PhD defense - Université Paris Diderot

The annotation of transposable elements through the understanding of their diversification

Timothée Flutre

Institut National de la Recherche Agronomique
Unité de Recherche en Génomique-Info

October the 28^{th}, 2010

PhD directors: Hadi Quesneville (INRA, URGI) and Catherine Feuillet (INRA, GDEC)

Outline

Outline

1. Introduction on TE discovery and their impacts on genomes
   - Brief history on TEs, their function and evolution
   - Impacts of TEs and their structural classification
2. Considering TE diversification in *de novo* annotation approaches
   - Aim, approach and pitfalls
   - Comparative analysis and combined approach
   - Identification of TE families and their structural variants

Outline

1. Introduction on TE discovery and their impacts on genomes
   - Brief history on TEs, their function and evolution
   - Impacts of TEs and their structural classification
2. Considering TE diversification in *de novo* annotation approaches
   - Aim, approach and pitfalls
   - Comparative analysis and combined approach
   - Identification of TE families and their structural variants
3. Discussion and perspectives
   - Application of the tools and practical consequences
   - A new method to detect TEs and study their diversification

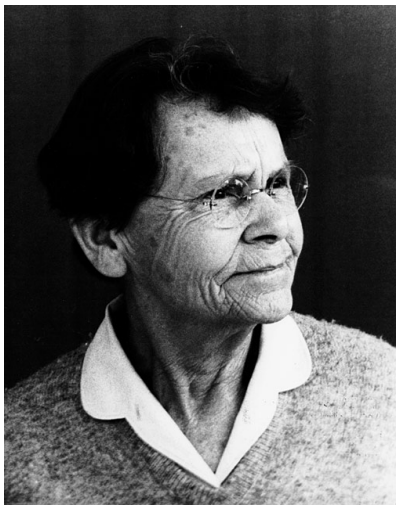
Outline

Outline

Context of the TE discovery

Barbara McClintock was a plant geneticist at Cold Spring Harbor:

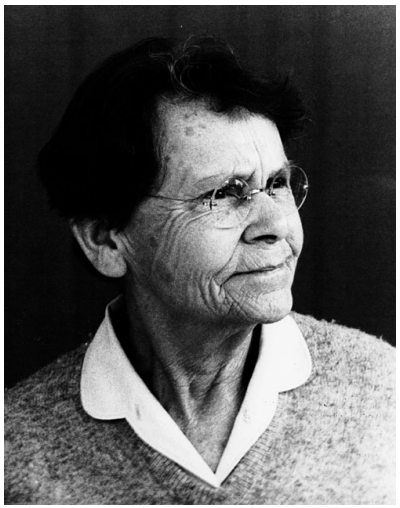
- in 1944, cross of two maize lines and "genetic earthquake";
- in 1950, publication on mutable loci (Ac-Ds);
- in 1956, publication on "controlling elements".



Context of the TE discovery

Barbara McClintock was a plant geneticist at Cold Spring Harbor:

- in 1944, cross of two maize lines and "genetic earthquake";
- in 1950, publication on mutable loci (Ac-Ds);
- in 1956, publication on "controlling elements".



DNA was known to carry genes only since 1944, and its structure was discovered only in 1953: what to think about "jumping" genes?

TE definition, function and evolution

- TEs are mobile genetic elements (DNA sequences).
- They can multiply within genomes (non-Mendelian inheritance).
- They are present in virtually all species (horizontal transfers).

TE definition, function and evolution

- TEs are mobile genetic elements (DNA sequences).
- They can multiply within genomes (non-Mendelian inheritance).
- They are present in virtually all species (horizontal transfers).

For decades since their discovery, TE existence, amount and ubiquity were key arguments in several major debates in theoretical biology:

- archetypes of the selfish gene, seen as ultimate parasites;
- only "junk DNA" or potential exaptations;
- possible role in the regulation of gene expression.

Annotation of TE content in sequenced genomes

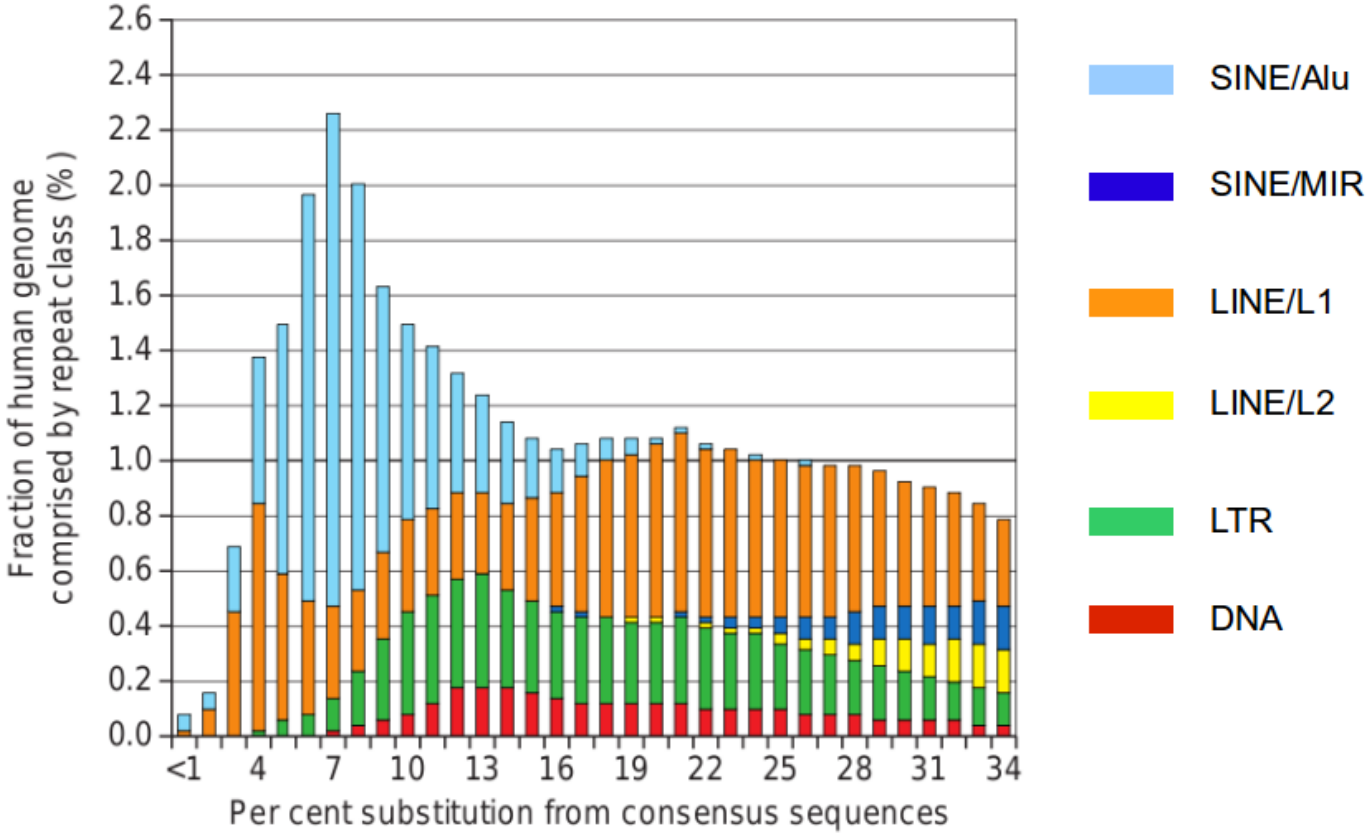


Figure: Relative age distribution of TE copies representing $\approx$ 45% of the 3-Gb human genome, annotated with known TE sequences (Lander *et al.*, 2001).

Annotation of TE content in sequenced genomes

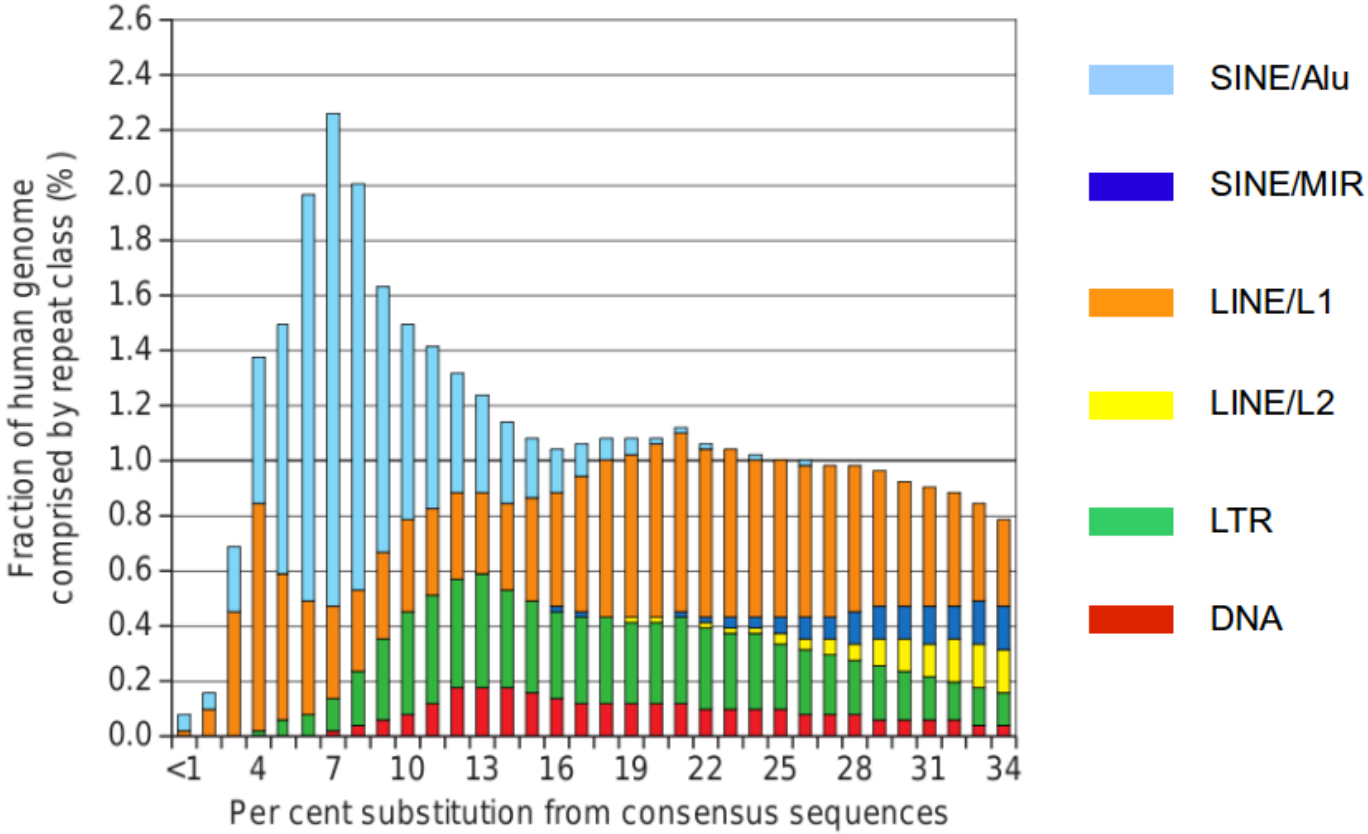


Figure: Relative age distribution of TE copies representing $\approx$ 45% of the 3-Gb human genome, annotated with known TE sequences (Lander *et al.*, 2001).

For non-model species, *de novo* approaches are required.

Outline

Impacts of TEs on gene expression

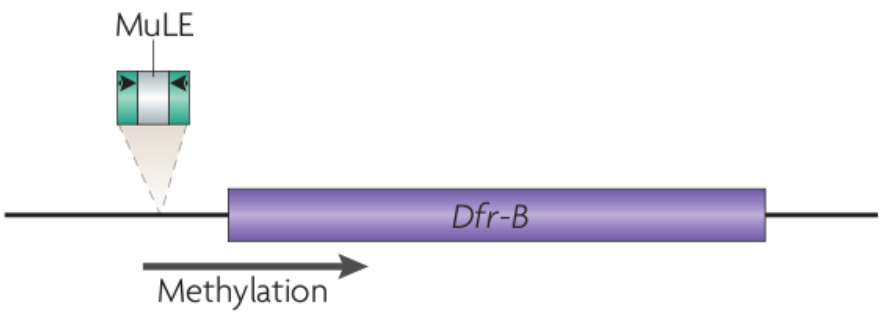


Figure: DNA methylation of a TE can spread to the promoter of a neighbouring gene, such as in morning glory flowers, causing petal-colour streaks (Slotkin and Martienssen, 2007).

Impacts of TEs on gene expression

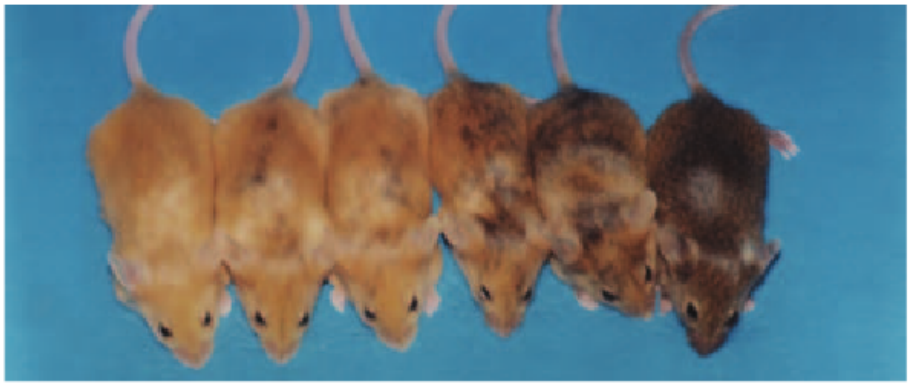


Figure: TEs can produce outward-reading transcripts that extend into neighboring genes, like here in mice where the transcription level of the *agouti* gene depends on the epigenetic status of a TE, thereby influencing coat darkness (Slotkin and Martienssen, 2007).

Impacts of TEs on gene expression

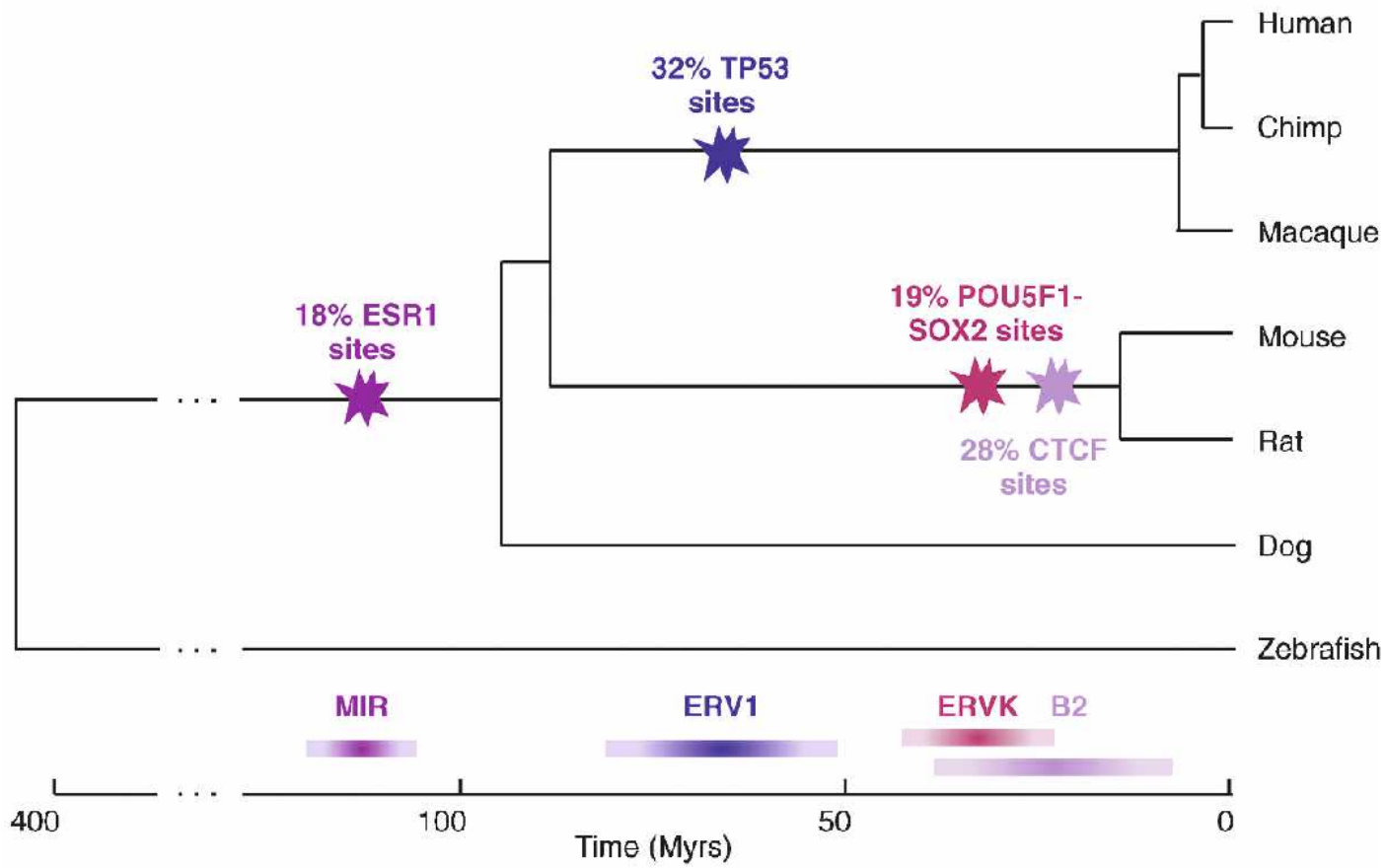


Figure: TE dynamics generate sequence diversity with potential impacts on regulatory networks, as shown here with the evolution of some mammalian transcription factor binding sites embedded within TEs (Bourque *et al.*, 2008).

