Use cases and improvements of REPET package : pipeline and tools to identify and annotate TEs in genomic sequences

Véronique Jamilloux, Olivier Inizan, Sandie Arnoux, Timothée Flutre, Claire Hoede, Nathalie Choisne, Hadi Quesneville

Use cases and improvements of REPET package : pipelines and tools to identify and annotate TEs in genomic sequences

CNET Lyon 4th – 6th July 2011

V. Jamilloux

Introduction

- **TEs play a major role in structure, functions and evolution of genomes.**
- **This role is relative to their particular dynamics: burst and degeneration.**
- **Their nature and their mobile capability make them very particular biological elements**

Goals

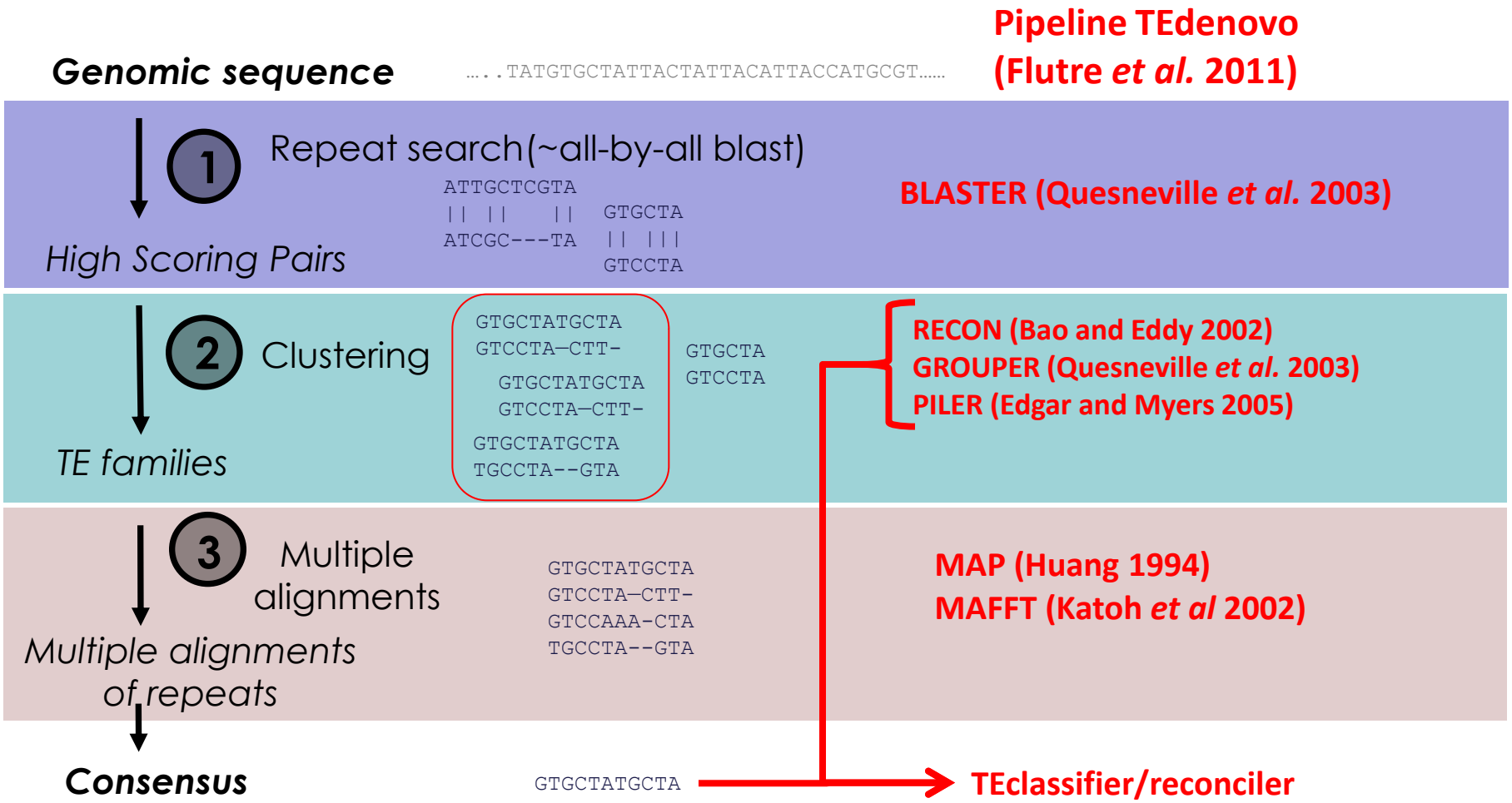
- **To correctly annotate TEs, the tools have to take into account these specificities and the TE diversity**

