REPET: pipelines for the identification and annotation of transposable elements in genomic sequences

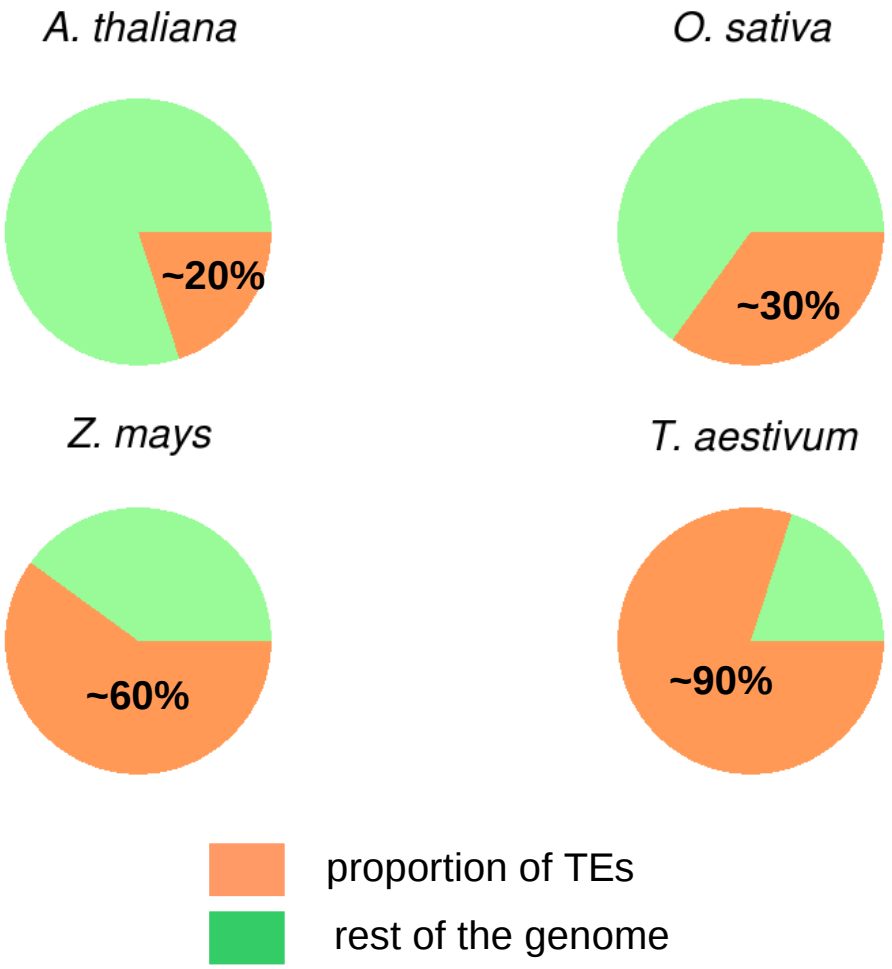
Timothée Flutre

INRA, Unité de Recherche en Génomique-Info (URGI)