Unified classification of eukaryotic TEs

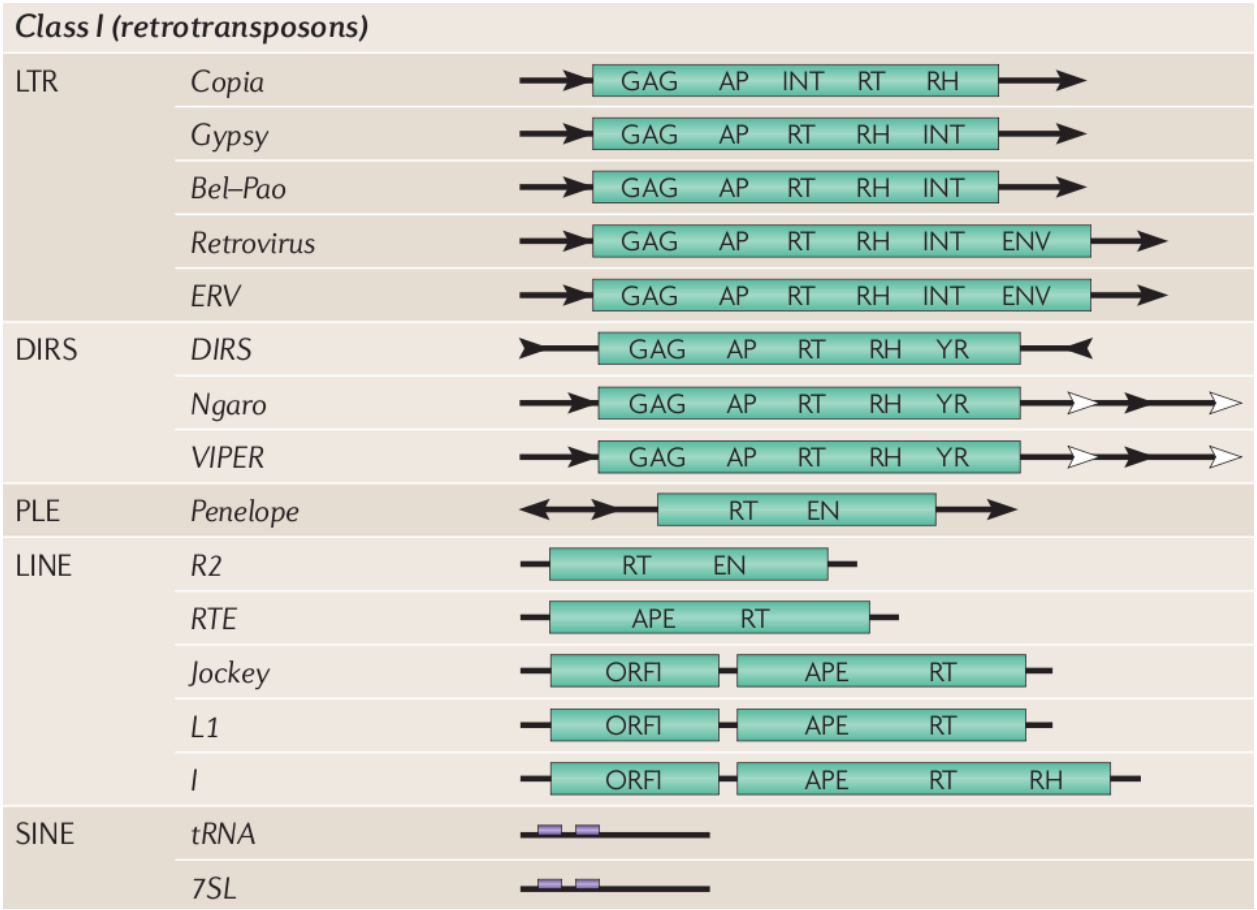


Figure: TEs from class I (Wicker *et al.*, 2007).

Unified classification of eukaryotic TEs

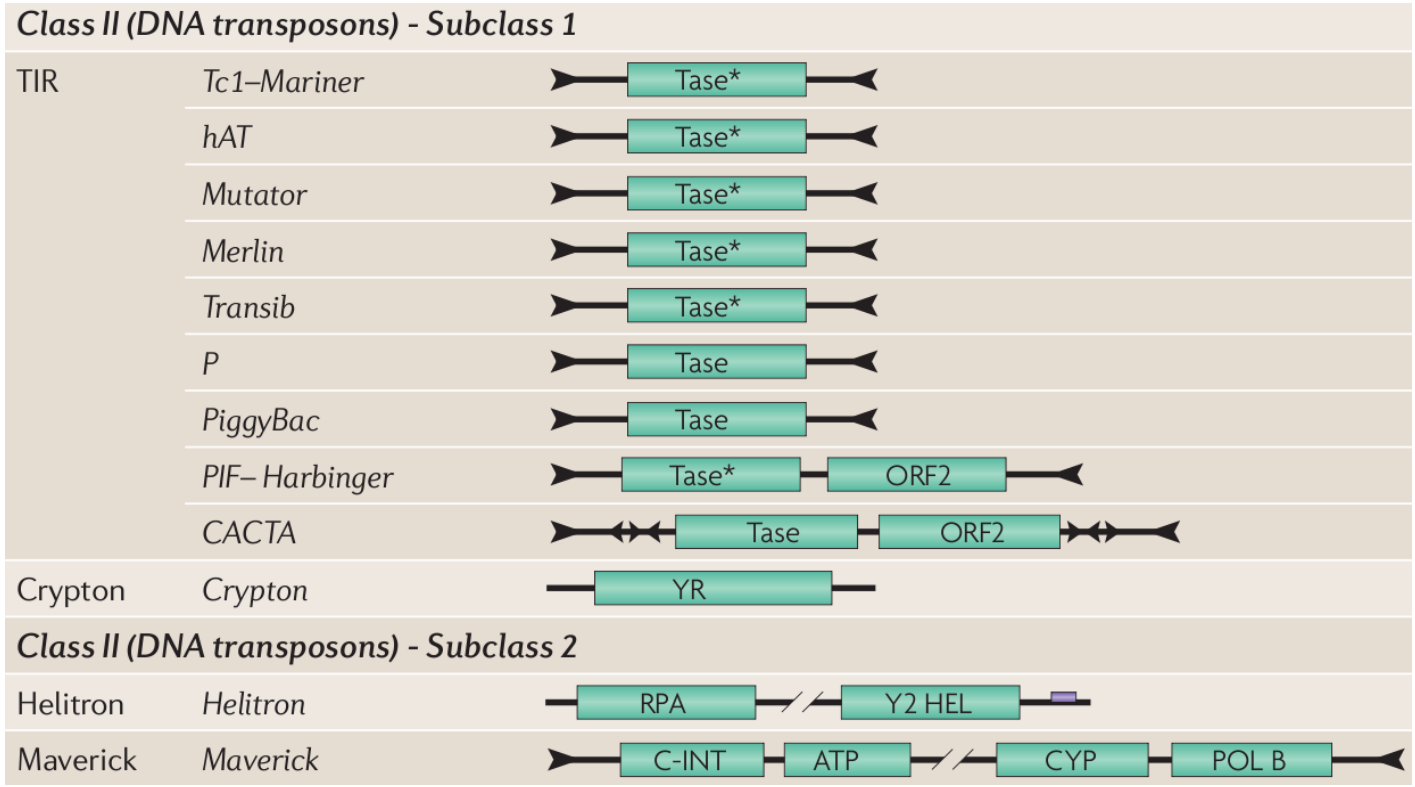


Figure: TEs from class II (Wicker *et al.*, 2007).

Presence of TE structural variants in genomes

However, several examples of TE structural variants have been found, which are not all easily accounted for by the current classification:

- non-autonomous TEs deriving from autonomous ones;
- chimeras resulting from the combination of two different TEs;
- TEs containing gene fragments from one or several genes.

Presence of TE structural variants in genomes

However, several examples of TE structural variants have been found, which are not all easily accounted for by the current classification:

- non-autonomous TEs deriving from autonomous ones;
- chimeras resulting from the combination of two different TEs;
- TEs containing gene fragments from one or several genes.

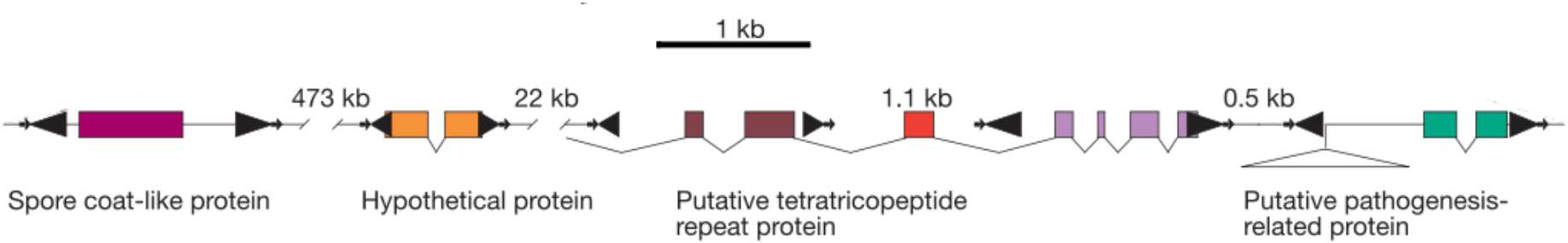


Figure: Examples of Pack-MULEs in the rice genome (Jiang *et al.*, 2004).

Questions

- What are the basic evolutionary principles from which the existing *de novo* programs were built?

Questions

- What are the basic evolutionary principles from which the existing *de novo* programs were built?
- How reliable are their predictions? How is their quality assessed?

Questions

- What are the basic evolutionary principles from which the existing *de novo* programs were built?
- How reliable are their predictions? How is their quality assessed?
- Are they able to recover and distinguish structural variants belonging to the same TE family?

Questions

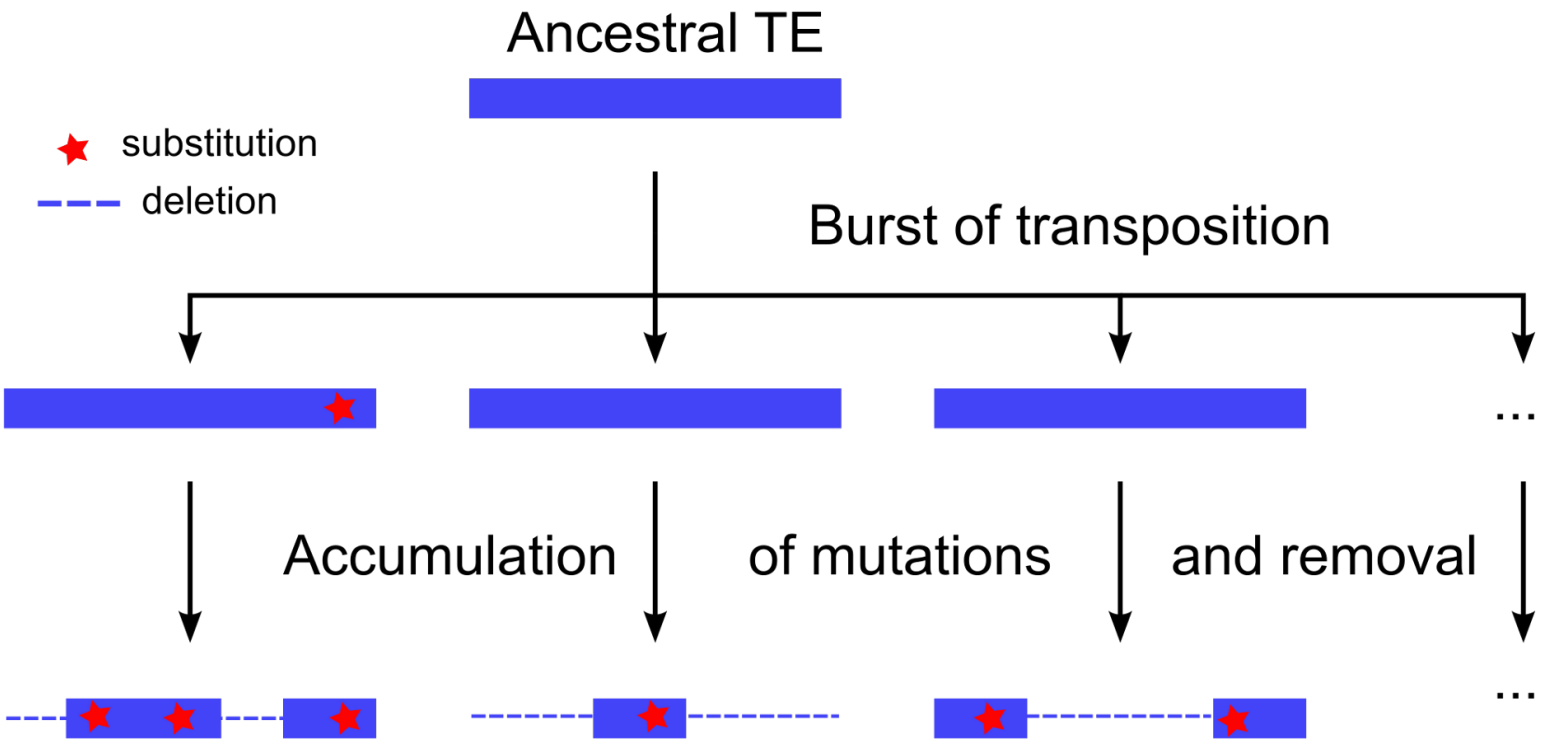
- What are the basic evolutionary principles from which the existing *de novo* programs were built?
- How reliable are their predictions? How is their quality assessed?
- Are they able to recover and distinguish structural variants belonging to the same TE family?
- How can we extend their scope and implement a robust tool to automatically annotate the TE content of newly sequenced genomes?

