosomal introns in 'primitive' unicellular flagellates

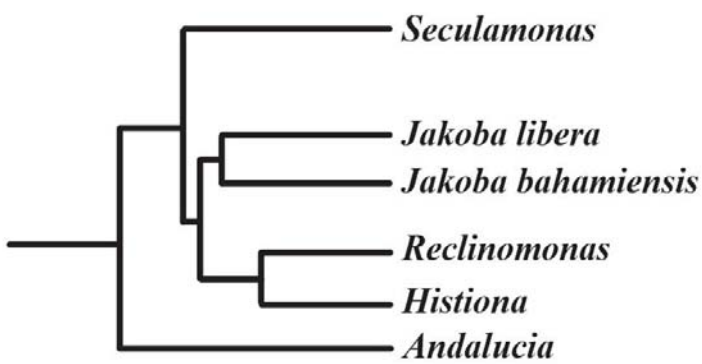
**M. Sarrasin, G. Burger and BF. Lang**

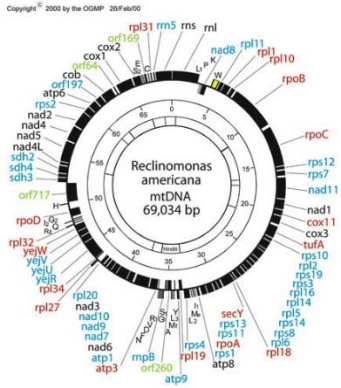
*Department of Biochemistry, Robert-Cedergren Centre of Bioinformatics and Genomics, Université de Montréal, Montréal, QC, H3C 3J7, Canada*



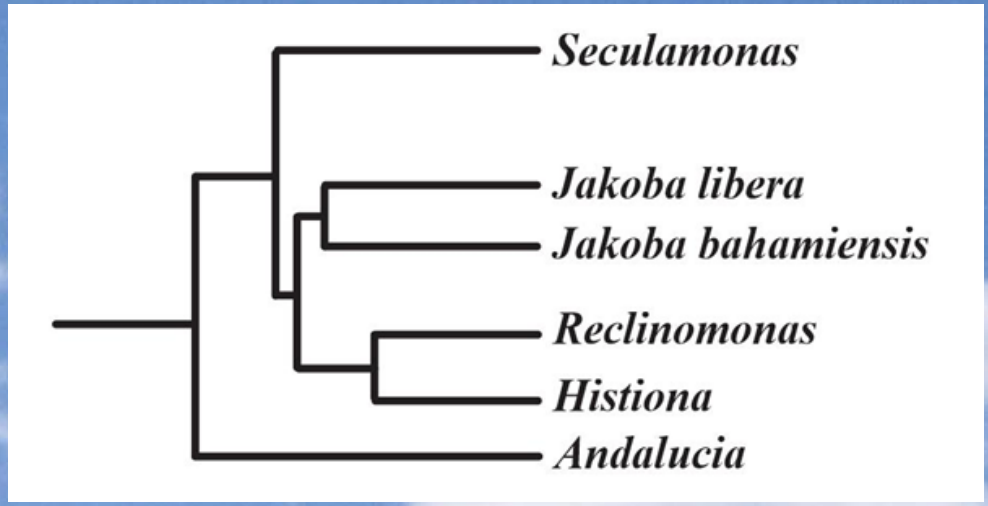
# Topics of my presentation

- **Why analyze nuclear genomes/transcriptomes of the chosen unicellular flagellates**
- **How to select eukaryotic contigs within a sea of bacterial contaminants?**
- **Improved nuclear genome assembly**
- **Improved gene/intron modeling procedures**
- **Results :**
  - **Genome size, number of genes and functional classes, introns**
  - **Do primitive eukaryotes have common structured RNAs**
  - **… including regular spliceosomal RNAs of the two types?**
  - **More than one type of major and minor spliceosome?**