03/09/2009

Context

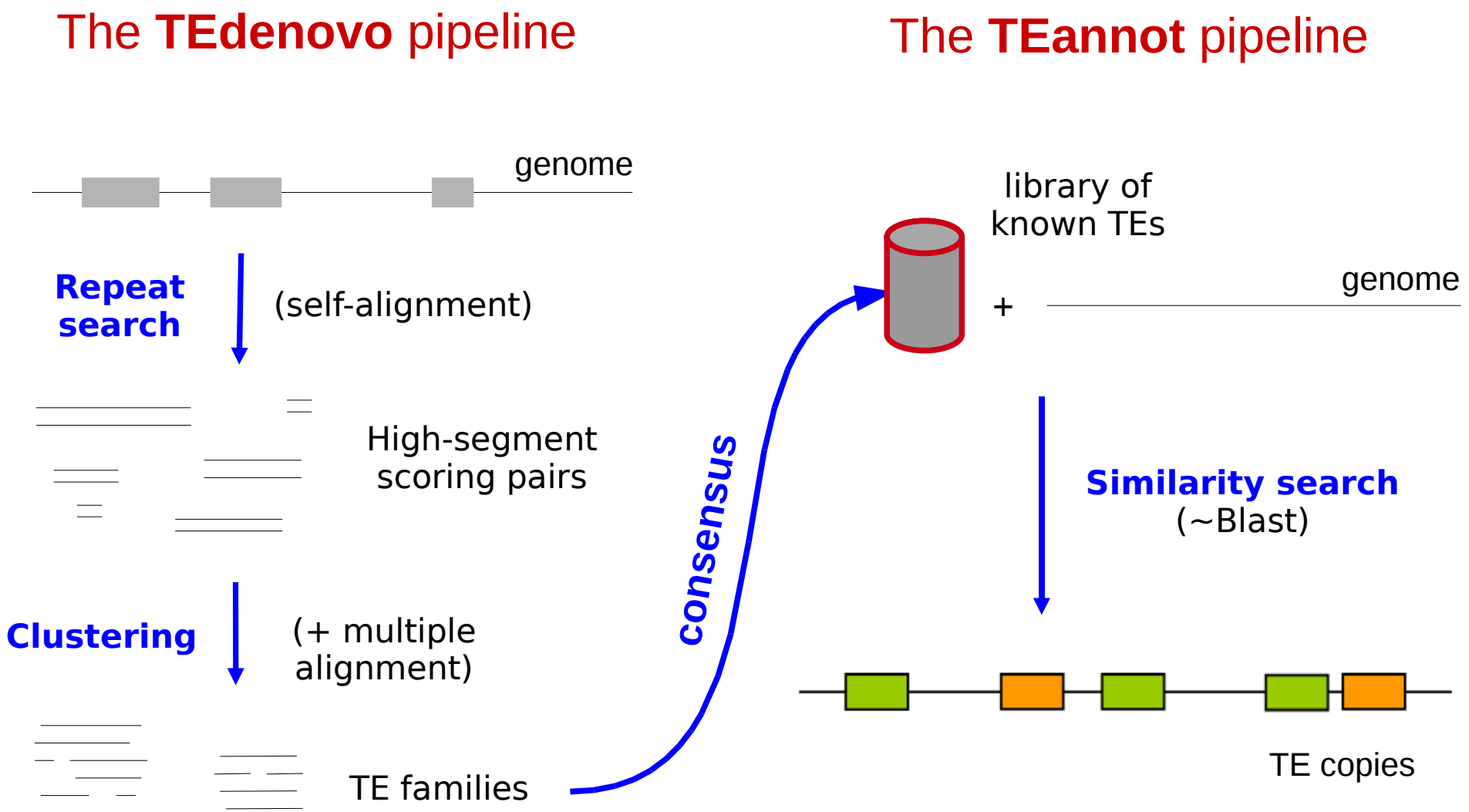
TEs in plant genomes



Genome sizes

- *A. thaliana*: ~180 Mb
- *O. sativa*: ~400 Mb
- *Z. mays*: ~2,400 Mb
- *T. aestivum*: ~16,000 Mb