Questions

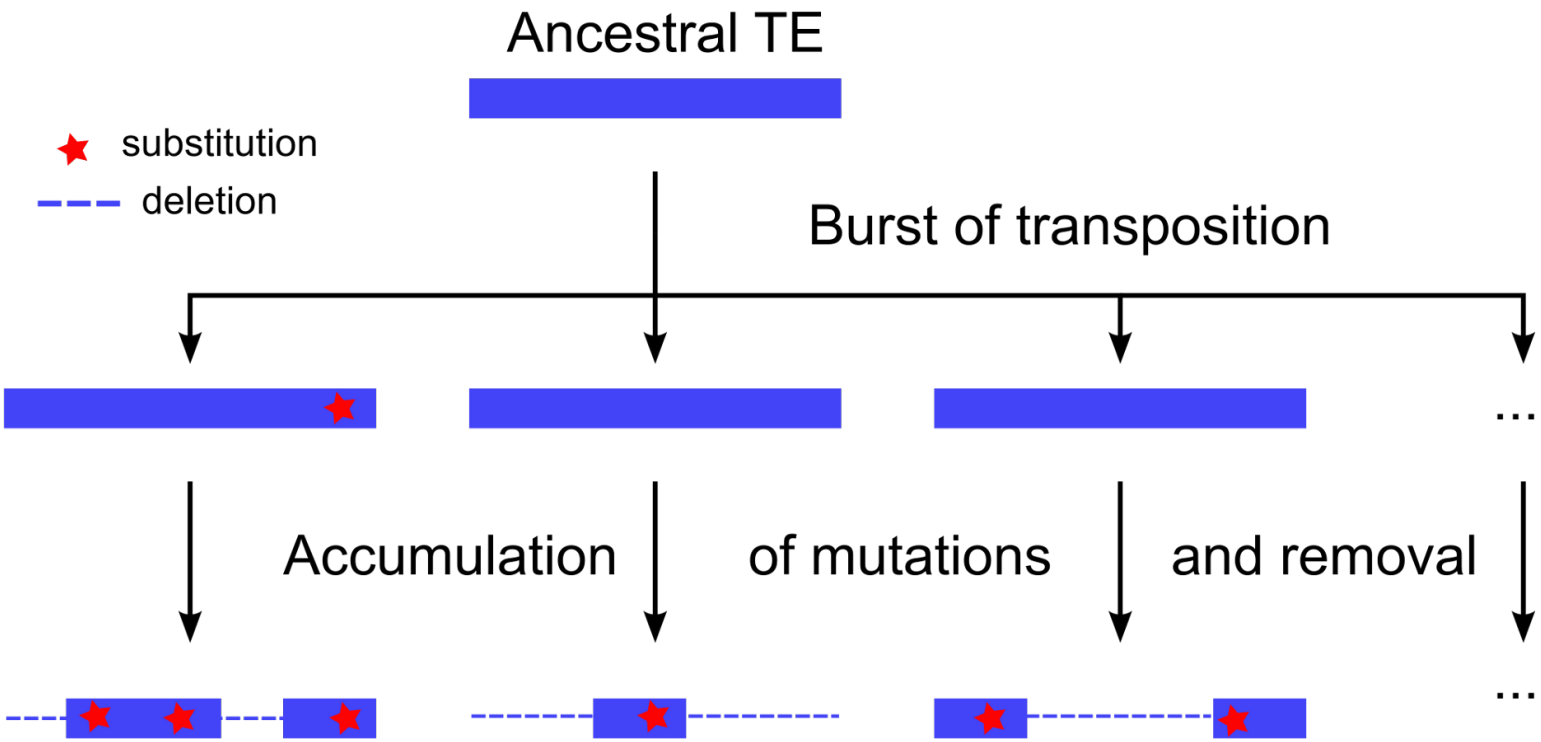
- What are the basic evolutionary principles from which the existing *de novo* programs were built?
- How reliable are their predictions? How is their quality assessed?
- Are they able to recover and distinguish structural variants belonging to the same TE family?
- How can we extend their scope and implement a robust tool to automatically annotate the TE content of newly sequenced genomes?
- What can we learn on TEs and genome biology in general using a *de novo* approach for TE detection?

Outline

Usual model of TE dynamics



Usual model of TE dynamics



In this model of a TE family, the aim is to automatically recover the full ancestral TE as one consensus sequence.

Build *de novo* TE consensus

Raw genomic sequences
(pseudomolecules,
BACs, contigs, ...)

↓

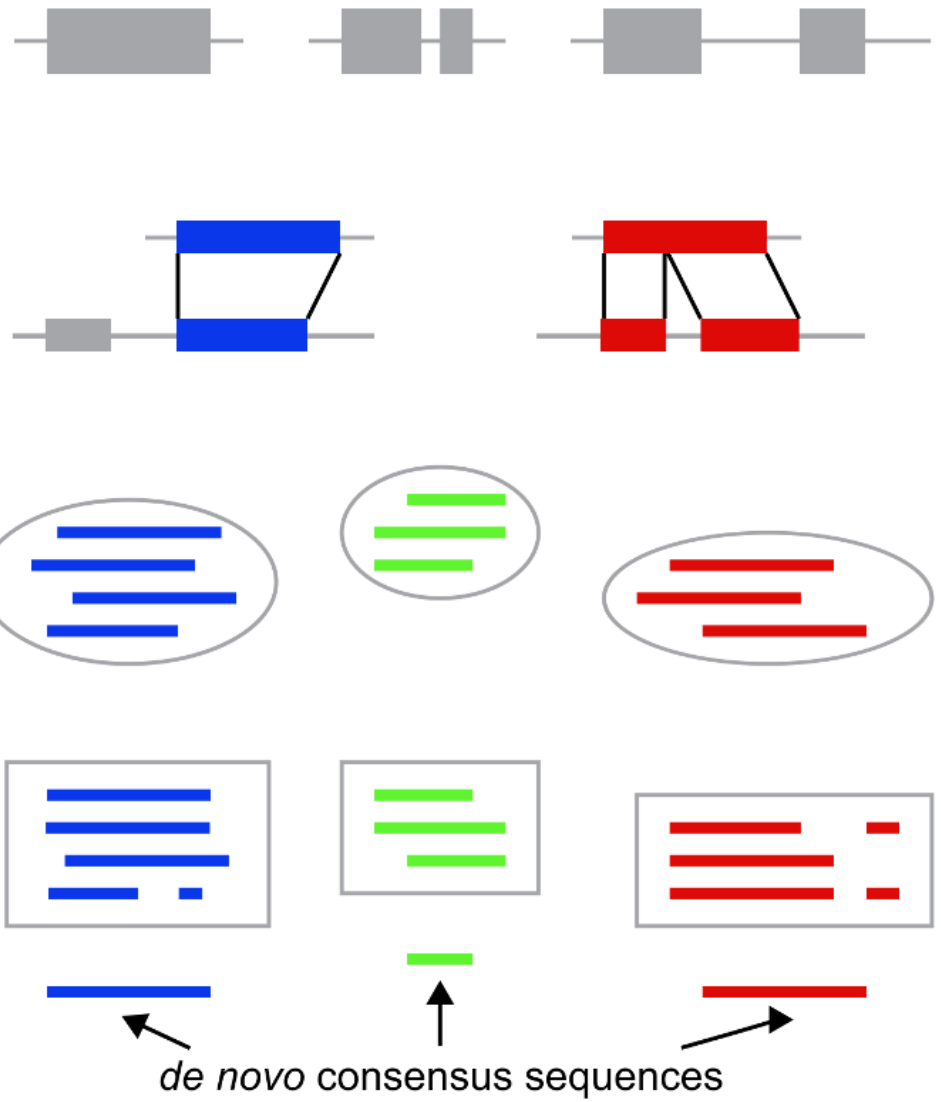
Self-alignment
(BLASTER or PALS)

↓

Clustering
(GROUPER, RECON
or PILER)

↓

Multiple alignment
(MAP, MAFFT, MUSCLE,
PRANK or CLUSTAL-W)



Material and datasets

We chose 2 well-known, small genomes to perform the analysis:

- *D. melanogaster*, release 4, 129 Mb (mainly euchromatin)
- *A. thaliana*, release 9, 119 Mb

Material and datasets

We chose 2 well-known, small genomes to perform the analysis:

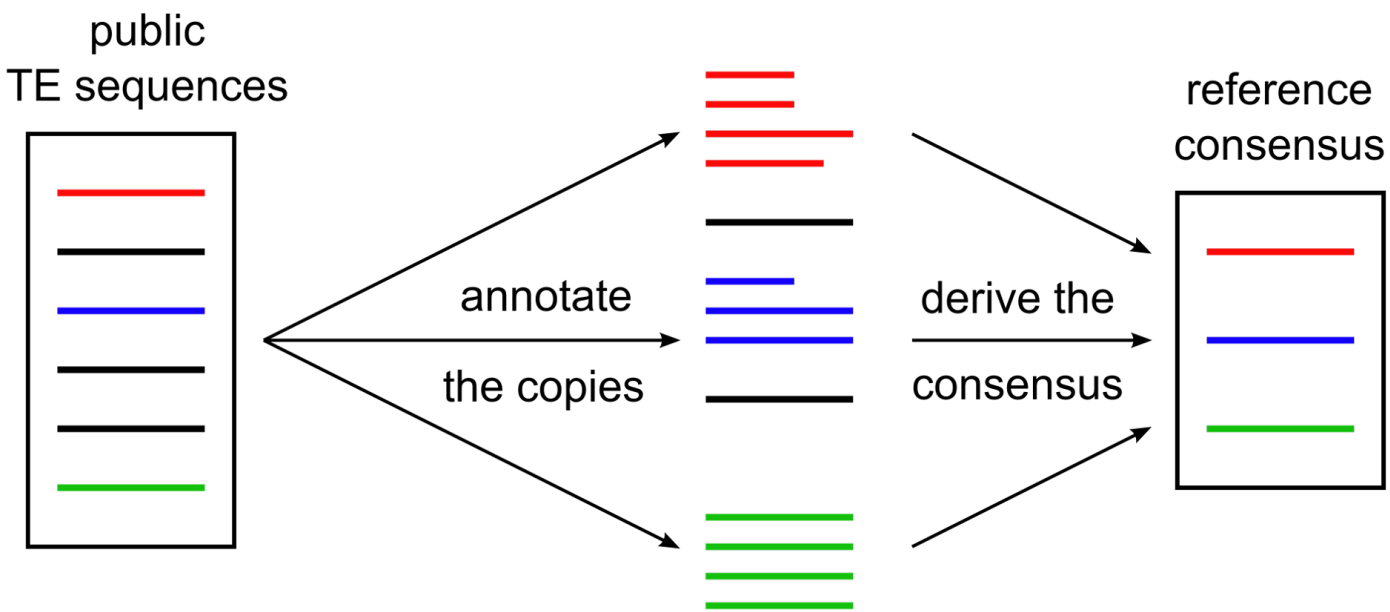
- *D. melanogaster*, release 4, 129 Mb (mainly euchromatin)
- *A. thaliana*, release 9, 119 Mb

Moreover, known TE sequences are publicly available for these genomes:

- Berkeley Drosophila Genome Project: 126 sequences
- Repbase for *A. thaliana*: 318 sequences.

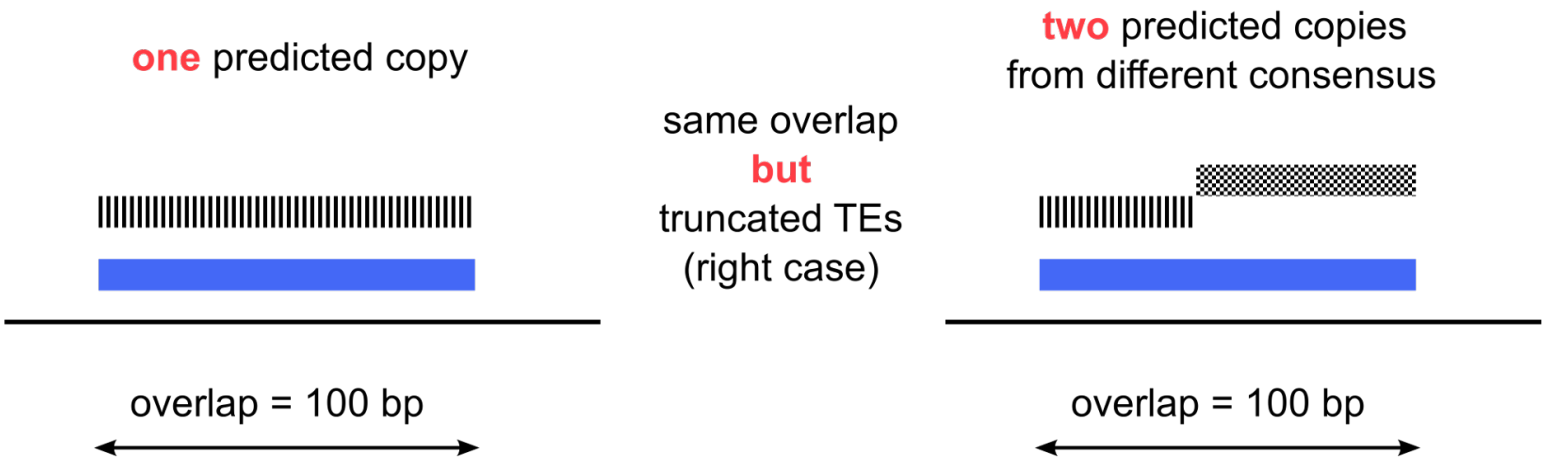
Build appropriate reference datasets

To prevent building *de novo* consensus not corresponding to TEs (e.g. duplications), only TEs with 3 copies are considered. Therefore, I built appropriate reference databanks (*ref*) for both model genomes.



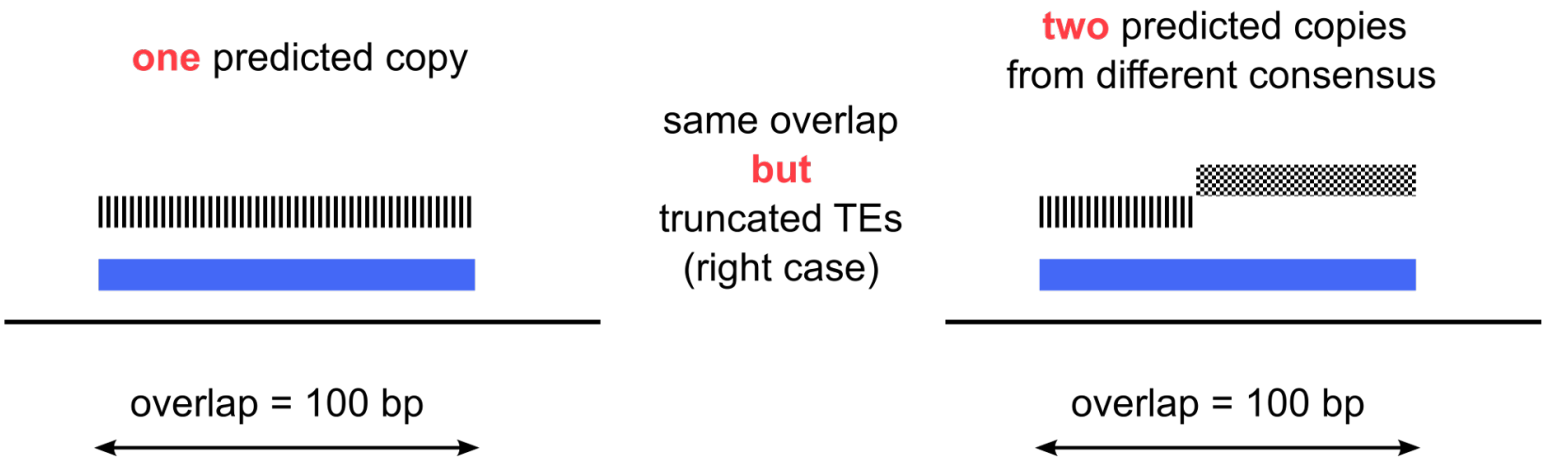
Improve the test protocol of the *de novo* approach

Usually, to check the quality of a *de novo* approach, one computes the nucleotide overlap between predicted copies and known copies:



Improve the test protocol of the *de novo* approach

Usually, to check the quality of a *de novo* approach, one computes the nucleotide overlap between predicted copies and known copies:



Here, need also to align the *de novo* consensus with the *ref* consensus:

- S'_n: proportion of *ref* consensus matching *de novo* consensus;
- S'_p: proportion of *de novo* consensus matching *ref* consensus;
- R_{CC}: close to 1 $\rightarrow$ many *ref* consensus are fully recovered.