# Species selection:

## Jakobids:

***Andalucia godoyi***
***Jakoba bahamiensis***
***Jakoba libera***
***Reclinomonas americana***
***Seculamonas***




## Malawimonads:

***Malawimonas californiana***
***Malawimonas jakobiformis***
***Malawimonas sp.***

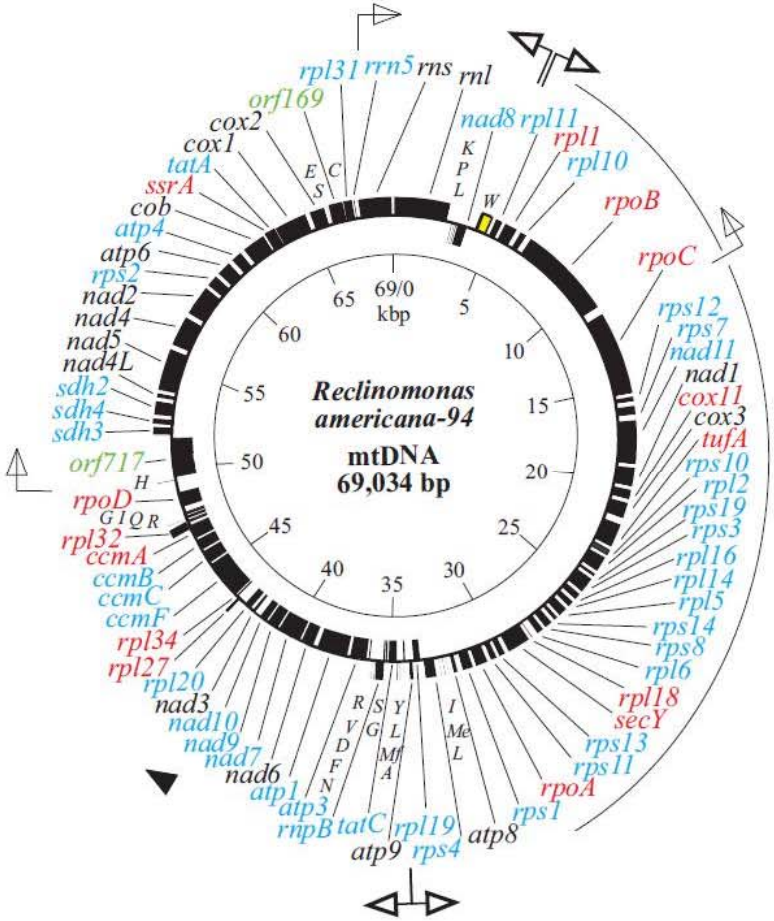
## Close to malawimonads:

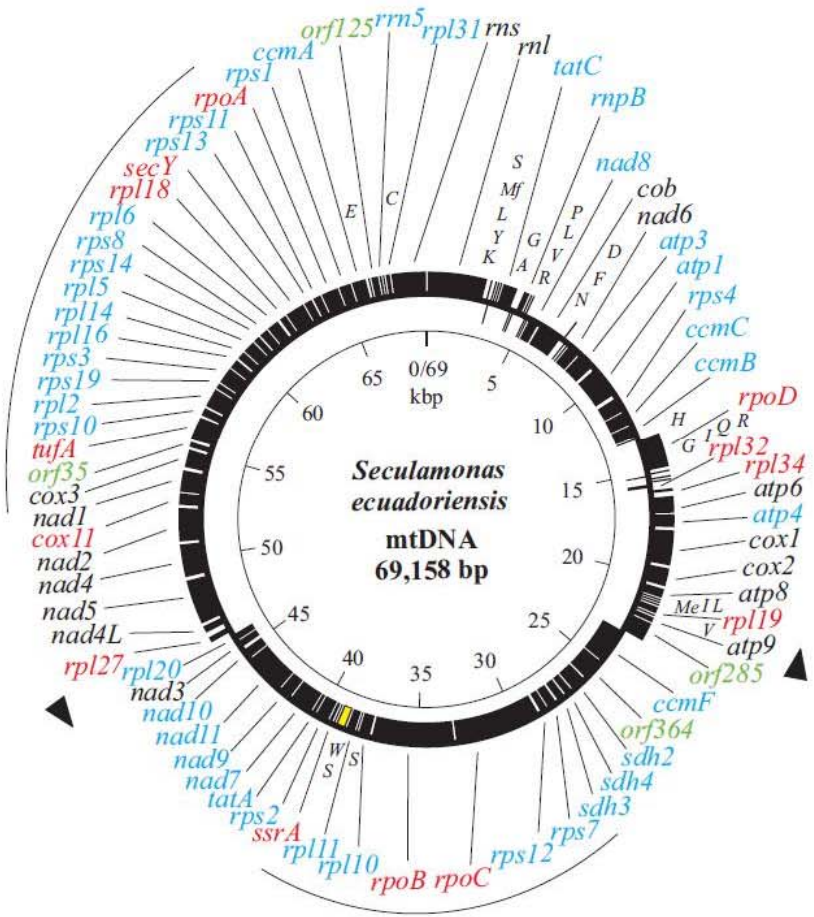
***Planomonas micra***

**All require live bacteria as food source – contamination issues**

# Why analyze genomes (plus transcriptomes) of the chosen unicellular flagellates?

## Jakobids have most 'primitive', gene-rich mitochondrial genomes

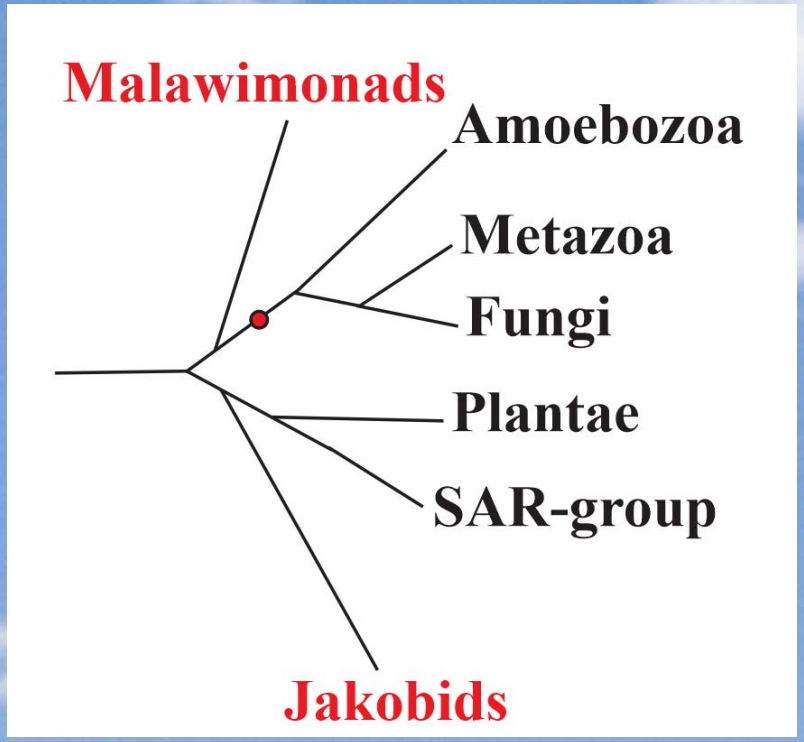


# Why analyze nuclear genomes plus transcriptomes of the chosen unicellular flagellates?

**Malawimonads have very short branch length in phylogenies, branch deeply in eukaryotic tree ancestral to animal/fungi/amoebozoans; and far away from jakobids.**

**Recent phylogenomics suggests that *Planomonas* maps deeply in the tree, not far from the malawimonad divergence.**



# Selected questions

**Do jakobids and malawimonads have**

- **the basic set of nuclear protein coding genes and**
- **structured RNAs (RNase P, MRP, SRP etc)**
- **spliceosomal introns (major and U12-type)**
- **recognizable U RNAs and associated proteins**
- **typical splice junctions**

**… etc …**

# Challenges in genome analysis: incorrect genome assemblies and gene models

**(1) Main issue: hybrid sequence reads (from library ligation reaction) cause incorrect joining : mixed eukaryote – bacterial contigs**