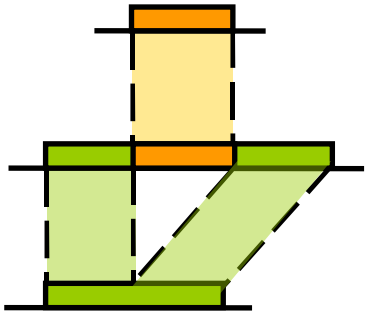
TE genomic annotation

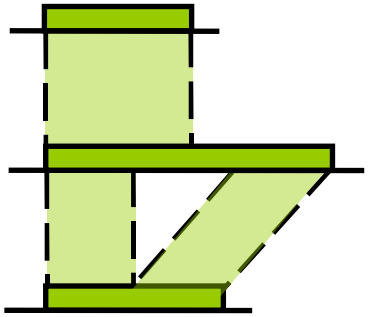


Biological challenges

… nested or degenerated elements

nested

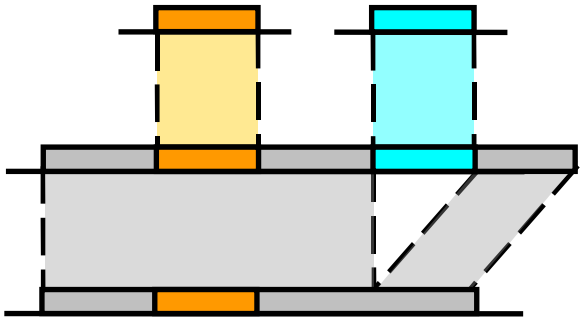




degenerated

… other genomic repeats

segmental duplications



satellites

tandem repeats

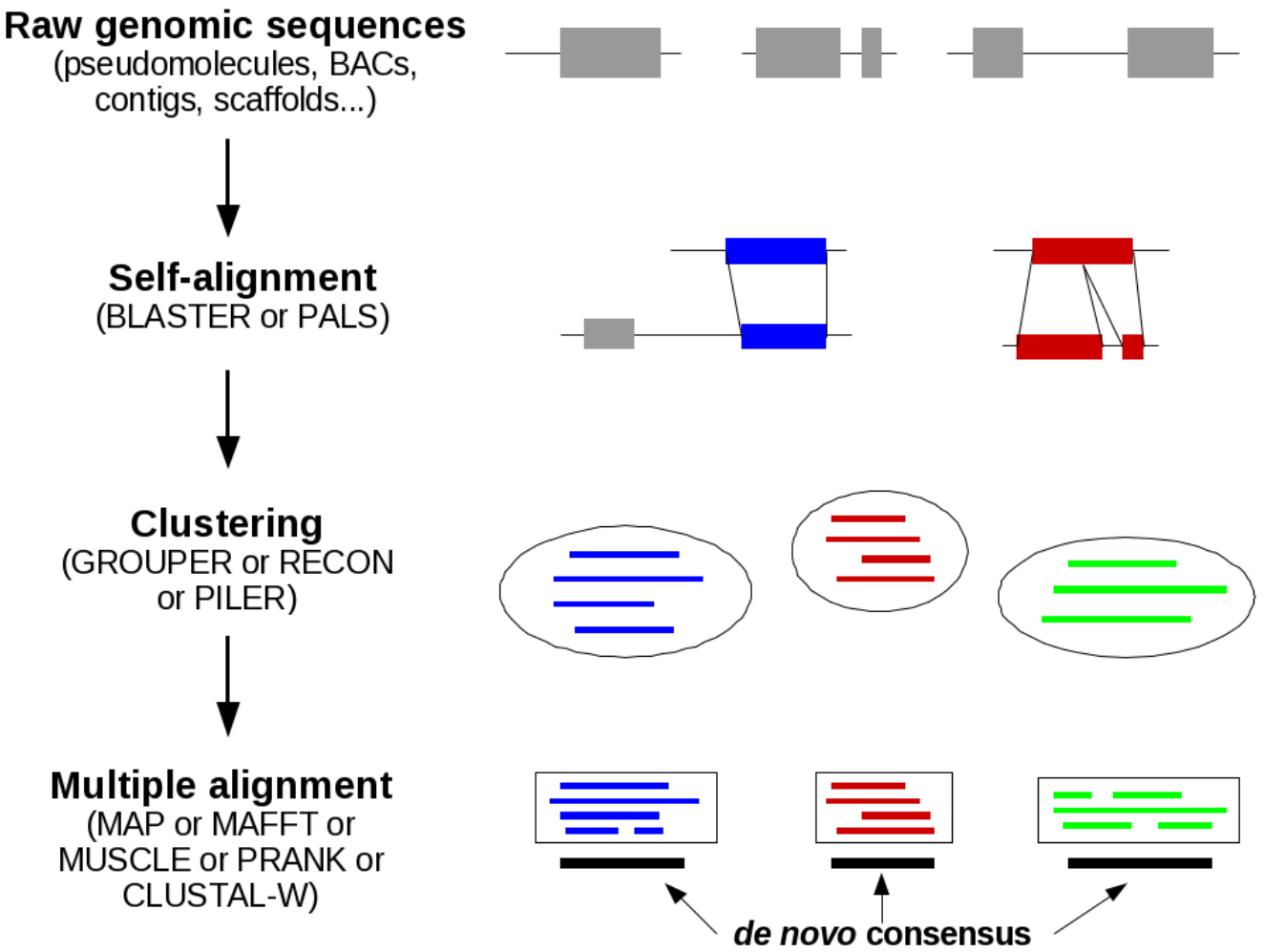
Technical challenges

- Handle big volumes of data:
  - Storage in MySQL tables

- Handle long computation times:
  - Interaction with computer clusters

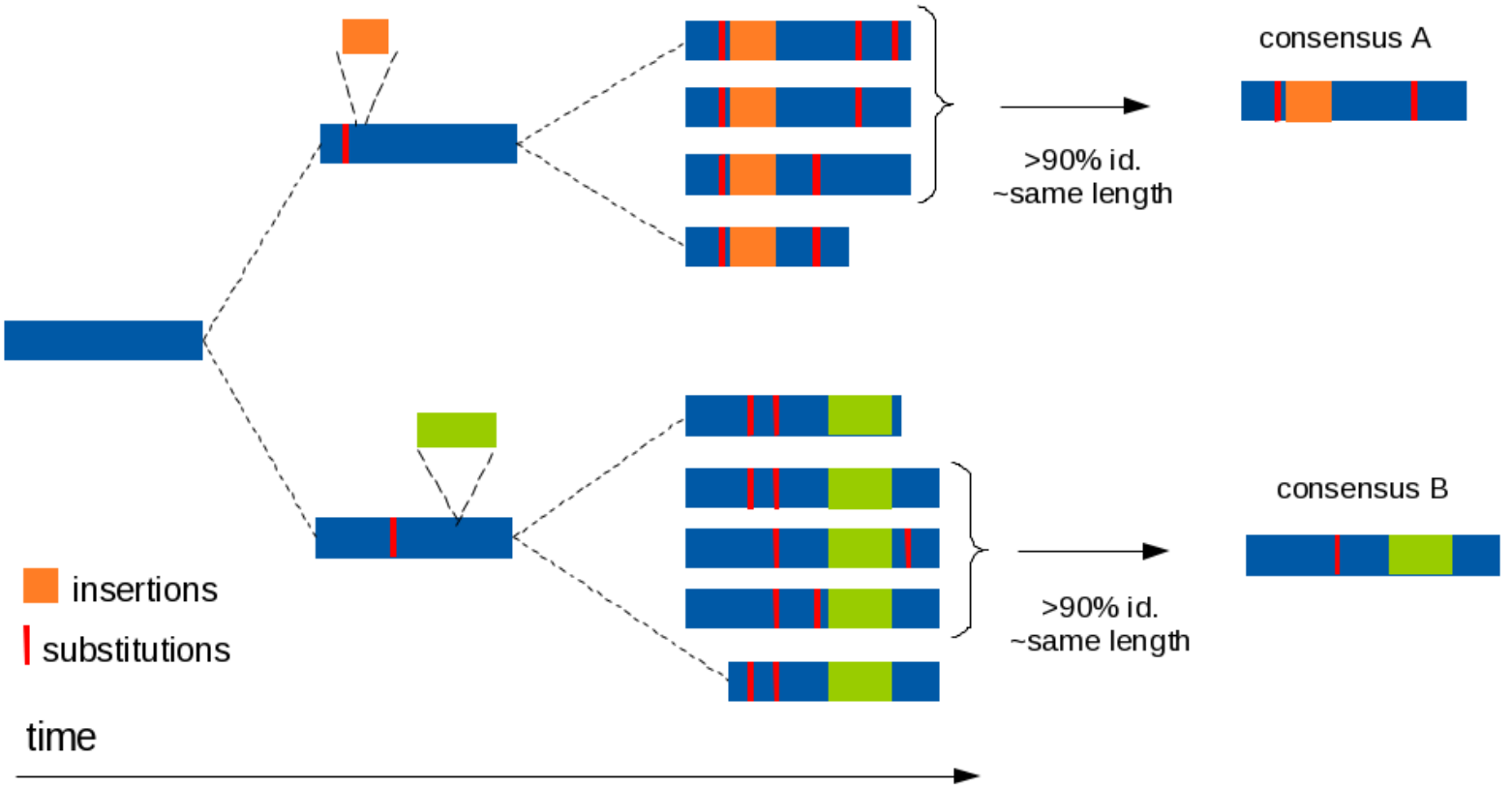
De novo identification of TE families

Build *de novo* TE consensus



Flutre *et al.*, *in prep.*

Clustering the variants



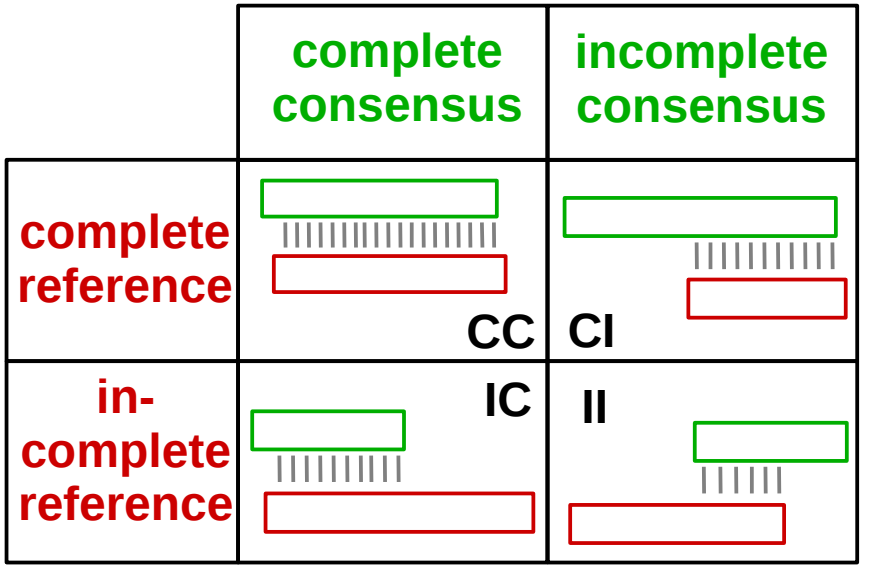
Benchmark

- *Drosophila melanogaster* genome sequence
  - Release 4: ~ 130 Mb (mainly euchromatic)

- Berkeley Drosophila Genome Project:

126 reference sequences of TEs found in the *D. melanogaster* genome

➞ high-quality library