Outline

Self-alignment of the genomic sequences

- Align the whole genome with itself using stringent parameters:
  - BLASTER (Quesneville *et al.*, 2003): BLAST wrapper, heuristic
  - PALS (Rasmussen *et al.*, 2006): find all local alignments over a given length with an error rate lower than a given threshold, exact algorithm

Self-alignment of the genomic sequences

- Align the whole genome with itself using stringent parameters:
  - BLASTER (Quesneville *et al.*, 2003): BLAST wrapper, heuristic
  - PALS (Rasmussen *et al.*, 2006): find all local alignments over a given length with an error rate lower than a given threshold, exact algorithm
- Filter out long repeats (segmental duplications).

Self-alignment of the genomic sequences

- Align the whole genome with itself using stringent parameters:
  - BLASTER (Quesneville *et al.*, 2003): BLAST wrapper, heuristic
  - PALS (Rasmussen *et al.*, 2006): find all local alignments over a given length with an error rate lower than a given threshold, exact algorithm
- Filter out long repeats (segmental duplications).

Genome	Program	Number of matches	Genome coverage
D. mel.	BLASTER	109,882	7.41%
	PALS	105,059	7.38%

Self-alignment of the genomic sequences

- Align the whole genome with itself using stringent parameters:
  - BLASTER (Quesneville *et al.*, 2003): BLAST wrapper, heuristic
  - PALS (Rasmussen *et al.*, 2006): find all local alignments over a given length with an error rate lower than a given threshold, exact algorithm
- Filter out long repeats (segmental duplications).

Genome	Program	Number of matches	Genome coverage
D. mel.	BLASTER	109,882	7.41%
	PALS	105,059	7.38%
A. tha.	BLASTER	103,728	13.48%
	PALS	51,023	10.53%

Clustering of the all-by-all matches

- Launch clustering programs specific of interspersed repeats:
  - GROUPER (Quesneville *et al.*, 2003): chain the matches, then single-link clustering (95% coverage)
  - RECON (Bao and Eddy, 2002): single-link clustering (50% coverage) and handle segmental duplications, then single-link clustering (90% coverage) and handle related TE families
  - PILER (Edgar and Myers, 2005): build lists of matches covering a maximal contiguous region, then globally align these lists (95% coverage)

Clustering of the all-by-all matches

- Launch clustering programs specific of interspersed repeats:
  - GROUPER (Quesneville *et al.*, 2003): chain the matches, then single-link clustering (95% coverage)
  - RECON (Bao and Eddy, 2002): single-link clustering (50% coverage) and handle segmental duplications, then single-link clustering (90% coverage) and handle related TE families
  - PILER (Edgar and Myers, 2005): build lists of matches covering a maximal contiguous region, then globally align these lists (95% coverage)
- Filter out clusters with less than 3 members.

Clustering of the all-by-all matches

- Launch clustering programs specific of interspersed repeats:
  - GROUPER (Quesneville *et al.*, 2003): chain the matches, then single-link clustering (95% coverage)
  - RECON (Bao and Eddy, 2002): single-link clustering (50% coverage) and handle segmental duplications, then single-link clustering (90% coverage) and handle related TE families
  - PILER (Edgar and Myers, 2005): build lists of matches covering a maximal contiguous region, then globally align these lists (95% coverage)
- Filter out clusters with less than 3 members.

Genome	all-by-all	GROUPER	RECON	PILER	Reference consensus
D. mel.	BLASTER	730	451	120	117
A. tha.	BLASTER	1428	1021	300	305

Multiple sequence alignment for each cluster

- Launch progressive multiple alignment programs:
  - MAP (Huang, 1994): no penalty for gaps beyond a given length
  - CLUSTAL-W (Thompson *et al.*, 1994): first to propose position-specific gap penalties
  - MAFFT (Katoh and Toh, 2008): FFT to quickly detect homologous segments, and normalized similarity matrix
  - PRANK (Loytynoja and Goldman, 2005): take indels into account

Multiple sequence alignment for each cluster

- Launch progressive multiple alignment programs:
  - MAP (Huang, 1994): no penalty for gaps beyond a given length
  - CLUSTAL-W (Thompson *et al.*, 1994): first to propose position-specific gap penalties
  - MAFFT (Katoh and Toh, 2008): FFT to quickly detect homologous segments, and normalized similarity matrix
  - PRANK (Loytynoja and Goldman, 2005): take indels into account
- Build a consensus from each multiple alignment.

Multiple sequence alignment for each cluster

- Launch progressive multiple alignment programs:
  - MAP (Huang, 1994): no penalty for gaps beyond a given length
  - CLUSTAL-W (Thompson *et al.*, 1994): first to propose position-specific gap penalties
  - MAFFT (Katoh and Toh, 2008): FFT to quickly detect homologous segments, and normalized similarity matrix
  - PRANK (Loytynoja and Goldman, 2005): take indels into account
- Build a consensus from each multiple alignment.

Genome	all-by-all	clustering	MSA	R_{CC}
D. mel.	BLASTER	RECON	MAP	66.2%
			CLUSTAL-W	20.6%
			MAFFT	54.4%
			PRANK	61.8%

Results and comparison with RepeatScout

Genome	Program	Consensus	$S_n^{'}$	$S_p^{'}$	R_{CC}
D. mel.	GROUPER	730	80%	86%	66%
	RECON	451	92%	73%	66%
	PILER	120	62%	84%	51%

Results and comparison with RepeatScout

Genome	Program	Consensus	$S_n^{'}$	$S_p^{'}$	R_{CC}
D. mel.	GROUPER	730	80%	86%	66%
	RECON	451	92%	73%	66%
	PILER	120	62%	84%	51%
A. tha.	GROUPER	1428	60%	82%	39%
	RECON	1021	74%	62%	43%
	PILER	300	47%	57%	32%

Results and comparison with RepeatScout

Genome	Program	Consensus	S_n'	S_p'	R_{CC}
D. mel.	GROUPER	730	80%	86%	66%
	RECON	451	92%	73%	66%
	PILER	120	62%	84%	51%
A. tha.	GROUPER	1428	60%	82%	39%
	RECON	1021	74%	62%	43%
	PILER	300	47%	57%	32%

GROUPER and RECON have the best overall results.
Moreover they appear to be complementary.

Results and comparison with RepeatScout

Genome	Program	Consensus	$S_n^{'}$	$S_p^{'}$	R_{CC}
D. mel.	GROUPER	730	80%	86%	66%
	RECON	451	92%	73%	66%
	PILER	120	62%	84%	51%
A. tha.	GROUPER	1428	60%	82%	39%
	RECON	1021	74%	62%	43%
	PILER	300	47%	57%	32%

GROUPER and RECON have the best overall results.
Moreover they appear to be complementary.

Genome	Program	Consensus	$S_n^{'}$	$S_p^{'}$	R_{CC}
D. mel.	RepeatScout	1770	95%	58%	25%
A. tha.	RepeatScout	3417	83%	40%	13%

Combined approach in the TEdenovo pipeline

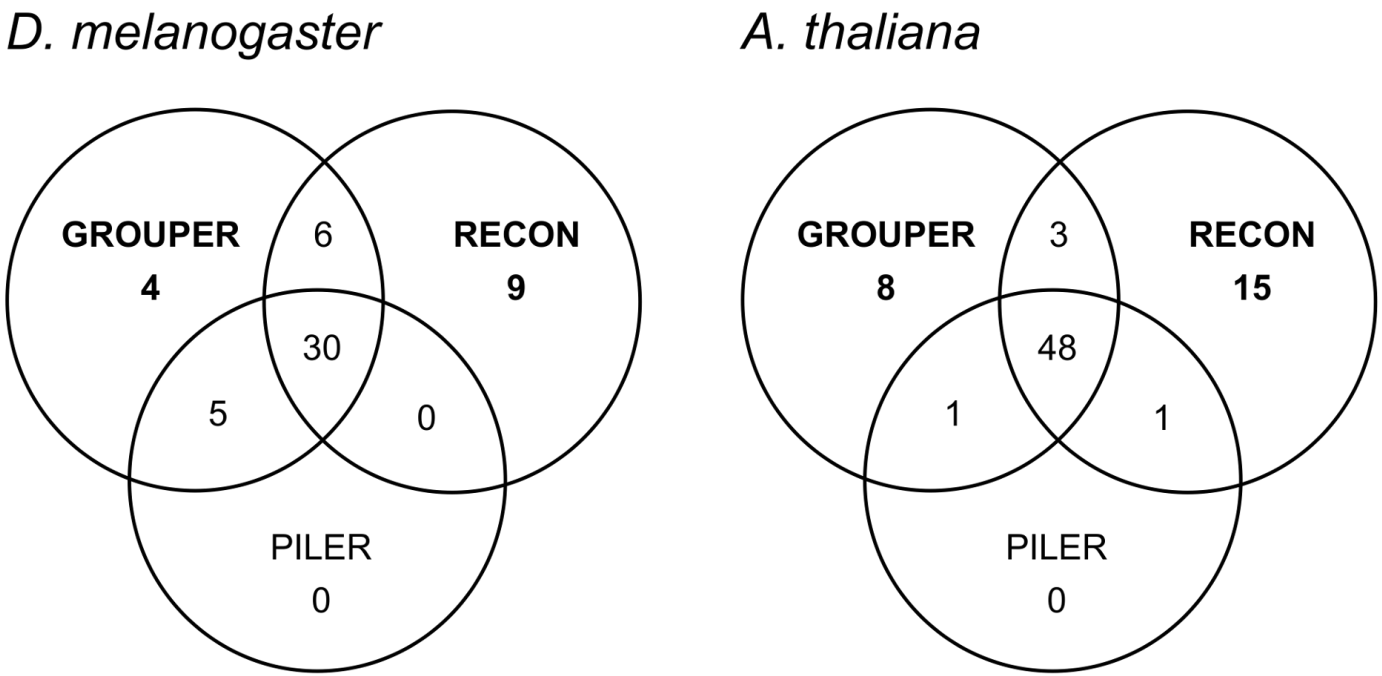


Figure: Number of TE reference sequences fully recovered by a *de novo* consensus.

Combined approach in the TEdenovo pipeline

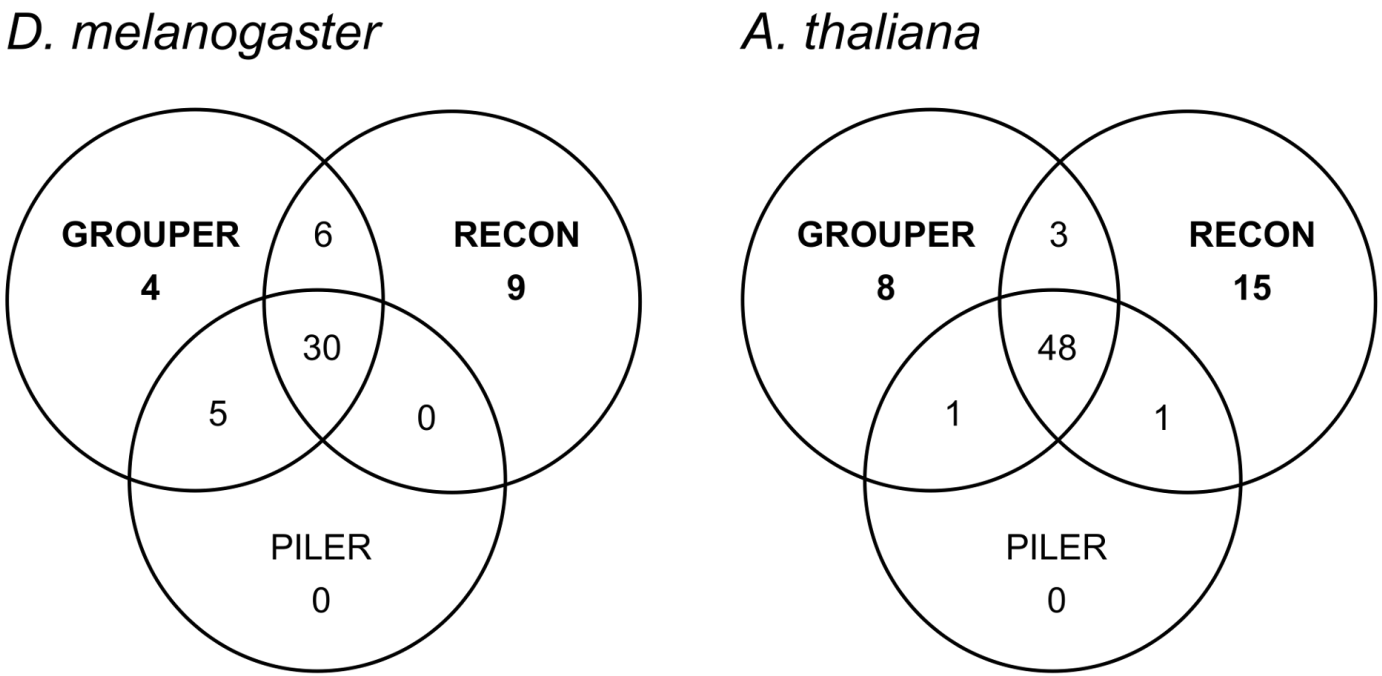


Figure: Number of TE reference sequences fully recovered by a *de novo* consensus.

Combining several clustering programs enables to recover more full-length TE reference sequences.

Classification of consensus sequences

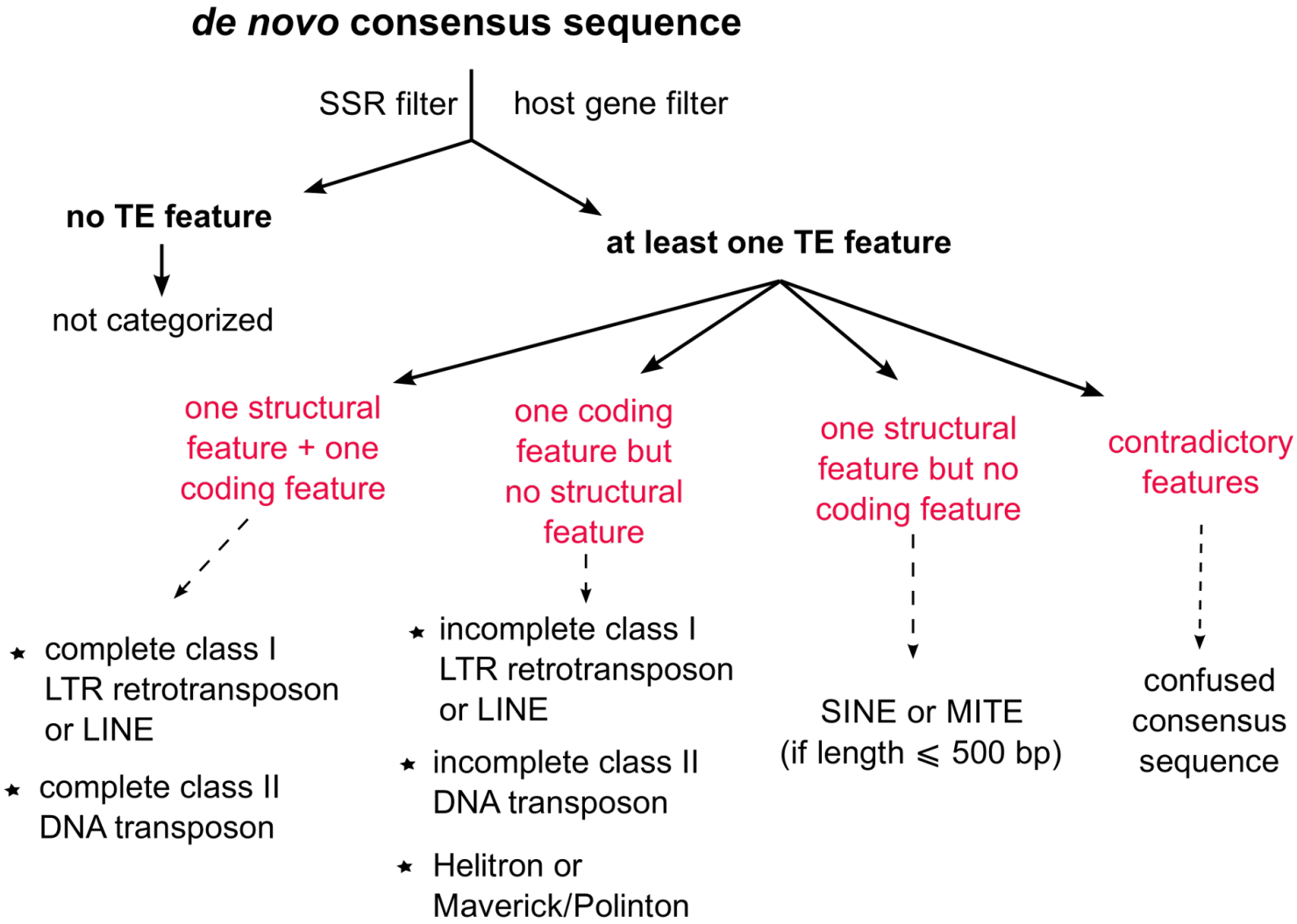


Figure: Simplified decision tree implemented in the TE classifier

Redundancy removal and validation

For instance, an "incomplete" LTR retrotransposon is removed when included over 98% of its length into a "complete" LTR retrotransposon with an identity over 95%.