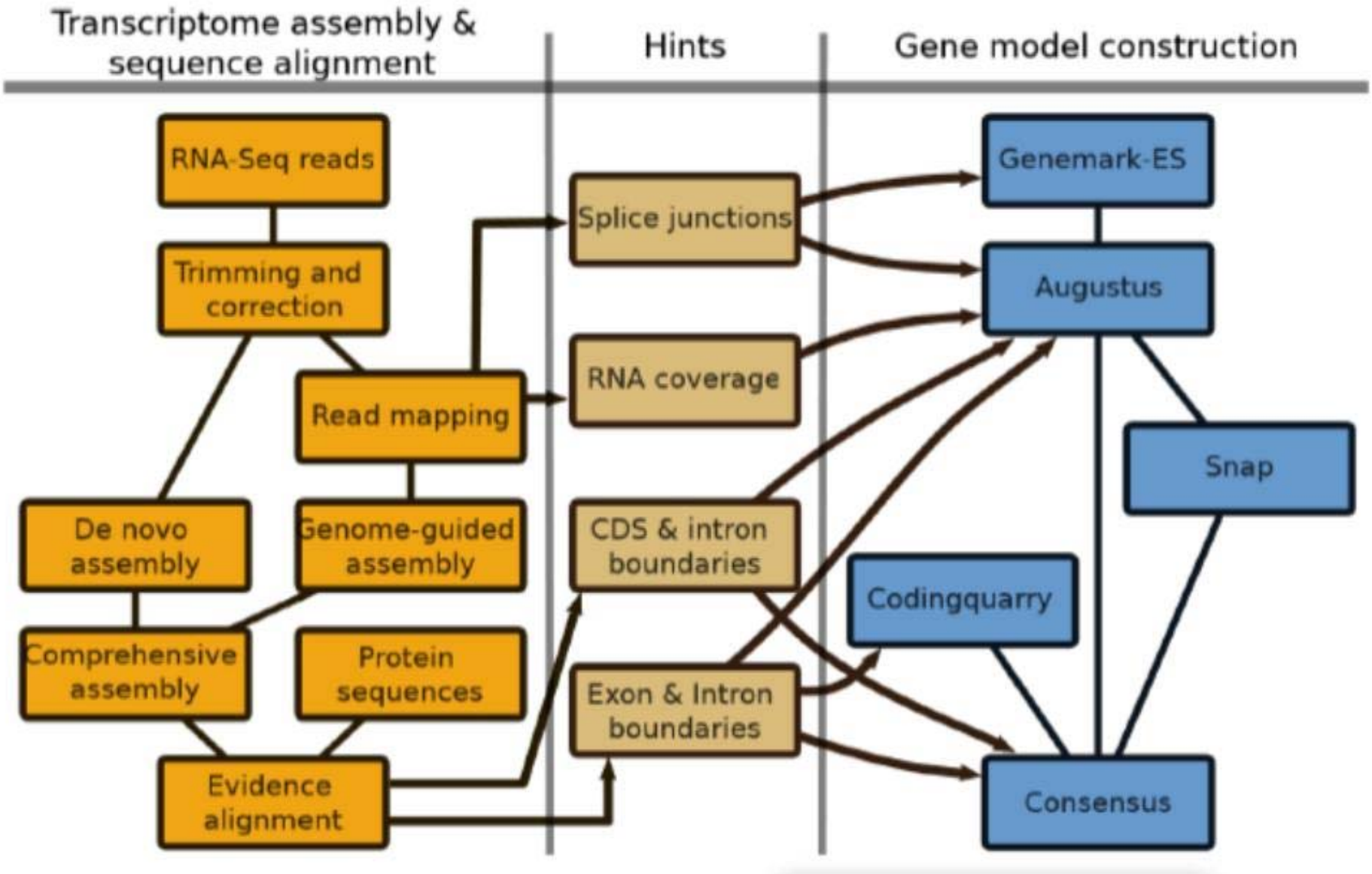
**Solutions: identify and remove hybrid reads, and selectively assemble eukaryotic reads only (new iterative assembly procedure)**

# (2) How to filter out nuclear eukaryotic contigs?

- mapping of deep RNAseq data (poly-A mRNAs)
- gene models indicate spliceosomal introns (if present)
- AT content
- read coverage
- similarity of eukaryotic *versus* bacteria-specific genes (issue with organelles …)
- comparative genomics – sequence several protist species from the same clade

Solution: we assembled a pipeline that decides, based on combined evidence. Yet, a human expert is still needed for the final check.

# (3) Improving gene/intron modeling procedure



# Current, comparative view of nuclear genomes

| Species | Size (Mbp) | Contigs | Genes (protein) | Introns GT-AG | Introns AT-AC | GT-AG U-RNAs | AT-AC U RNAs |
|---|---|---|---|---|---|---|---|
| *Ancyromonas* | 29,6 | 7 181 | 13 433 | 7 540 | - | + | - |
| *Malawi_calif* | 50,1 | 953 | 13 559 | 46 428 | + | + | + |
| *Malawi_jakobi* | 70,8 | 8 106 | 25 693 | 143 089 | + | + | + |
| *Malawi_sp* | 40,9 | 2 678 | 18 991 | 54 155 | - | + | - |
| *Andalucia* | 20,1 | 66 | 8 642 | 1 280 | - | + | - |
| *Jakoba_baha* | 28,6 | 6 085 | 11 870 | 69 350 | 74 | + | + |
| *Jakoba_libera* | 80,7 | 33 265 | 27 121 | 59 089 | + | + | + |
| *Reclinomonas* | 51,3 | 15 554 | 21 039 | 111 752 | + | + | + |
| *Seculamonas* | 47,9 | 3 739 | 10 256 | 97 199 | + | + | + |