Complete = match length >= 95%

Results on *D. melanogaster*

- **First steps of the TEdenovo pipeline:**
  - 220,000 HSPs from BLASTER (7.4% coverage)
  - 4004 clusters and consensus:
    - 3443 from GROUPER
    - 441 from RECON
    - 120 from PILER

Results on *D. melanogaster*

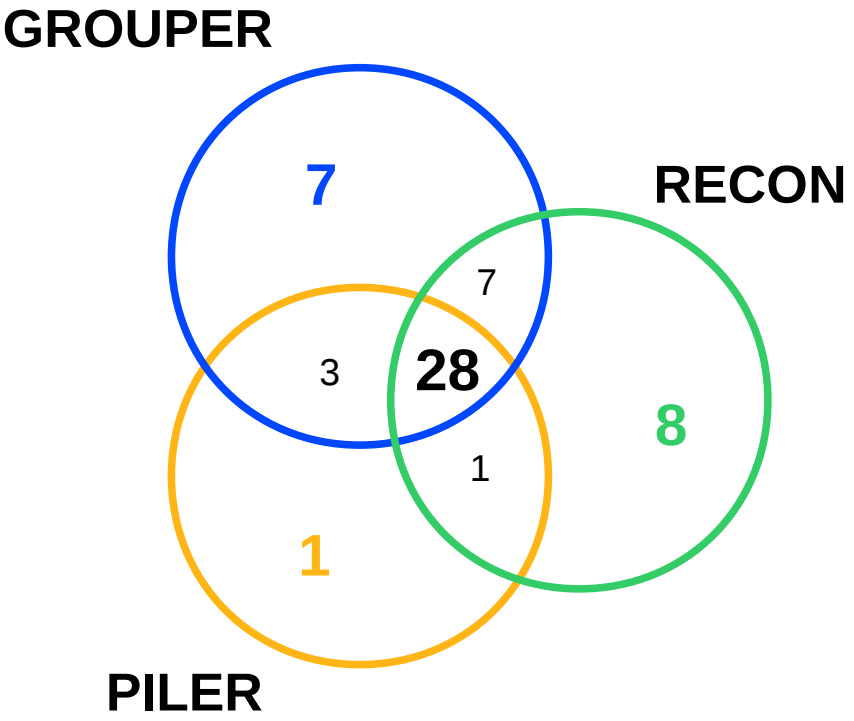
- **First steps of the TEdenovo pipeline:**
  - 220,000 HSPs from BLASTER (7.4% coverage)
  - 4004 clusters and consensus:
    - 3443 from GROUPER
    - 441 from RECON
    - 120 from PILER

⇒ **54** reference sequences retrieved with **precise boundaries** (“CC” alignments with *de novo* consensus)

⇒ **close to the optimum**: 62 families with 2 complete copies, 51 with 3 complete copies

Combined approach

Clustering methods	Ref. sequences in CC matches
Grouper	**45**
Recon	**44**
Piler	33
Grp + Rec	54
Grp + Pil	46
Rec + Pil	48
Grp + Rec + Pil	**54**



⇒ Combine several programs at the clustering step gives better results !