- **With the acceleration of genome sequencing, the tools have to treat lot of data and to annotate automatically to help the biologist**

REPET package

- **Automatic detection and annotation of repeats in genomic sequences**
- **Version 1.4 available http://urgi.versailles.inra.fr/tools/REPET**
- **2 pipelines TEdenovo and TEannot**
- **Several tools to analyze the different results of the 2 pipelines**
- **Run on a compute cluster, SGE**
- **Results files and database (MySQL)**
- **Development and validation phases on benchmark with *D. melanogaster* (release 4) and *A. thaliana* (release 9)**

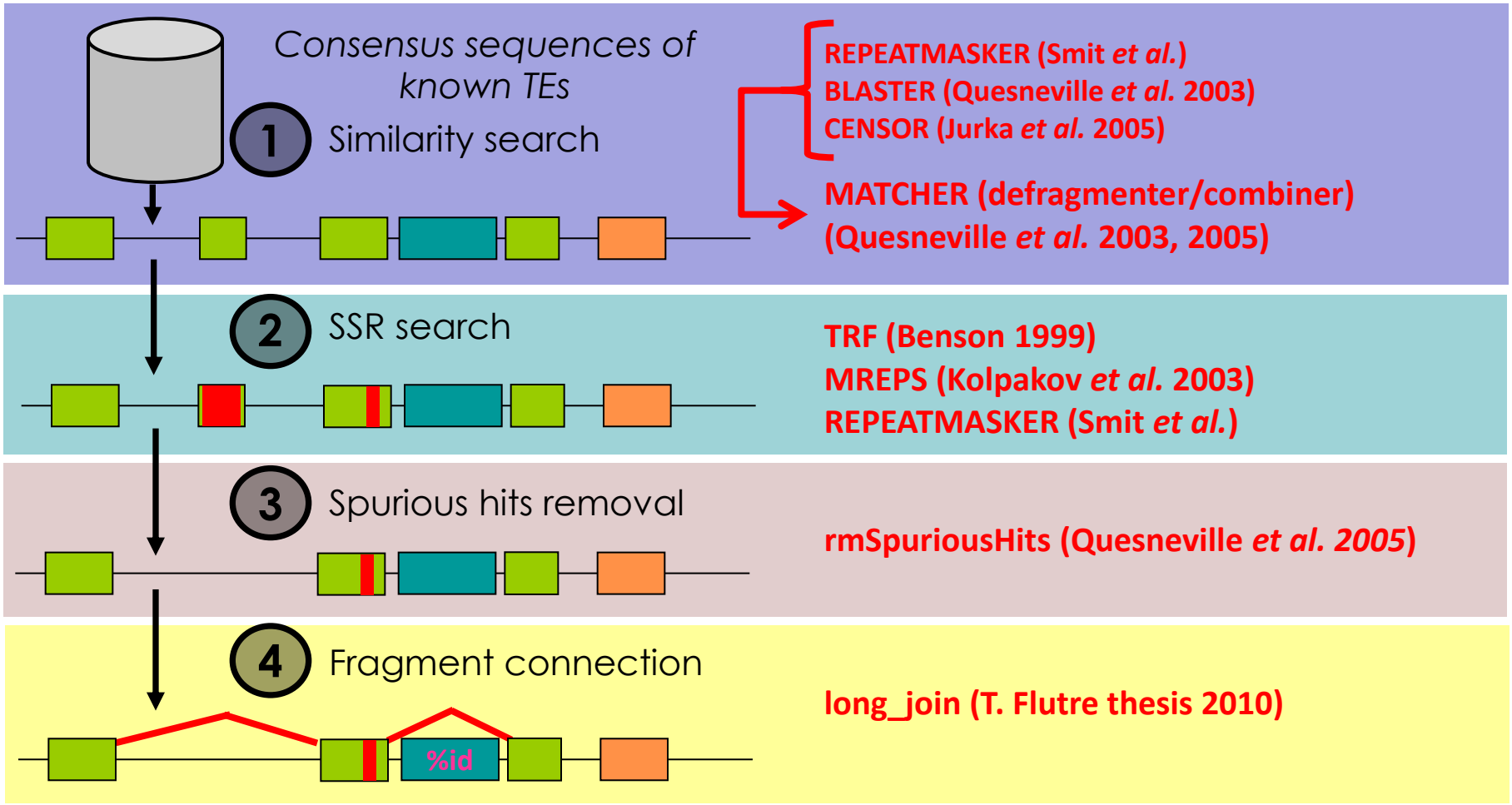
TEdenovo


Pipeline TEdenovo
(Flutre et al. 2011)
Genomic sequence
…..TATGTGCTATTACTATTACATTACCATGCGT……
1
Repeat search(~all-by-all blast)
ATTGCTCGTA
|| || || GTGCTA
ATCGC---TA || |||
GTCCTA
BLASTER (Quesneville et al. 2003)
High Scoring Pairs
2
Clustering
GTGCTATGCTA
GTCCTA—CTT-
GTGCTATGCTA
GTCCTA—CTT-
GTGCTATGCTA
TGCCTA--GTA
GTGCTA
GTCCTA
RECON (Bao and Eddy 2002)
GROUPER (Quesneville et al. 2003)
PILER (Edgar and Myers 2005)
TE families
3
Multiple alignments
GTGCTATGCTA
GTCCTA—CTT-
GTCCAAA-CTA
TGCCTA--GTA
MAP (Huang 1994)
MAFFT (Katoh et al 2002)
Multiple alignments of repeats
Consensus
GTGCTATGCTA
TEclassifier/reconciler

2 new steps included

- **Filter out SSR and not classified consensus (built from less than 10 sequences).**
- **Consensus clustering (with Blastclust) to investigate the similarity relationships among the consensus that have been built.**

TEannot


Consensus sequences of known TEs
1
Similarity search
REPEATMASKER (Smit et al.)
BLASTER (Quesneville et al. 2003)
CENSOR (Jurka et al. 2005)
MATCHER (defragmenter/combiner) (Quesneville et al. 2003, 2005)
2
SSR search