**Gene numbers in *M. jakobiformis*, *J. libera* and *Reclinomonas* are inflated due either to genome duplication or ploidy. In *Reclinomonas* , the distribution of variants is consistent with a diploid genome, *J. libera* seems like a more complex mixed situation.**

# Current, comparative view

## Mito proteome (in nuclear genes) mostly standard

- a few functions more bacteria-like (analyses by Mike Gray)
- Phage-like mitochondrial RNA polymerase in *Ancyromonas* and malawimonads, jakobids have bacterial subunits encoded in mtDNA
- most of mitochondrial import machinery (TIM complex)

## Other major functions also fairly conventional

Including: proteasome, peroxisome, golgi, nuclear pore, some meiosis and sex-related genes, dyneins, other cytoskeleton structures, RNase P and signal recognition complex …

Presence of RNase P, MRP, SRP RNAs, yet several are only found after improving/updating RFAM CM models

*Andalucia* has streamlined gene sets; seems secondarily derived.

# Presence of spliceosomal U RNAs

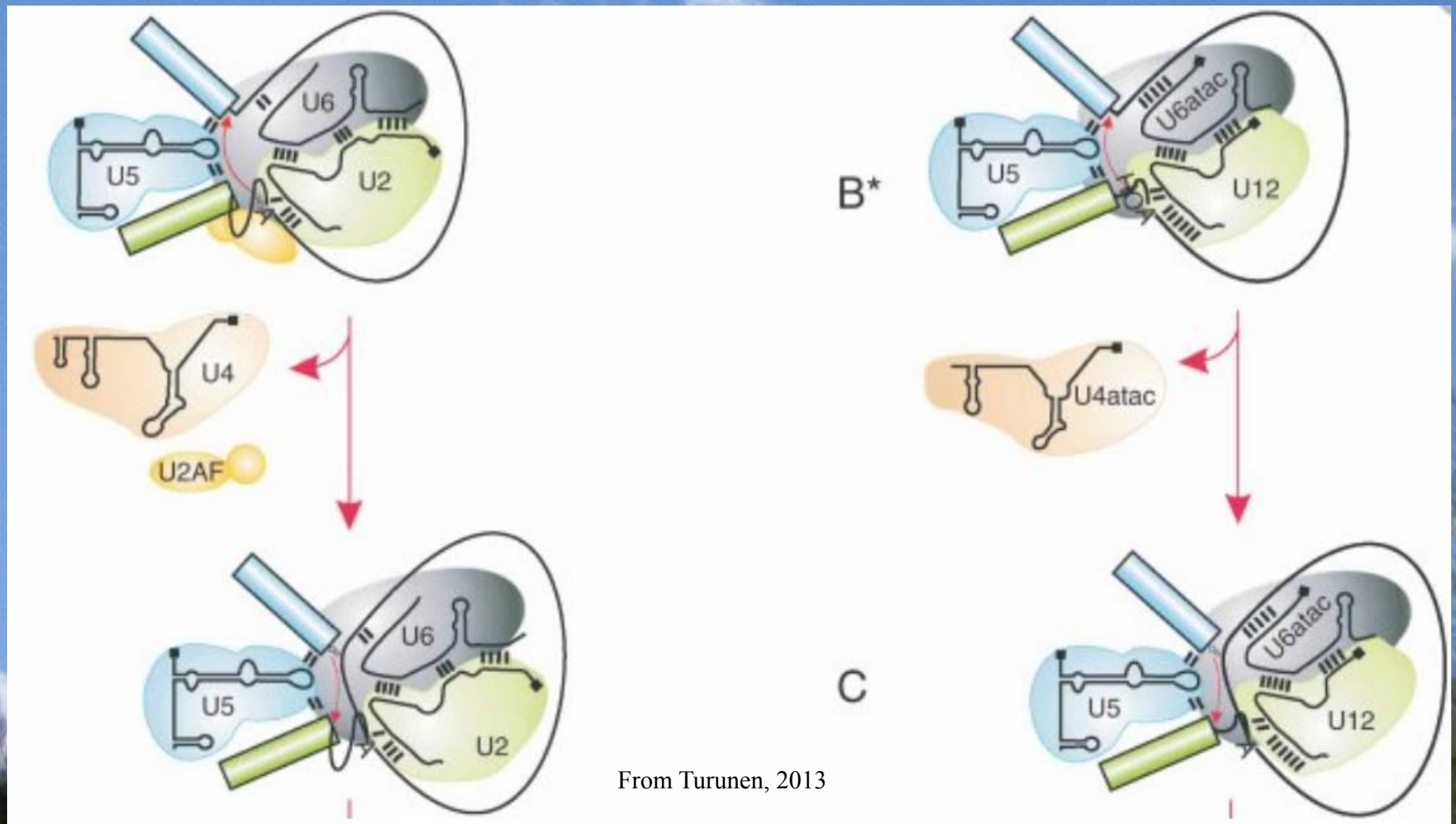
**U1,2,4,5.6: major spliceosome, GT – AG boundaries**
**U4atac, U5, U6atac,U11,12: minor spliceosome, AT – AC and GT - AG**