A. thaliana and O. sativa

- TAIR release 9 for *A. thaliana* and IRGSP build 5 for *O. sativa*

- **First steps of the TEdenovo pipeline:**
  - HSPs from BLASTER:
    - *A. thaliana*: 207,000 (coverage 13.5%)
    - *O. sativa*: 12,086,000 (coverage 3.1%)
  - clusters and consensus:
    - *A. thaliana*: **6,639** (5,344 from GROUPER, 994 from RECON and 301 from PILER)
    - *O. sativa*: **88,042** (81,400 from GROUPER, 5,071 from RECON and 1,571 from PILER)

TE classification

- False-positive filtering based on consensus analysis (TE structural features)
- Removal of redundant consensus

TE classification

- False-positive filtering based on consensus analysis (TE structural features)
- Removal of redundant consensus

- Search for terminal repeats (LTR and TIR)
- Search for polyA and SSR-like tails
- Detect matches with known TEs via tblastx and blastx
- Detect matches with host's genes via blastn
- Detect matches with TE HMM profiles

TE classification

- False-positive filtering based on consensus analysis (TE structural features)
- Removal of redundant consensus

- Search for terminal repeats (LTR and TIR)
- Search for polyA and SSR-like tails
- Detect matches with known TEs via tblastx and blastx
- Detect matches with host's genes via blastn
- Detect matches with TE HMM profiles

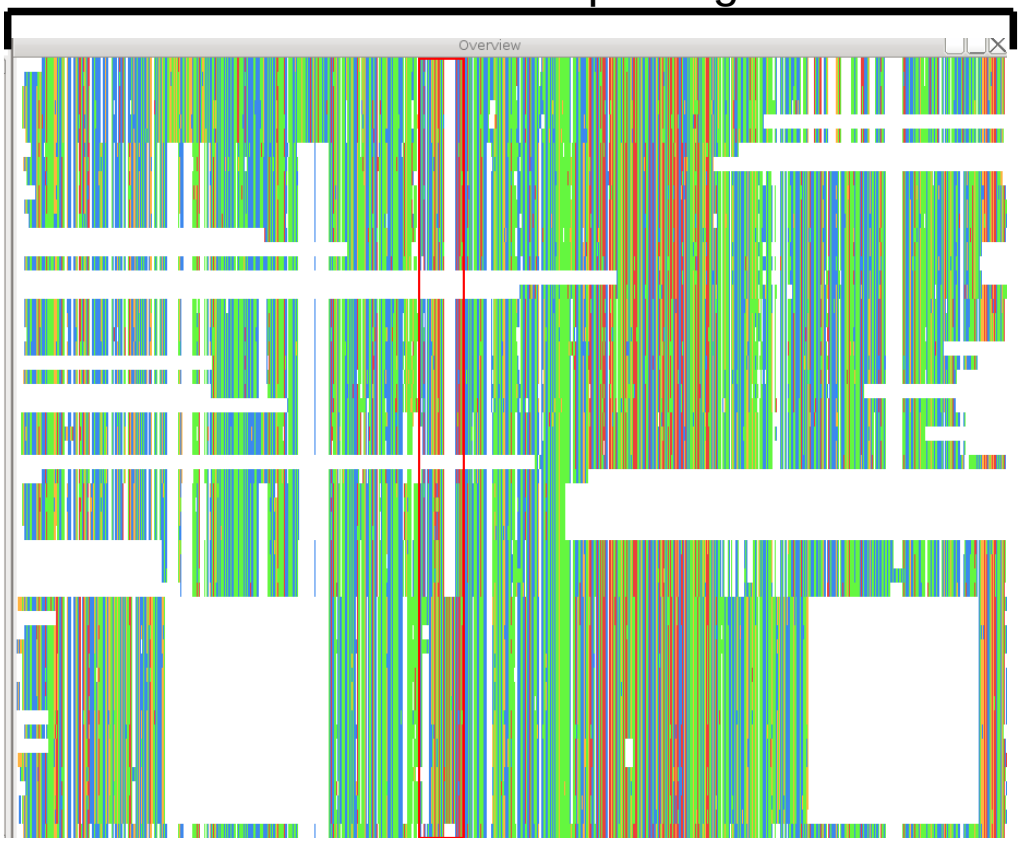
- **Category**: 'class I' or 'class II' for the TEs, 'SSR', 'HostGene', 'NoCat' or '?' for the others
- **Type**: 'LTR' / 'LARD' / 'LINE' / 'SINE' / 'TIR' / 'MITE' / 'Helitron' / 'Polinton' for the TEs, 'NA' or '?' for the other categories
- **Completeness**: 'yes', 'no', 'NA' or '?'
- **Comments**: briefly explain the classification
- **Confusedness**: 'yes' or 'no'

Flutre *et al.*, *in prep.*

TE families

- ATREP11, non-autonomous Helitron (1053 nt)