TRF (Benson 1999)
MREPS (Kolpakov et al. 2003)
REPEATMASKER (Smit et al.)
3
Spurious hits removal
rmSpuriousHits (Quesneville et al. 2005)
4
Fragment connection
%id
long_join (T. Flutre thesis 2010)

Tools

- **ClusterConsensus, 3 steps :**
  - Consensus clustering with Blastclust
  - Data organization : 1 directory by cluster (considered like a TE family), containing fasta files (consensus, “all-by-all” sequences)
  - For each cluster, building a multiple alignment and phylogeny (between consensus and “all-by-all” sequences or annotated copies)

- **GiveInfoTEannot**
  - After the TE annotation, this tool gives some statistics about genome annotation (coverage, copies, …)

More than 35 annotated genomes

Espèces	*Publications*
D. melanogaster	*Quesneville et al.* ***PLoS Comp. Biol 2005*** *Bergman et al.* ***Genome Biology 2006*** *Browsable at:* ***http://flybase.org***
12 Drosophila genomes	*Clark AG et al.* ***Nature. 2007***
A. thaliana	*Buisine et al.* ***Genomics. 2008*** *Browsable at:* ***http://www.arabidopsis.org***
Fusarium graminearum	*Cuomo et al.* ***Science. 2007***
Rice	
Laccaria bicolor	*Martin et al.* ***Nature. 2008***
Meloidogyne incognita	*Abad et al.* ***Nature Biotech 2008***
Acyrthosiphon pisum	*IAGC* ***PLoS Biology 2010***
Aedes aegypti	*Nene et al.* ***Science. 2007***
Spodoptera frugiperda, Helicoverpa armigera	*Alençon et al.* ***PNAS 2010*** *Browsable at:* ***http://genouest.org/***
Tuber melanosporum	*Martin et al.* ***Science 2010***
Ectocarpus siliculosus	*Cock et al.* ***Nature 2010.***
Leptospheria maculans	*Rouxel et al* ***Nature Comm 2011***
Blumeria graminis	*Spanu et al* ***Science 2010***
Melampsora larici-populina and Puccinia graminis	*Duplessis et al. (submitted)*