| Species | U1 | U2 | U4 | U5 | U6 | U4atac | U6atac | U11 | U12 |
|---|---|---|---|---|---|---|---|---|---|
| *Planomonas* | + | + | + | - | - | - | - | - | - |
| *Malawi_calif* | + | + | + | + | + | - | + | + | + |
| *Malawi_jakobif* | + | + | + | + | + | - | + | + | + |
| *Malawi_sp* | + | + | + | + | + | - | - | - | - |
| *Andalucia* | + | + | + | + | + | - | - | - | - |
| *Jakoba_baha* | + | + | + | - | + | - | + | + | - |
| *Jakoba_libera* | + | + | + | - | + | - | + | + | + |
| *Reclinomonas* | + | + | + | - | + | - | + | + | + |
| *Seculamonas* | + | + | + | + | + | - | + | + | + |

**Occurrence of U-RNAs correlates with observed presence of AT/AC introns, however lack of U4atac and U5 ???**

# U4/U4atac and U5 have essential roles in spliceosome assembly and structure.

## Are there other members of this RNA family?




From Turunen, 2013

# Families of distinct U RNAs?

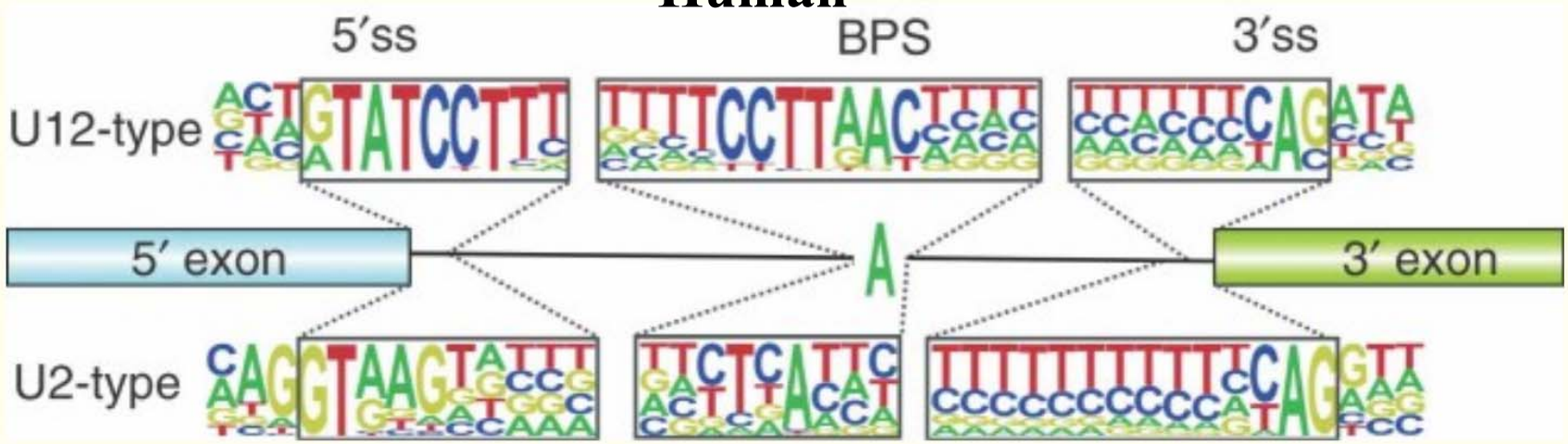
| Species | U1 | U2 | U4 | U5 | U6 | U4atac | U6atac | U11 | U12 | U11/12 proteins |
|---|---|---|---|---|---|---|---|---|---|---|
| Sacch.cerevisiae | (1) | 1 | 1 | 1 | 1 | - | - | - | - | - |
| Schizo.pombe | 1 | 1 | 1 | 1 | 1 | - | - | - | - | - |
| Aspergillus | 1 | 1 | 1 | 1 | 1 | - | - | - | - | - |
| Homo -genome | 186 | 530 | 106 | 30 | 1527 | 17 | 45 | 3 | 1 | + |
| Homo -transcriptome | 51 | 63 | 32 | 12 | 254 | 5 | 6 | 3 | 1 | + |
| Acanthamoeba | 9 | 18 | 2 | 2 | 1 | - | 1 | 1 | 1 | + |
| | | | | | | | | | | |
| Andalucia | 1 | 1 | 1 | 1 | 1 | - | - | - | - | - |
| Reclinomonas | 2 | 1 | 1 | - | 2 | - | 2 | 2 | 1 | + |
| Jakoba bahiemensis | 1 | 1 | 1 | - | 1 | - | 1 | 1 | - | + |
| Jakoba libera | 2 | 1 | 1 | - | 1 | - | 1 | 1 | 1 | + |
| Seculamonas | 1 | 1 | 1 | 1 | 1 | - | 1 | 1 | 1 | + |
| | | | | | | | | | | |
| Malawi_jak | 3 | 1 | 2 | 1 | 1 | - | 1 | 2 | 1 | + |
| Malawi_cal | 2 | 1 | 1 | 2 | 1 | - | 1 | 1 | 1 | + |
| Malawi_sp | 2 | 3 | 2 | 2 | 2 | - | - | - | - | - |
| Planomonas | 1 | 1 | 1 | - | - | - | - | - | - | - |