Redundancy removal and validation

For instance, an "incomplete" LTR retrotransposon is removed when included over 98% of its length into a "complete" LTR retrotransposon with an identity over 95%.

Genome	Redundancy removal	Consensus	S_n'	S_p'	R_{CC}
D. mel.	no	1301	93%	81%	79%
	95-98	593	92%	75%	78%
A. tha.	no	2749	74%	72%	49%
	95-98	1275	74%	67%	49%

Redundancy removal and validation

For instance, an "incomplete" LTR retrotransposon is removed when included over 98% of its length into a "complete" LTR retrotransposon with an identity over 95%.

Genome	Redundancy removal	Consensus	S_n'	S_p'	R_{CC}
D. mel.	no	1301	93%	81%	79%
	95-98	593	92%	75%	78%
A. tha.	no	2749	74%	72%	49%
	95-98	1275	74%	67%	49%

At the end of the TEdenovo pipeline, a databank of TE *de novo* consensus sequences is generated, of high quality and low redundancy.

TE copy annotation with TEannot

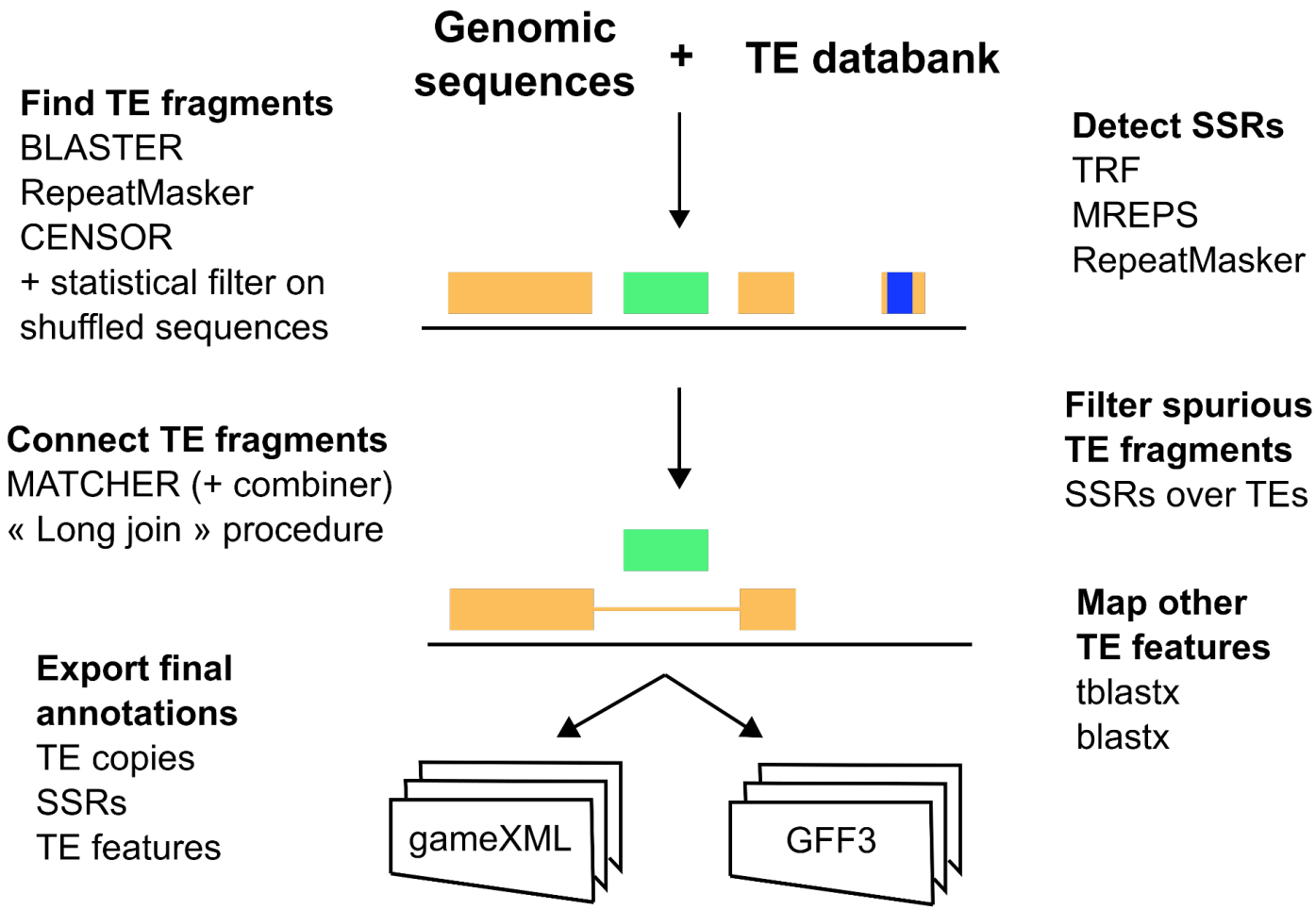
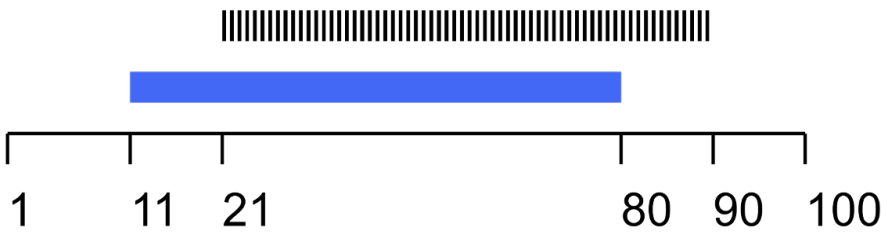


Figure: Flow chart of the TEannot pipeline (Quesneville *et al.*, 2005) on which I improved several steps to better handle the databanks of *de novo* consensus.

Validation of TE annotation with *de novo* consensus



Sensitivity = TP / (TP + FN)
Sn = 60 / (60 + 10) = 86%

Specificity = TN / (TN + FP)
Sp = 20 / (20 + 10) = 67%

Validation of TE annotation with *de novo* consensus

Sensitivity = TP / (TP + FN)
Sn = 60 / (60 + 10) = 86%

Specificity = TN / (TN + FP)
Sp = 20 / (20 + 10) = 67%

Genome	Bank	Consensus	Coverage	S_n	S_p
D. mel.	BDGP	125	11%	na	na
	TEdenovo	568	12%	91%	97%
A. tha.	Repbase	318	19%	na	na
	TEdenovo	1232	23%	87%	92%

Validation of TE annotation with *de novo* consensus

Sensitivity = TP / (TP + FN)
Sn = 60 / (60 + 10) = 86%

Specificity = TN / (TN + FP)
Sp = 20 / (20 + 10) = 67%

Genome	Bank	Consensus	Coverage	S_n	S_p
D. mel.	BDGP	125	11%	na	na
	TEdenovo	568	12%	91%	97%
A. tha.	Repbase	318	19%	na	na
	TEdenovo	1232	23%	87%	92%

Using TEdenovo and TEannot enables to automatically annotate the TE content of newly sequenced genomes with good results.

Outline

Extensive structural variation within TE families

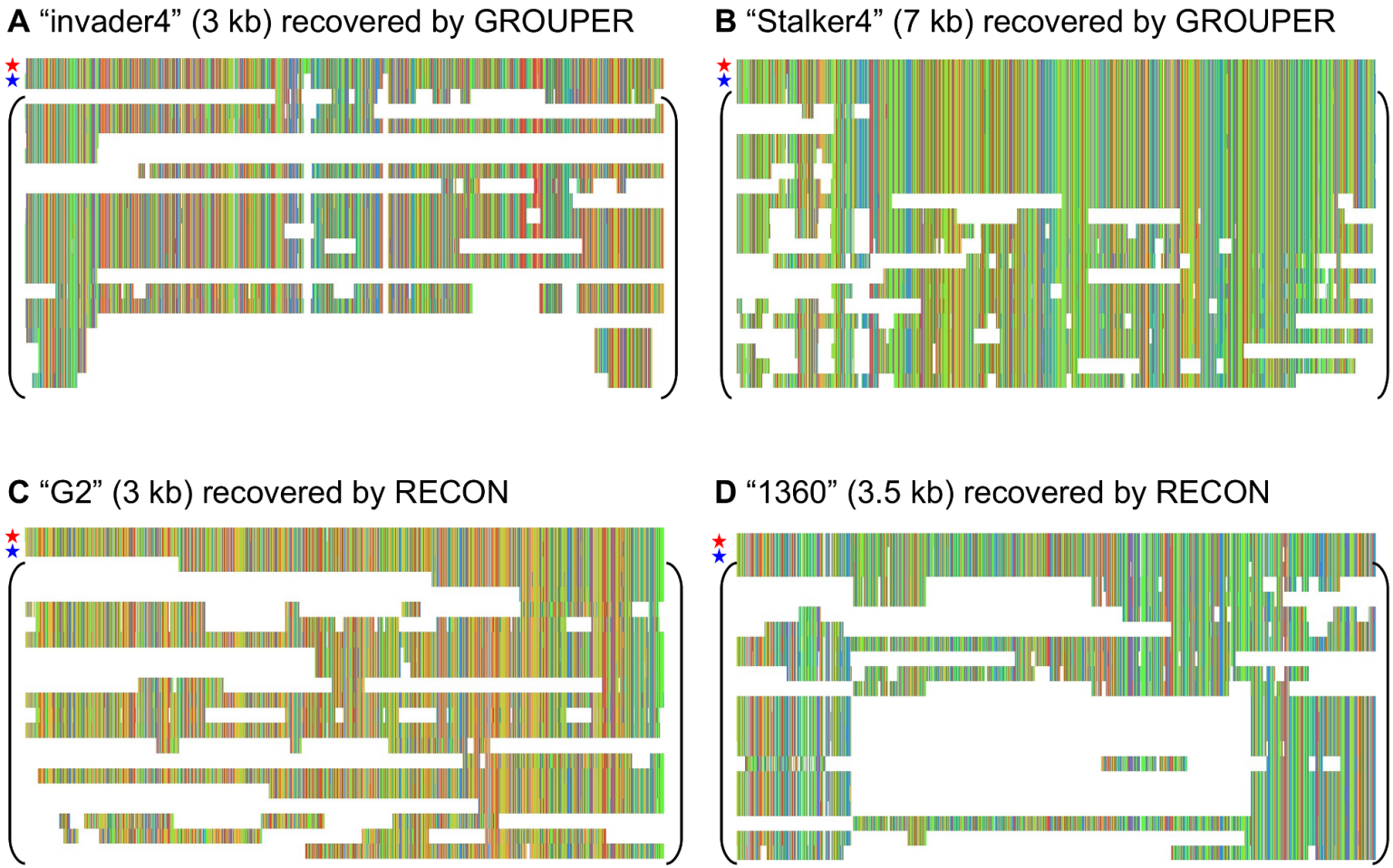


Figure: TE families for which only one clustering method fully recovered the reference sequence (red star: the reference sequence; blue star: the *de novo* consensus; in brackets: the genomic copies; one color per nucleotide).

Interpreting the diversification of a TE family?

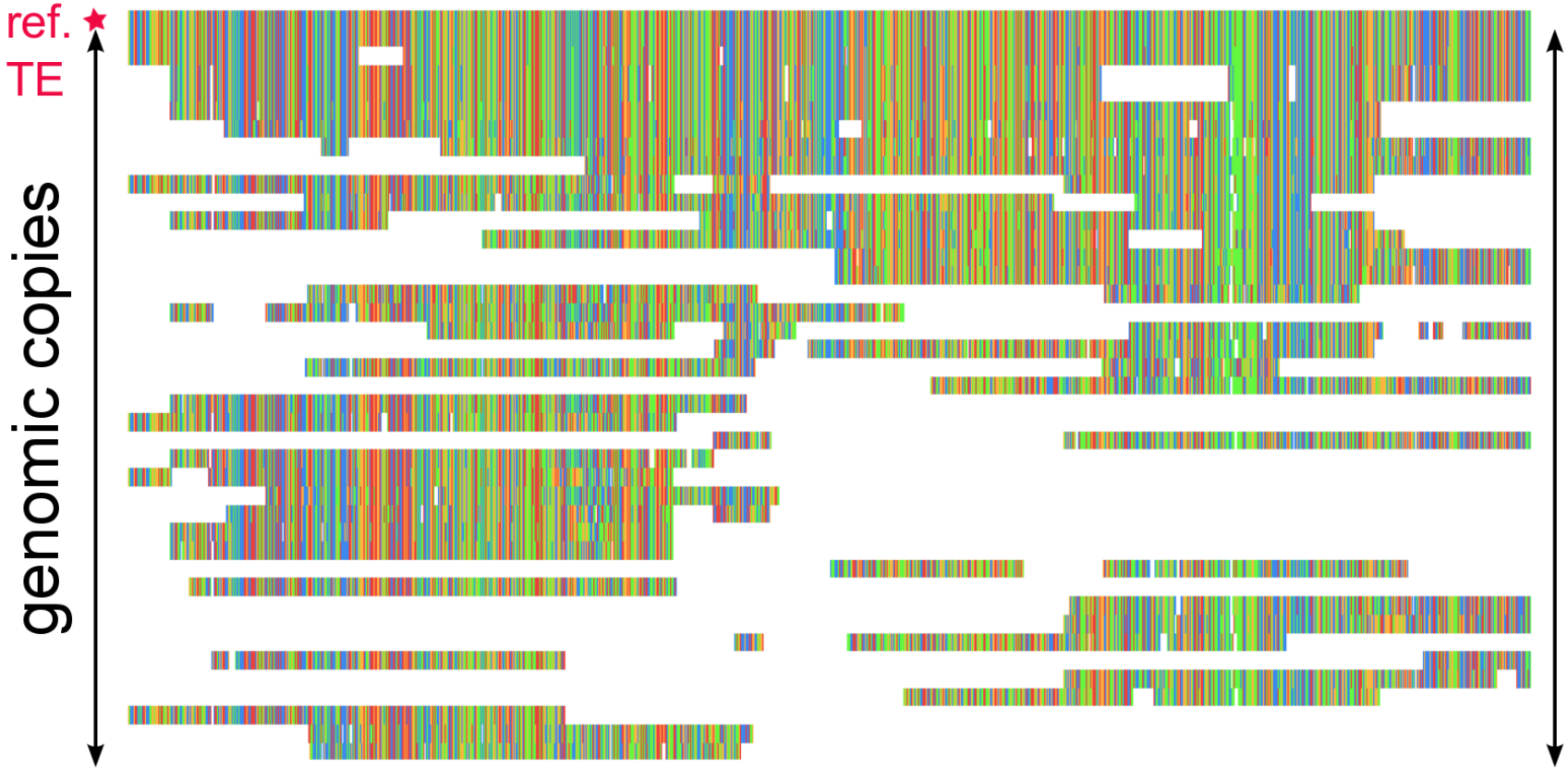


Figure: Multiple alignment of the "Athila" reference sequence (red star) with some of its longest genomic copies (black arrows).