Several use cases

- *Vitis vinifera* genome 12X
- *Blumeria graminis* v3
- *Aphid Acyrthosiphon*
- *Ectocarpus*
- *Nicotiana tabacum*
- *Ixodes scapularis*
- *Brassica rapa*

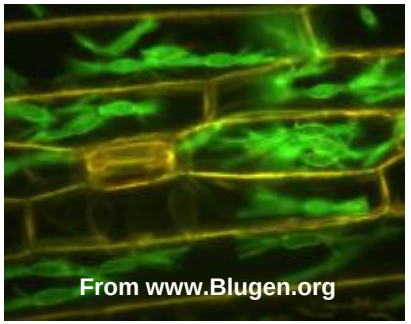
Vitis vinifera

- **Analysis aim :** Study the TE distribution and their evolution.
  For REPET, improvement for larger genomes
- **The genome :** 486 Mbp, whole genome 33 built chromosomes, 12X
- **REPET package v1.3.12.1** , adapt settings for sensitivity and parallel computing.
- **Results** :
  - TEdenovo : a library of 6315 TE consensus (11621 consensus before SSR and noCat filtering)
  - Classification of the 6315 TE consensus :

Class I => 69,90% ; Class II => 15,15% ; Confused/Unknown => 14,95%

  - TEannot : 375737 copies (8425 full length copies) using 6315 TE consensus => 65,07% of the genome

Blumeria graminis



- **Analysis aim :** to annotate the genome and to learn about the TE impact on genome dynamics and evolution.
- **The genome :** 130 Mbp, whole genome sequences 8X
- **REPET package v1.4** with a new setting on Grouper (TEdenovo step 2). This specific setting improves clustering in order to avoid false positive chimeric consensus.
- **Results** :
  - TEdenovo : a library of 2226 TE consensus
  - Classification of the 2215 TE consensus :

  Class I => 69,03% ; Class II => 1,40% ; Confused/Unknown => 29,57%
  - TEannot :165505 copies using 2215 TE consensus => 75,87% of the genome
  - Genome coverage according to classification :

  Class I => 60,33% ; Class II => 1,18% ; Confused/Unknown => 14,36%

See J. Amselem's talk : "Fungal genome analysis of TEs".

Aphid acyrtosiphon

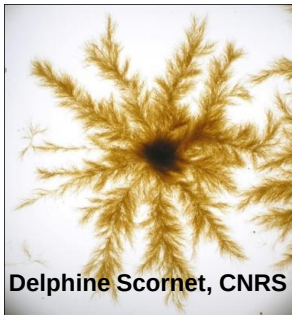
- **Analysis aim :** This genome is holocentric, make the TEs annotation and see their distribution.
- **The genome :** 464 Mbp, (1130 scaffolds)
- **REPET package V1 with default settings**
- **Results** :
  - TEdenovo : a library of 1883 TE consensus (with manual curation)
  - Classification of the 1883 TE consensus :

  Class I => 19,60% ; Class II => 16,46% ; unclassified => 64,58%
  - TEannot : 506309 copies using these 1883 consensus => 37,86% of the genome.
  - Genome coverage according to classification :

  Class I => 5,43% ; Class II =>5,31% ; unclassified => 27,12%;

See : "Genome sequence of the Pea Aphid *Acyrthosiphon pisum*". Plos biology, Feb 2010

Ectocarpus siliculosus


Delphine Scornet, CNRS

- **Analysis aim :** identify the TE complement to evaluate their contribution to the genome size and to analyze the phylogenetic relationship with LTR retrotransposons from other micro algae.
- **The genome :** 195 Mbp, 1561 scaffolds (N50 = 504,468)
- **REPET package V1**
- **Results** :
  - TEdenovo : a library of 770 TE consensus (with manual curation)
  - Classification of the 770 TE consensus :

  Class I => 9,35% ; Class II => 8,18% ; Tandem repeat => 2,99% ; unclassified => 79,48%
  - TEannot : 95004 copies using these 770 consensus => 22.4% of the genome.
  - Genome coverage according to classification :

  Class I => 9,4% ; Class II => 2,7% ; Tandem repeat => 0,4% ; unclassified => 9,9%

See F. Maumus's talk : "Transposable elements in marine stramenopiles".