**Function of these variants? More than two spliceosomes? Lack of U5 in some jakobids?**

**Needs modeling of potential U-RNA/intron splice site interactions**

# What about intron splice junctions ?

## Human




From Turunen, 2013

## *Jakoba bahamiensis* – extended motifs

| | | | | |
|---|---|---|---|---|
| **U12-type:** | **ATATCCTC** | **…** | **…** | **GTGTGCAC** |
| | **GTATC** | **…** | **…** | **YUCAG** |
| **U2-type** | **GTGCGT** | **…** | **…** | **YUCAG** |
| **group II** | **GTGCGA** | **…** | | |

# Conclusions

- More precise genome assembly and annotation procedures for highly contaminated total DNAs
- Jakobids and malawimonads have a basic set of about 9,000 - 13,000 nuclear protein coding genes, as well as common eukaryotic genes for structured RNAs. No surprising lack of general-function genes.
- Intron numbers vary wildly among jakobids, and splice-site motifs are unusually long
- Jakobid and malawimonad U RNAs are fairly typical, and frequently occur in more than one variant. U4atac remains unidentified, as well as U5 in three out of five jakobids.
- Modeling of U RNA – intron sequence interactions is required to better understand changes of the spliceosomal machineries.

## Thanks to collaborators of this project …

**Romain Derelle (Barcelona)**
**Alastair Simpson (Halifax, Canada)**
**Marek Elias' group (Czech Republic)**
**Andrew Roger's group (Halifax, Canada)**
**Mike W. Gray (Halifax, Canada)**
**David Morse (Montreal, Canada)**

## and financing from

NSERC CRSNG

Université de Montréal

Fonds de recherche Nature et technologies Québec

# U11 in human transcriptome (NCBI)

```
Query:       new  [CLEN=120]
Hit scores:
 rank     E-value  score  bias  sequence  start    end    mdl trunc   gc  description
 ----   --------- ------ -----  -------- ------ ------    --- ----- ----  -----------
  (1) !   3.1e-12   82.8   0.0  AK292656      2    130 +  cm    no 0.51  1
  (2) !   1.5e-10   75.1   0.0  FV525965      2     96 +  cm    3' 0.47  1
  (3) !   3.4e-09   68.9   0.0  J04118        1    131 +  cm    no 0.52  1
  (4) !   2.4e-07   60.5   0.0  LF385294   9773   9641 -  cm    no 0.46  1
  (5) !     3e-06   55.5   0.0  LF383785 125089 125221 +  cm    no 0.44  1
```