Recovering several *de novo* consensus per TE family

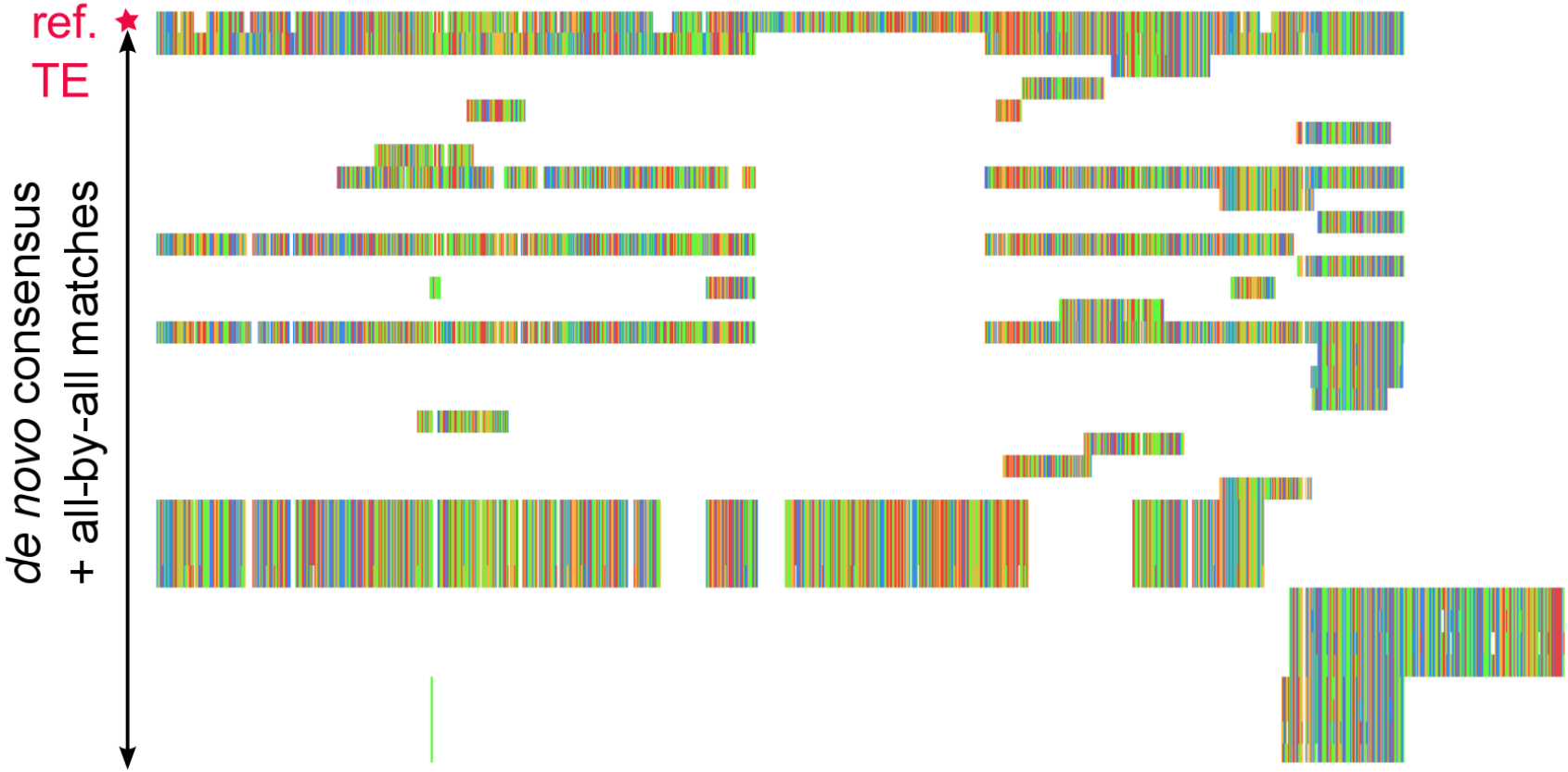


Figure: Multiple alignment of the "Athila" reference sequence (red star) and its 4 *de novo* consensus sequences, to which are added the all-by-all matches from which they were derived.

Recovering several *de novo* consensus per TE family

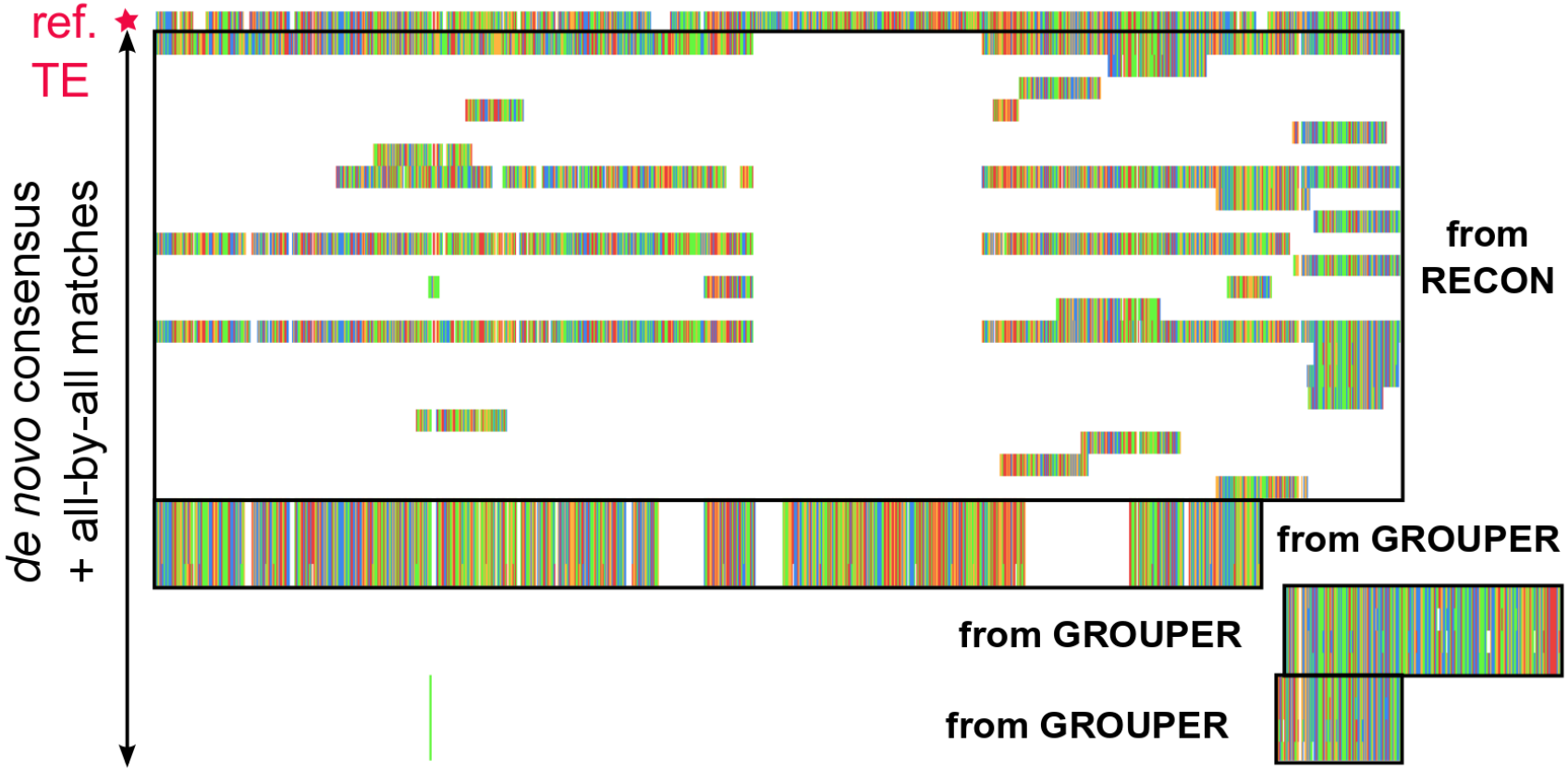
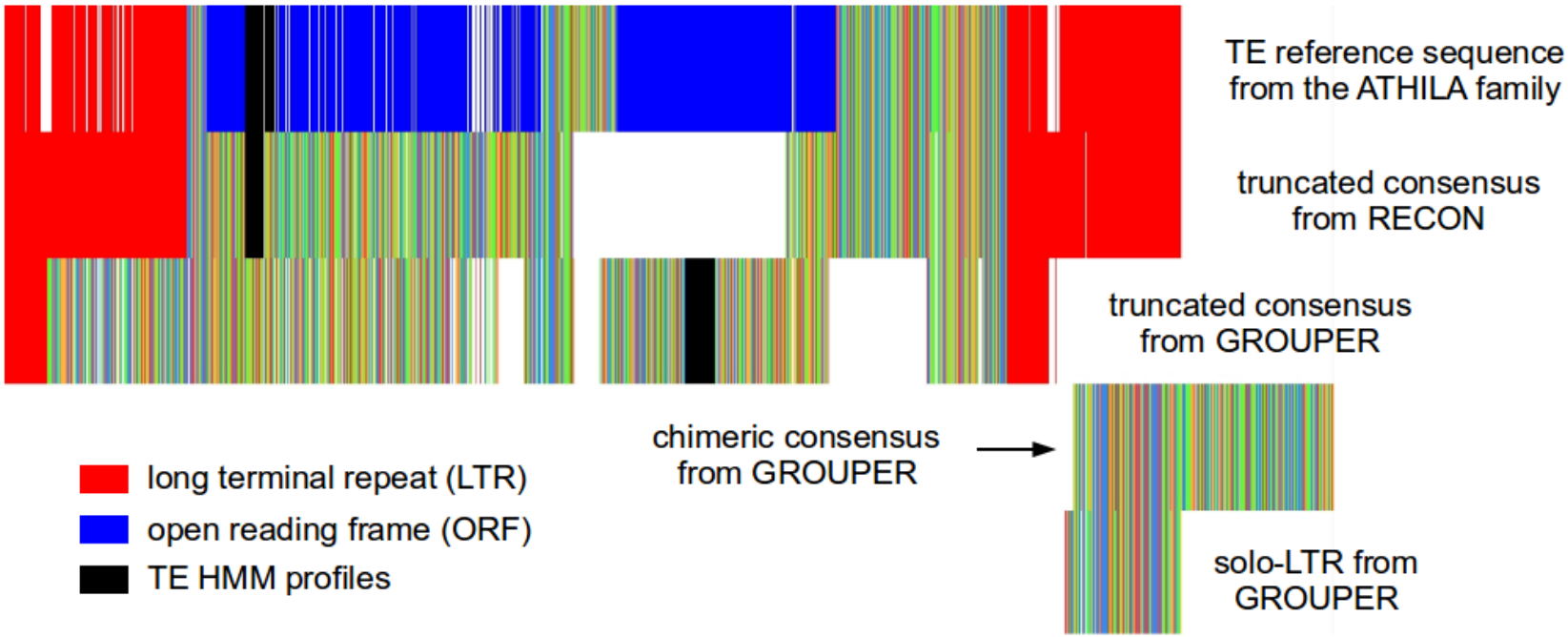


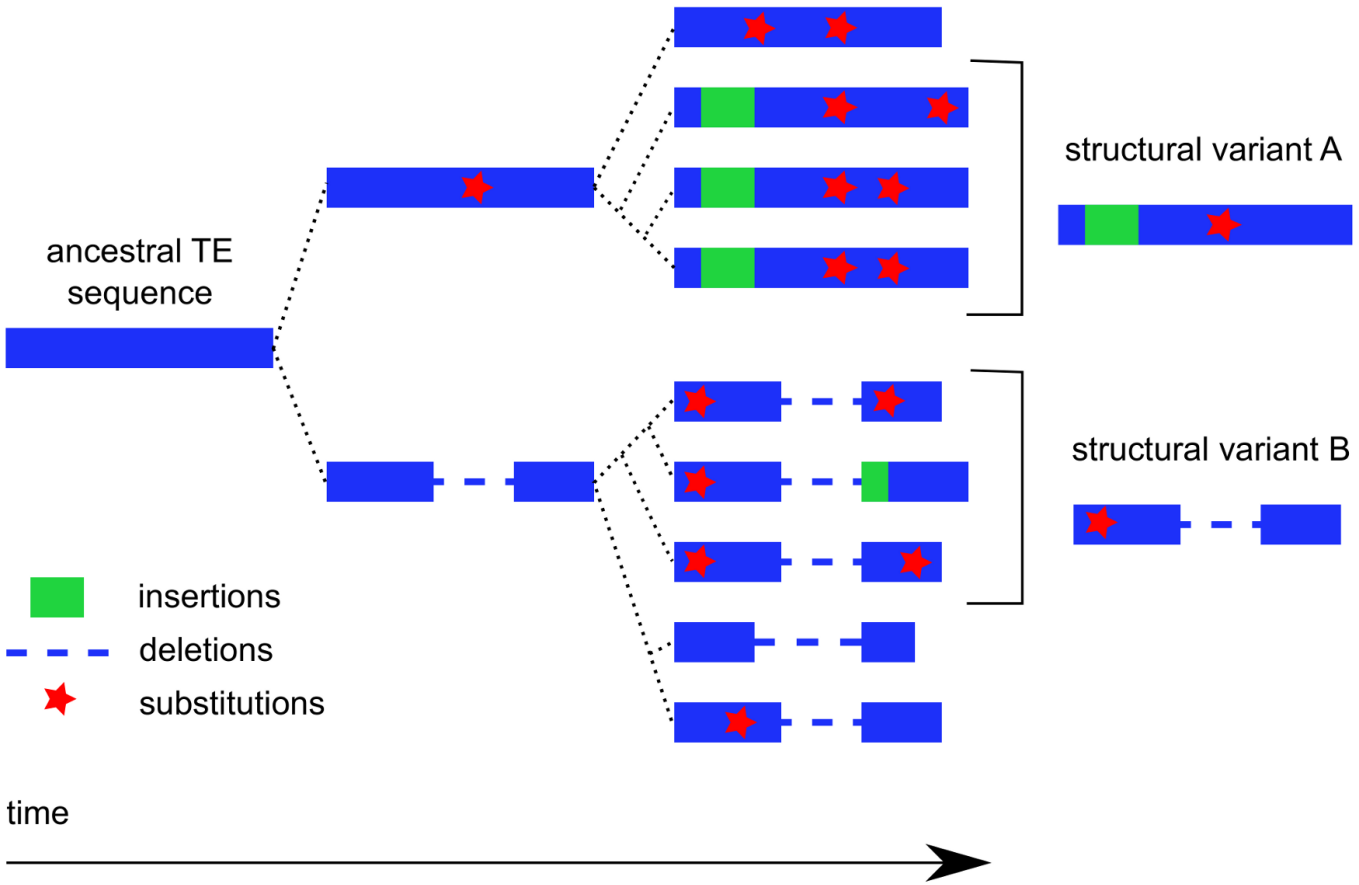
Figure: Multiple alignment of the "Athila" reference sequence (red star) and its 4 *de novo* consensus sequences, to which are added the all-by-all matches from which they were derived.