Overview of the multiple alignment

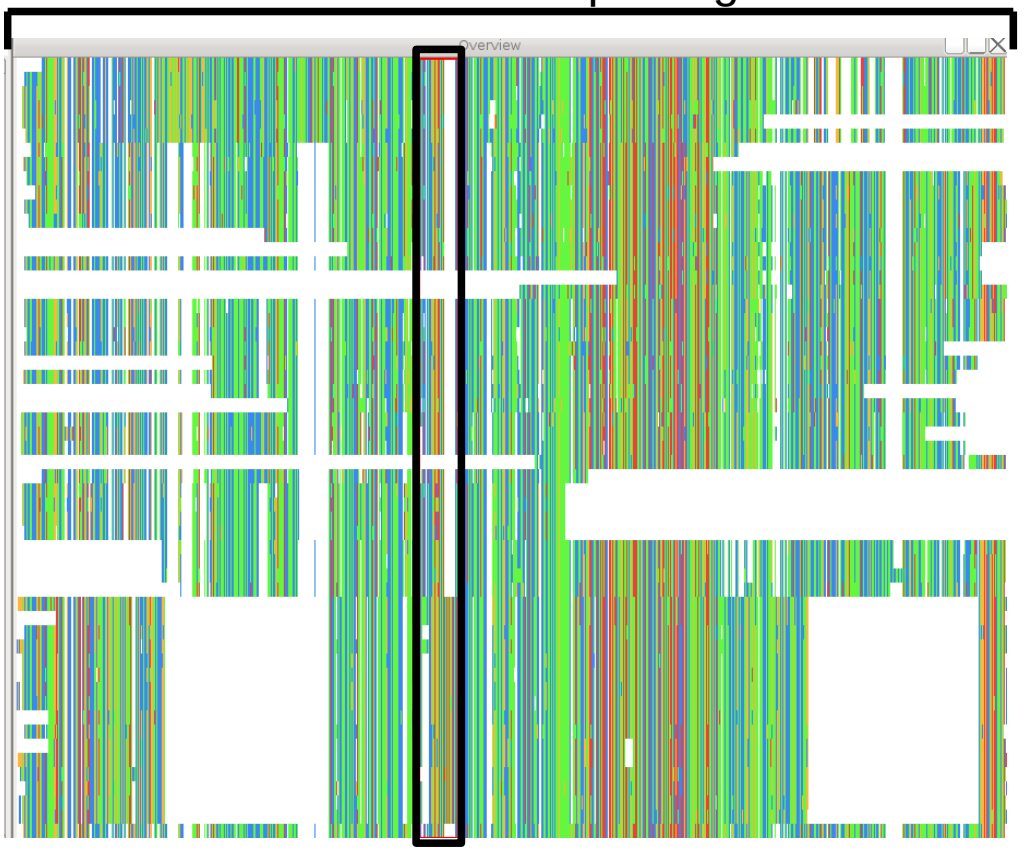


all-by-all matches + consensus + reference sequence

TE families

- ATREP11, non-autonomous Helitron (1053 nt)