Nicotiana tabacum

- **Analysis aim :** Identification of new retrotransposons families and estimation of their copy numbers
- **Initial genome data** : 382 Mbp (300158 seq from SGN assembly)
- **Part of genome analyzed :** 72,2Mbp (~1,6% of estimated genome size); (16892 contigs > 3 kpb) , N50 => 4116;
- **REPET package : TEdenovo v1.3.12.1 and TEannot v1.4** with default settings.
- **Results** :
  - TEdenovo : a library of 1022 TE consensus
  - Classification of the 1022 TE consensus :

Class I => 3,52% ; Class II => 1,17% ; SSR => 0,19%; unclassified => 95,1%

  - TEannot : 60504 copies using 1022 TE consensus => 36,30% of the genome.

Ixodes scapularis

- **Analysis aim :** TEs like characterization tool for polymorphism, MITEs because they are near the genes
- **Initial genome data** : 1,39 Gbp (570 637 contigs); genome estimated size 2,1 Gbp
- **Part of genome analyzed :** 694,2 Mbp; (88 851contigs)
- **REPET package V 1.3.9** with setting MITE_max_length: 800 for TEclassifier.
- **Results** :
  - TEdenovo : a library of 3802 TE consensus
  - Classification of the 3802 TE consensus :

  Class I => 28,09% ; Class II => 18,62% (MITEs =>5,6%) ; unclassified => 53,29%
  - TEannot : Only on contigs > 10 Kbp, 24035 copies using 1789 TE consensus (LTR, LINE, TIR, MITE=> 9%) => 10,26% of the analyzed genome (2,05% of estimated genome).

Master 2 report "Recherche *in silico* des éléments transposables dans le génome de la tique *ixodes scapularis*" S. Aubernon

Brassica rapa

- **Analysis aim :** To predict Brassica TEs and then study their activation during the polyploidization in *brassica napus.*
- **Initial genome data :** 529 Mpb
- **Part of genome :** 184,4Mbp; (1377 BACs) , N50 => 130740; (34,82% of the genome)
- **REPET package : V1.3.12.1** with default settings.
- **Results** :
  - TEdenovo : a library of 4123 TE consensus after redundancy elimination
  - Classification of the 4123 TE consensus :

Class I => 15,45% ; Class II => 9,99% ; SSR => 0,41%; unclassified => 74,14%

Note : V. Sarilar's thesis "Activation des éléments transposables suite à la polyploïdisation du colza *Brassica napus*"

Few useful tips

- **Analyze your data**
  - Sort your sequences according to their size and filter out short ones
- **Specify the biological question**
  - Help for parameters settings
- **Contact: urgi-repet@versailles.inra.fr**
  - Support
  - Collaborations within projects

Perspectives

- **TEdenovo**
  - Add a new dedicated structural detection tool : LTRharvest
  - Improve combined approach. See poster P8 «Strategies for TEs detection among genomes » O. Inizan
  - Replace the classifier with PASTEC. See Poster P9 “PASTEC” S. Arnoux poster
- **ClusterConsensus:**
  - Improve the analysis of structural variants.
- **GiveInfoTEannot**
  - Improve statistics (e.g. by cluster of consensus)

Acknowledgments

- **Special thanks**

  O. Inizan, S. Arnoux, T. Flutre, C. Hoede, T. Chaumier and H. Quesneville

- **For their data**
  - N. Choisne *(Vitis)*, J. Amselem *(Blumeria)*
  - F. Maumus *(Ectocarpia)*, E. Permal *(A. pisum)*,
  - M. Deloger *(Nicotiana)*, V. Sarilar *(Brassica)*, S. Aubernon *(I. scapularis)*

- **All URGI team members**

Thank you for your attention