Advantage of this *de novo* approach

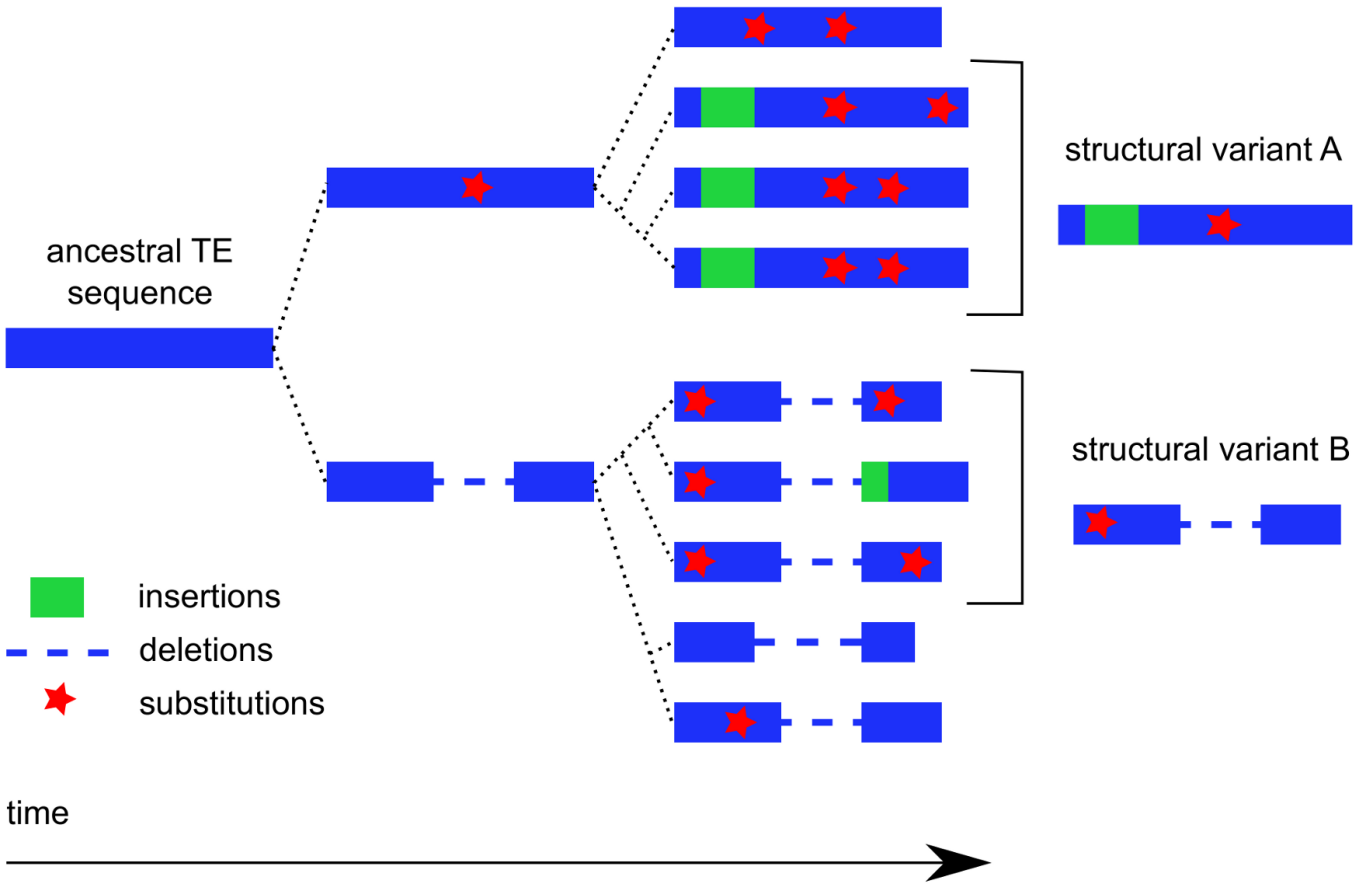


Enabling the interpretation of TE diversification, and building hypotheses concerning the mechanisms behind the emergence of structural variants.

Taking TE structural variations into account



Taking TE structural variations into account



This work, from the comparative analysis to the results on TE diversification, was recently submitted (currently under review).

Outline

From TE annotation to biological questions

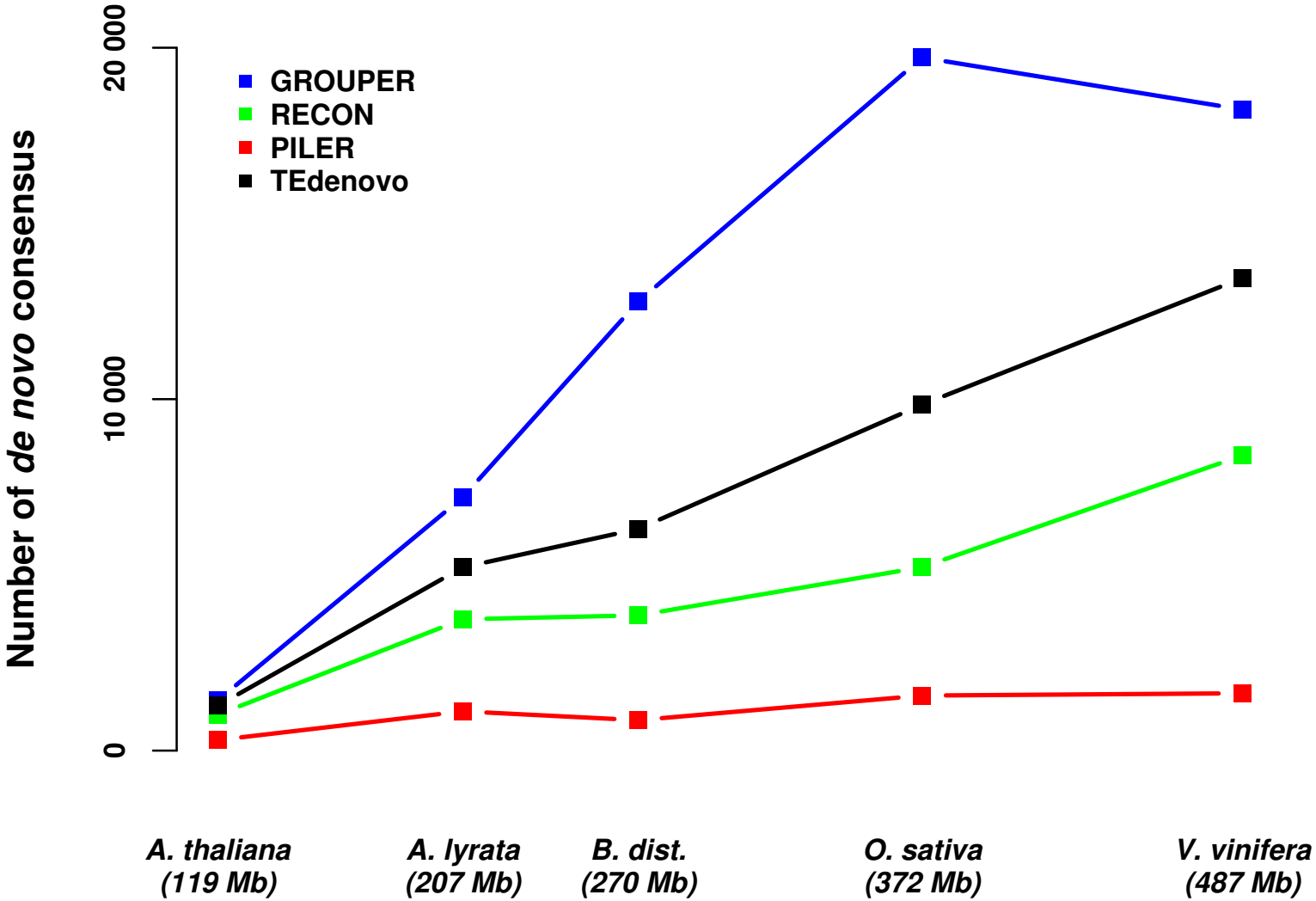
- TE annotation of the *M. incognita* genome
  - co-author, published in Nature Biotechnology (2008)
  - asexual plant-parasitic nematode
  - small genome (86 Mb) but high TE content (35%)
  - involved in horizontal transfers of plant cell-wall degrading genes?

From TE annotation to biological questions

- TE annotation of the *M. incognita* genome
  - co-author, published in Nature Biotechnology (2008)
  - asexual plant-parasitic nematode
  - small genome (86 Mb) but high TE content (35%)
  - involved in horizontal transfers of plant cell-wall degrading genes?

- Correlation between TEs and synteny breaks in holocentric genomes
  - co-author with E. Permal, published in PNAS (2010)
  - TE annotation in 15 syntenic BACs of Lepidopterans
  - correlation with synteny breaks which occurrence rate is 4x higher than in Drosophilidae, itself 2x higher than in mammals

Differential TE diversification in plant genomes



(Analysis on *V. vinifera*: with N. Choisne; for GROUPER, only clusters with 4 members were kept)

Projects benefiting from the pipelines

- TriAnnot, the web interface to annotate Triticeae BACs, with P. Leroy, F. Giacomoni and A. Bernard from GDEC (is underway);

Projects benefiting from the pipelines

- TriAnnot, the web interface to annotate Triticeae BACs, with P. Leroy, F. Giacomoni and A. Bernard from GDEC (is underway);
- TE annotation of *Acyrthosiphon pisum* with E. Permal from URGI (published in PLoS Biology, 2010);

Projects benefiting from the pipelines

- TriAnnot, the web interface to annotate Triticeae BACs, with P. Leroy, F. Giacomoni and A. Bernard from GDEC (is underway);
- TE annotation of *Acyrthosiphon pisum* with E. Permal from URGI (published in PLoS Biology, 2010);
- TE annotation of *Ectocarpus siliculosus* with C. Pommier from URGI (published in Nature, 2010);

Projects benefiting from the pipelines

- TriAnnot, the web interface to annotate Triticeae BACs, with P. Leroy, F. Giacomoni and A. Bernard from GDEC (is underway);
- TE annotation of *Acyrthosiphon pisum* with E. Permal from URGI (published in PLoS Biology, 2010);
- TE annotation of *Ectocarpus siliculosus* with C. Pommier from URGI (published in Nature, 2010);
- Comparative analysis of TEs in fungi genomes and impact of RIP mutations on TE copies with J. Amselem from URGI-BIOGER (is underway);

Projects benefiting from the pipelines

- TriAnnot, the web interface to annotate Triticeae BACs, with P. Leroy, F. Giacomoni and A. Bernard from GDEC (is underway);
- TE annotation of *Acyrthosiphon pisum* with E. Permal from URGI (published in PLoS Biology, 2010);
- TE annotation of *Ectocarpus siliculosus* with C. Pommier from URGI (published in Nature, 2010);
- Comparative analysis of TEs in fungi genomes and impact of RIP mutations on TE copies with J. Amselem from URGI-BIOGER (is underway);
- Population genetics of TEs in Drosophilidae genomes, with A-S. Fiston-Lavier and D. Petrov from Stanford (is underway).

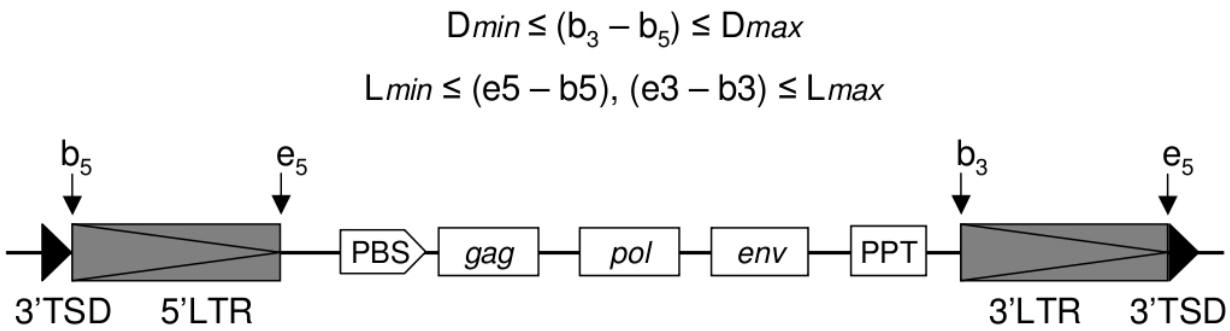
Outline

Preliminary analysis for future improvements

- Unification of TE clustering algorithms:
  - new algorithm combining the advantages of GROUPER and RECON
  - parallelize the clustering of the all-by-all alignments

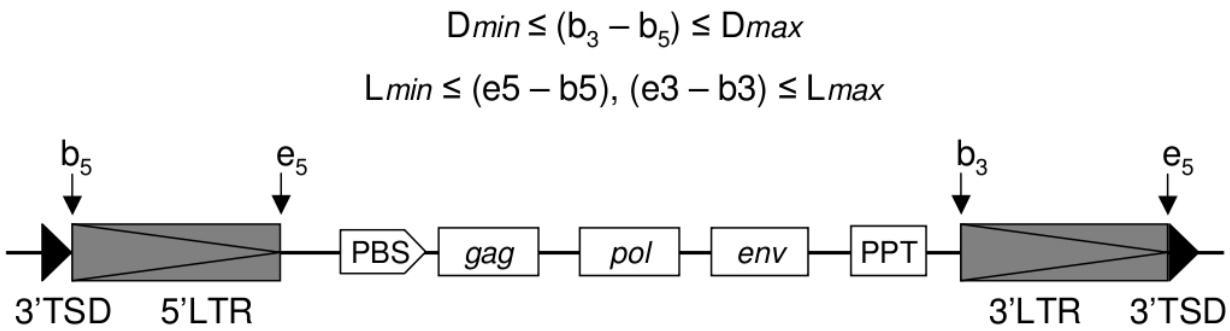
Preliminary analysis for future improvements

- Unification of TE clustering algorithms:
  - new algorithm combining the advantages of GROUPER and RECON
  - parallelize the clustering of the all-by-all alignments
- Integration of structure-based programs:
  - e.g. LTRharvest from Ellinghaus *et al*. (2010)



Preliminary analysis for future improvements

- Unification of TE clustering algorithms:
  - new algorithm combining the advantages of GROUPER and RECON
  - parallelize the clustering of the all-by-all alignments
- Integration of structure-based programs:
  - e.g. LTRharvest from Ellinghaus *et al*. (2010)



- Strategies to annotate large genomes:
  - "light" TEannot and "iterative splicing"
  - TE detection quality on genomes sequenced with NGS

Limits of the current *de novo* methodology

Main obstacle: in computational biology, how to test a *de novo* approach when no experiment can directly confirm or invalidate the predictions?

Limits of the current *de novo* methodology

Main obstacle: in computational biology, how to test a *de novo* approach when no experiment can directly confirm or invalidate the predictions?

- recover and distinguish structural variants from TE families with complex evolutionary trajectories?

Limits of the current *de novo* methodology

Main obstacle: in computational biology, how to test a *de novo* approach when no experiment can directly confirm or invalidate the predictions?

- recover and distinguish structural variants from TE families with complex evolutionary trajectories?
- integrate conflicting predictions from all-by-all and structure-based approaches?

Limits of the current *de novo* methodology

Main obstacle: in computational biology, how to test a *de novo* approach when no experiment can directly confirm or invalidate the predictions?

- recover and distinguish structural variants from TE families with complex evolutionary trajectories?
- integrate conflicting predictions from all-by-all and structure-based approaches?
- precisely quantify the TE detection quality as a function of sequencing coverage?