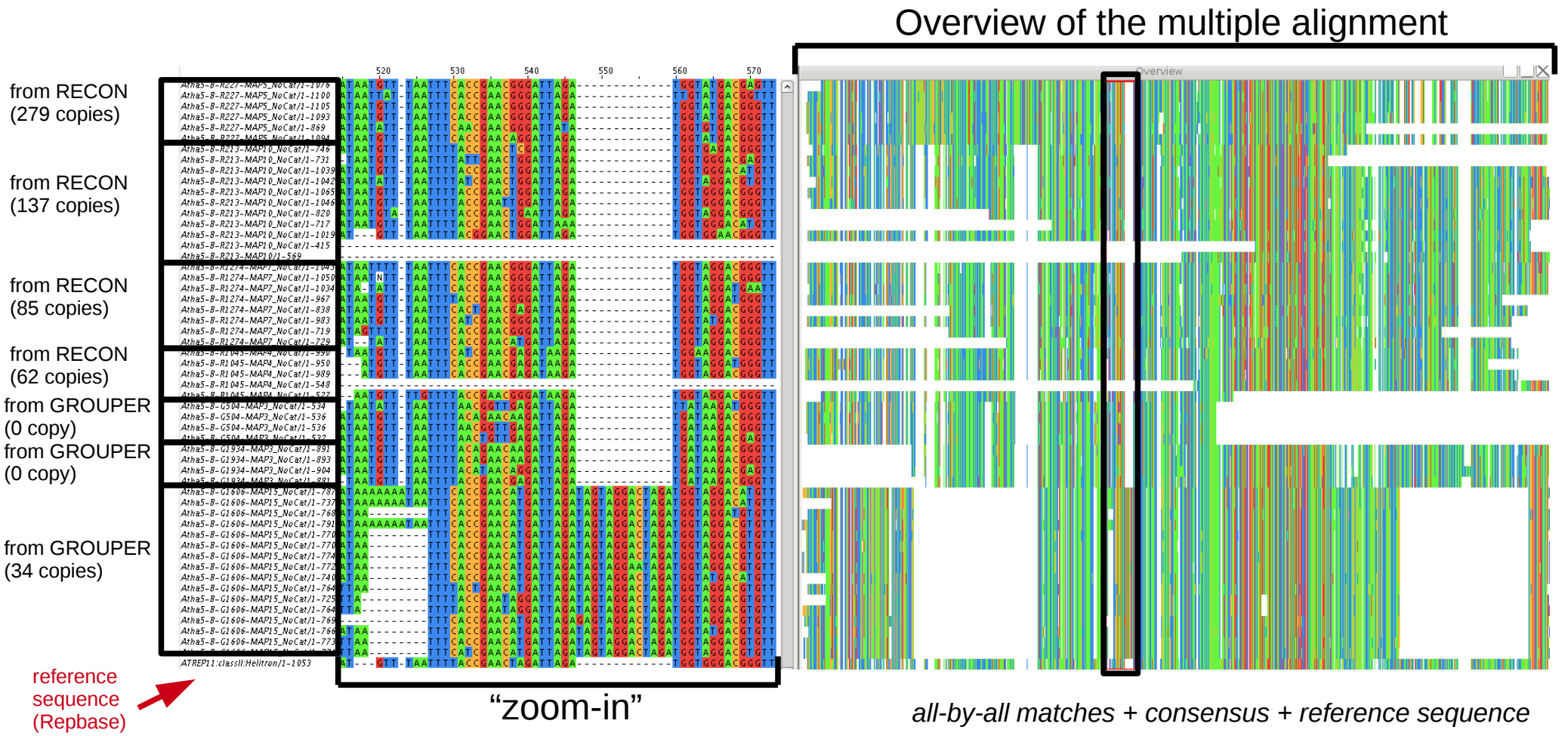
Overview of the multiple alignment



all-by-all matches + consensus + reference sequence

TE families

- ATREP11, non-autonomous Helitron (1053 nt)



TE families

- ATREP11, non-autonomous Helitron (1053 nt)

all-by-all matches + consensus + reference sequence

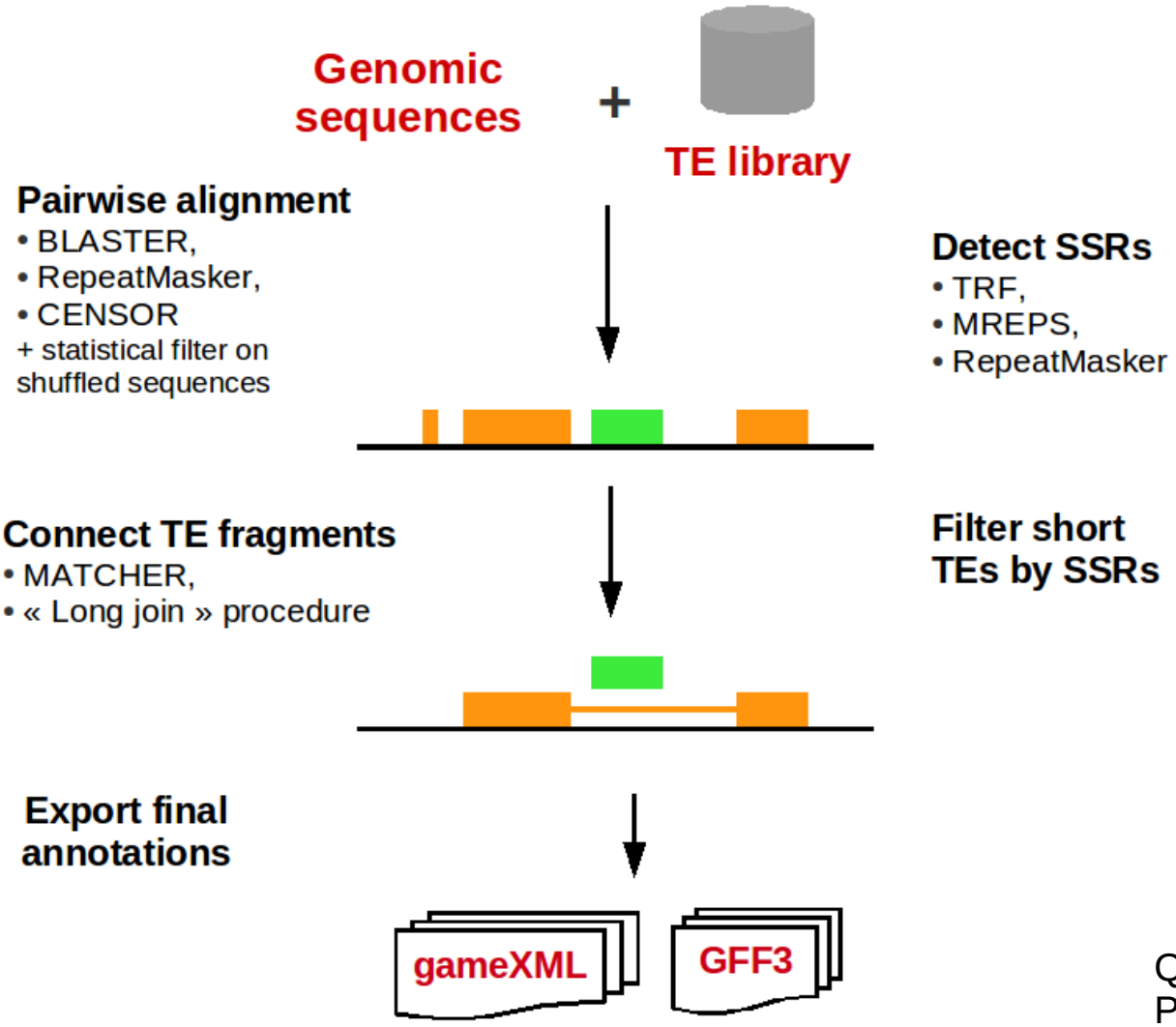
Take-home messages (1)

- Final TE library (classified, non-redundant consensus)
  - *D. melanogaster*: 748 consensus (461 for R+P)
  - *A. thaliana*: 1584 consensus (999 for R+P)
  - *O. sativa*: 5243 consensus (for R+P only)

- ➔ TEdenovo builds TE consensus with **well-defined boundaries**.
- ➔ A TE family can be represented by several consensus for its different **structural variants**.

Genomic annotation of TE copies

TE annotation pipeline



Quesneville *et al.*, PLoS Comp. Bio., 2005

TE fragment connections (1)

- MATCHER:
  - Take all matches found by BLASTER, RepeatMasker and CENSOR;
  - Connect the TE fragments via dynamic programming (find the optimal path);
  - Filter the redundancy and clean the conflicts.

TE fragment connections (1)

- MATCHER:
  - Take all matches found by BLASTER, RepeatMasker and CENSOR;
  - Connect the TE fragments via dynamic programming (find the optimal path);
  - Filter the redundancy and clean the conflicts.

TE annotations **before MATCHER**

TE annotations **after MATCHER**

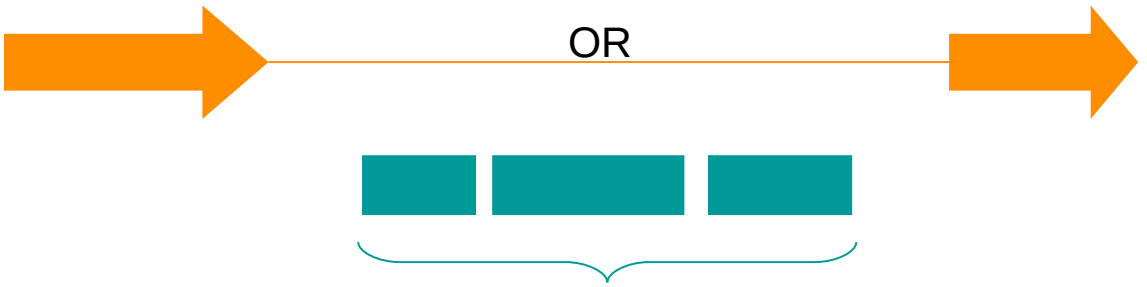
TE fragment connections (2)

- “Long join” procedure:

gap < 5kb

mismatches < 500bp

Δidentity < 2%



Σ TE lengths = 95%

all identities > max(identity1,identity2) - 2%

Results on *A. thaliana*

- Input:
  - TAIR release 9 (120 Mb)
  - TE library of 1531 *de novo* consensus
- Output:
  - 45,558 TE fragments
  - 39,675 TE copies (67 “long joins”)
  - 21.80% genome coverage
  - 115 consensus without any copy

Integration in TriAnnot

- Optimize the statistical filter parameters
- Build a library of consensus by combining TREP with custom sequences from F. Choulet
- Launch the TEannot pipeline as part of the TriAnnot pipeline (block 03 - TE modelling)

Take-home messages (2)

- A full TE annotation requires to **handle nested patterns** and thus to reconstruct TE copies

➔ TEannot implements an **efficient dual strategy** to connect TE fragments:

  - MATCHER
  - “long join” procedure

Authors & contributors: Hadi Quesneville, Timothée Flutre, Olivier Inizan, Claire Hoede, Anna-Sophie Fiston-Lavier, Elodie Duprat, Gaël Faroux, Delphine Autard, Benoît Bely

Acknowledgements

- **URGI, INRA Versailles:** Hadi Quesneville, Emmanuelle Permal, Joëlle Amselem, Olivier Inizan, Claire Hoede, Victoria Dominguez, Isabelle Luyten and Sébastien Reboux
- **GDEC, INRA Clermont:** Catherine Feuillet, Philippe Leroy, Frédéric Choulet
- **Univ. Perpignan:** Olivier Panaud, François Sabot, Cristian Chaparro
- Elodie Duprat (UPMC), Anna-Sophie Fiston-Lavier (Stanford)

Thank you !

Genomes annotated with REPET

Done (8)

- *D. melanogaster* (Flybase) with M. Ashburner (Univ. Cambridge) and C. Bergman (Univ. Manchester)
- *A. thaliana* (TAIR) with V. Colot (ENS) and N. Buisine (MNHN)
- *Fusarium graminearum* with C. Kistler (Univ. Minnesota)
- *Laccaria bicolor* with F. Martin (INRA-Nancy)
- *Meloidogyne incognita* with P. Abad (INRA Antibes)
- *Ectocarpus siliculosus* with M. Cock (Roscoff)
- *Leptospheria maculans* with T. Rouxel (INRA Versailles)
- *Acyrthosiphon pisum* with D. Tagu (INRA Rennes) and A. Wilson (Miami)

In progress …

- *O. sativa* (IRGSC) with O. Panaud (Perpignan)
- *T. aestivum* (IWGSC) with C. Feuillet (INRA Clermont-Ferrand)
- 12 Drosophilidae genomes with D. Petrov (Univ. Stanford)
- *Spodoptera frugiperda*, *Helicoverpa armigera* with P. Fournier (INRA Montpellier)
- *Melampsora larici-populina* with F. Martin (INRA-Nancy)
- *Tuber melanosporum* with F. Martin (INRA-Nancy)
- *Stagonospora nodorum* with T. Rouxel (INRA Versailles)
- *Blumeria graminis* with P. Spanu (Imp. Coll. London)
- *Botrytis cinerea* with M-H. Lebrun (INRA Versailles)
- *Emiliania huxleyi* with Betsy Read (Cal State University)

…