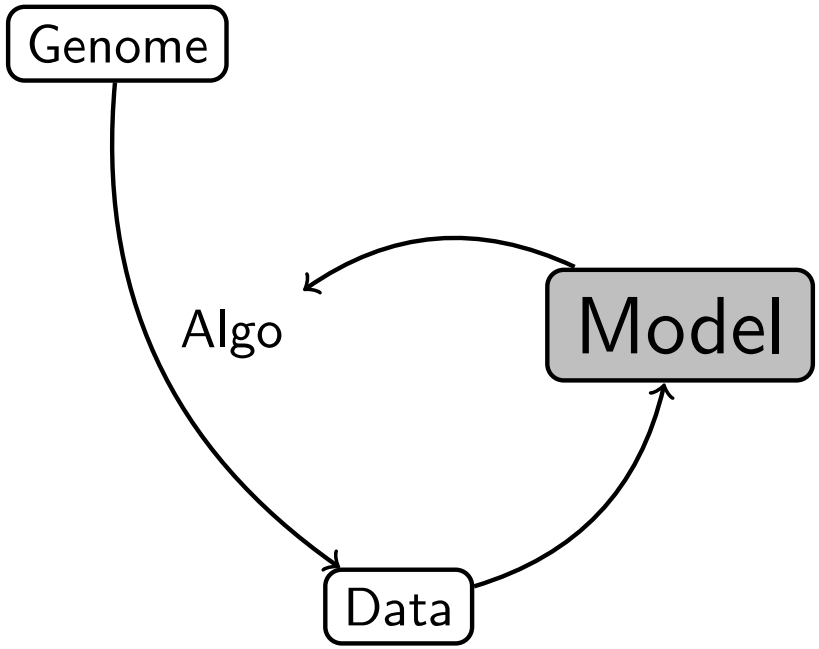
Limits of the current *de novo* methodology

Main obstacle: in computational biology, how to test a *de novo* approach when no experiment can directly confirm or invalidate the predictions?

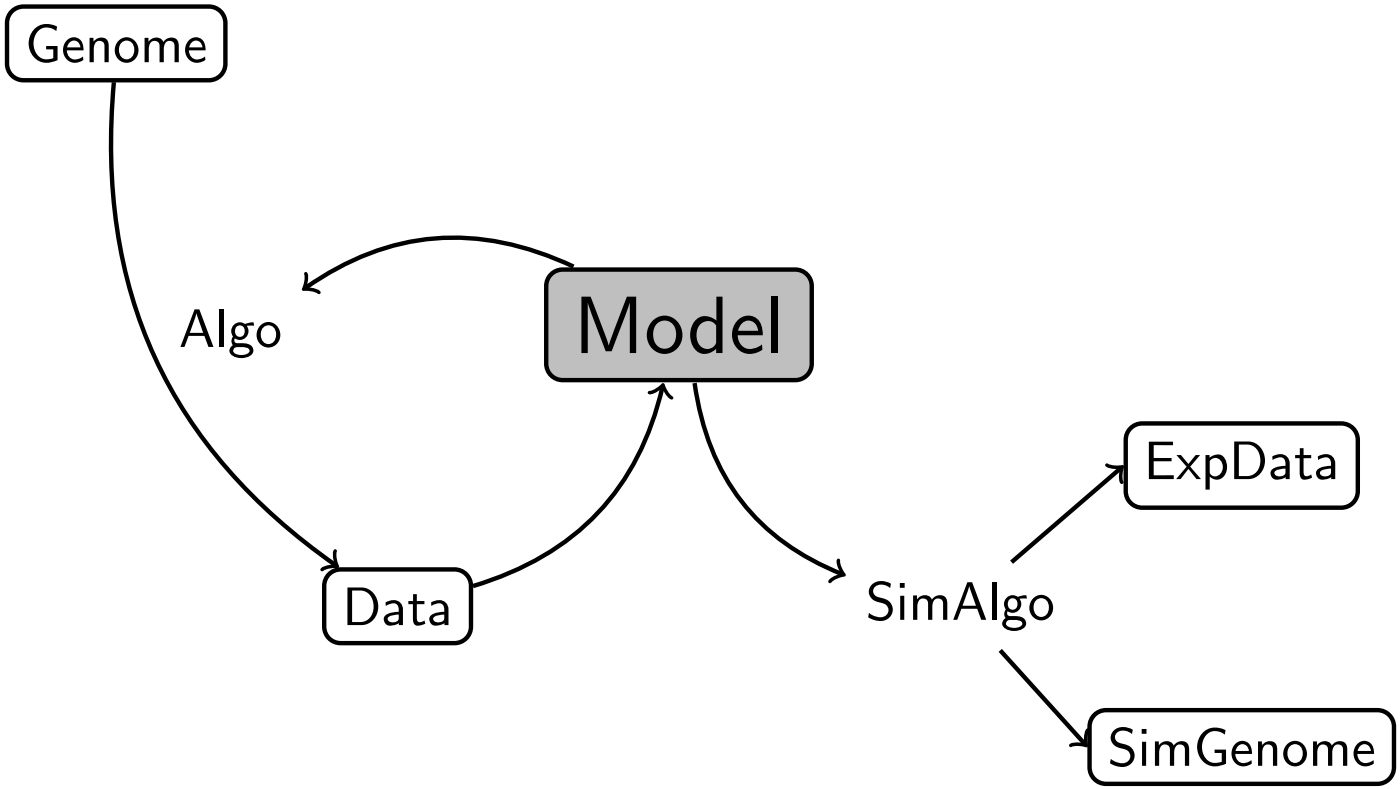
- recover and distinguish structural variants from TE families with complex evolutionary trajectories?
- integrate conflicting predictions from all-by-all and structure-based approaches?
- precisely quantify the TE detection quality as a function of sequencing coverage?

Need for pragmatic simulation, not as an end, but as a means.

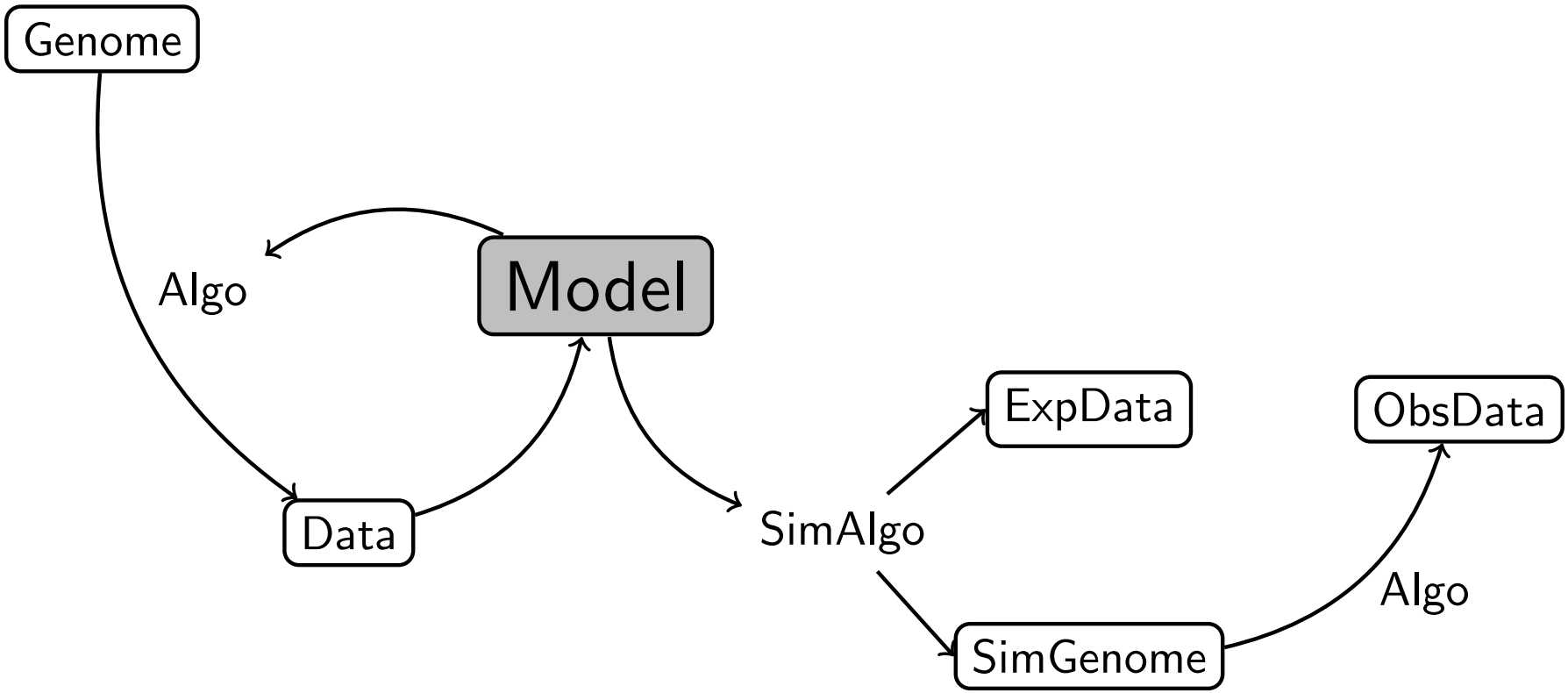
Interplay between detection and simulation algorithms



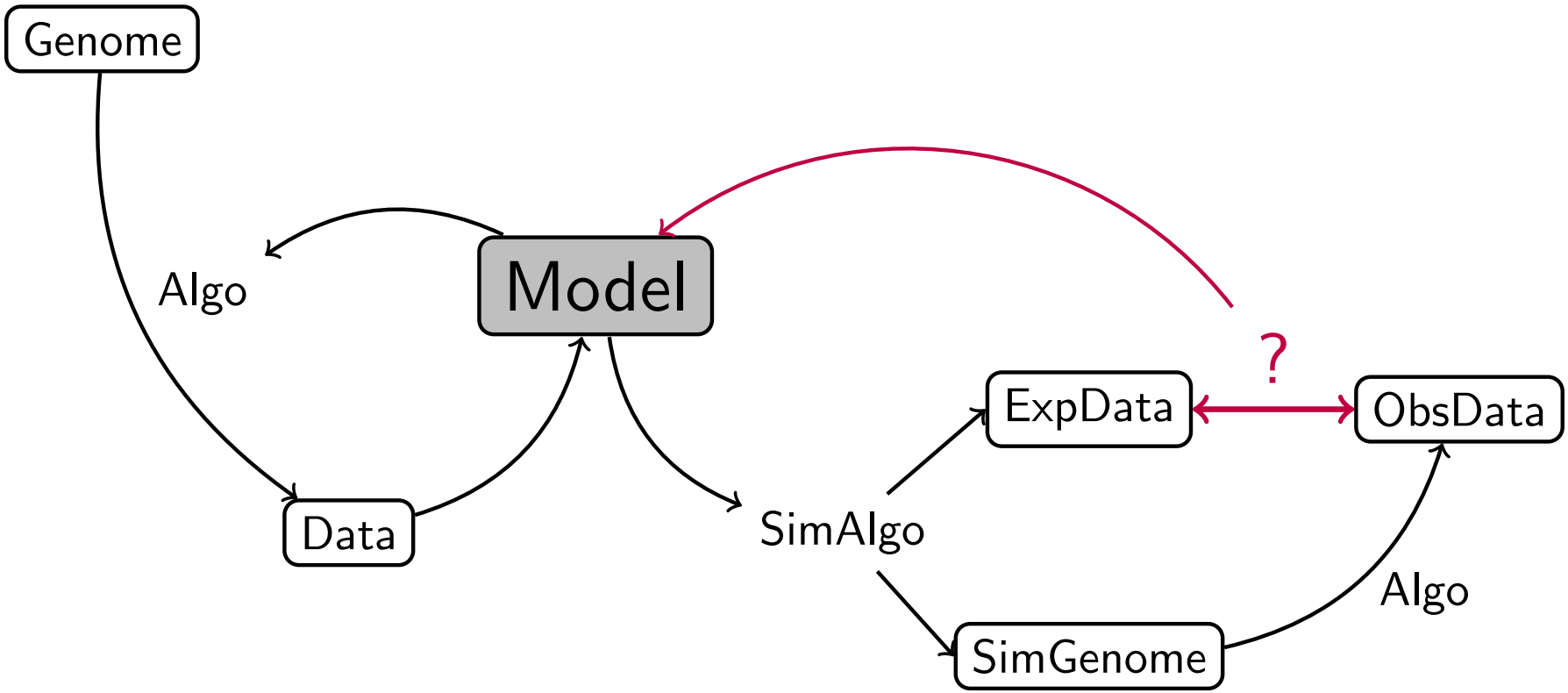
Interplay between detection and simulation algorithms



Interplay between detection and simulation algorithms



Interplay between detection and simulation algorithms



Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample
2. launch the TE detection algorithms on this sample

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample
2. launch the TE detection algorithms on this sample
3. compare the predicted results with the expectations

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample
2. launch the TE detection algorithms on this sample
3. compare the predicted results with the expectations
4. improve the detection algorithms or the evolutionary scenarios

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample
2. launch the TE detection algorithms on this sample
3. compare the predicted results with the expectations
4. improve the detection algorithms or the evolutionary scenarios
5. iterate steps 1 to 4 ($\approx$ one month)

Generic workflow

It is not about simulating genome evolution, rather generating an evolved genome. What does it mean in practice?

1. choose an evolutionary scenario and generate a sample
2. launch the TE detection algorithms on this sample
3. compare the predicted results with the expectations
4. improve the detection algorithms or the evolutionary scenarios
5. iterate steps 1 to 4 ($\approx$ one month)

Designing an algorithm to detect TE structural variants and identify their family is not only an improvement in terms of TE annotation, but more importantly, it is a first step towards a better understanding of TE biology.

Contribution of this PhD thesis

- Using model genomes, I compared programs to detect TEs *via* a *de novo* approach; this allowing me to explicit the conceptual model from which these programs were built.

Contribution of this PhD thesis

- Using model genomes, I compared programs to detect TEs *via* a *de novo* approach; this allowing me to explicit the conceptual model from which these programs were built.
- From this analysis, I showed that the *de novo* approach represents a typical TE family by several sequences, and interpret this as resulting from the pervasive presence of TE structural variants.

Contribution of this PhD thesis

- Using model genomes, I compared programs to detect TEs *via* a *de novo* approach; this allowing me to explicit the conceptual model from which these programs were built.
- From this analysis, I showed that the *de novo* approach represents a typical TE family by several sequences, and interpret this as resulting from the pervasive presence of TE structural variants.
- From results on model genomes, I extended the model of TE dynamics and implemented an *ad hoc* combined approach that take these variants into account for TE detection.

Contribution of this PhD thesis

- Using model genomes, I compared programs to detect TEs *via* a *de novo* approach; this allowing me to explicit the conceptual model from which these programs were built.
- From this analysis, I showed that the *de novo* approach represents a typical TE family by several sequences, and interpret this as resulting from the pervasive presence of TE structural variants.
- From results on model genomes, I extended the model of TE dynamics and implemented an *ad hoc* combined approach that take these variants into account for TE detection.
- The tools I developed have been applied since on various, newly sequenced genomes, confirming the importance of TE structural variants.

Contribution of this PhD thesis

- Using model genomes, I compared programs to detect TEs *via* a *de novo* approach; this allowing me to explicit the conceptual model from which these programs were built.
- From this analysis, I showed that the *de novo* approach represents a typical TE family by several sequences, and interpret this as resulting from the pervasive presence of TE structural variants.
- From results on model genomes, I extended the model of TE dynamics and implemented an *ad hoc* combined approach that take these variants into account for TE detection.
- The tools I developed have been applied since on various, newly sequenced genomes, confirming the importance of TE structural variants.
- Finally, I propose a methodology based on simulation to improve detection algorithms of the *de novo* approach and, through this, increase our understanding of TE diversification.

Outline

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet
- Defense jury: Khashayar Pakdaman, Olivier Panaud, Pierre Rouzé, Thomas Wicker

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet
- Defense jury: Khashayar Pakdaman, Olivier Panaud, Pierre Rouzé, Thomas Wicker
- Thesis advisory commitee: Sébastien Aubourg, Hugues Roest Crollius

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet
- Defense jury: Khashayar Pakdaman, Olivier Panaud, Pierre Rouzé, Thomas Wicker
- Thesis advisory commitee: Sébastien Aubourg, Hugues Roest Crollius
- URGI: Michael Alaux, Françoise Alfama, Joëlle Amselem, Sandie Arnoux, Nazneen Badroudine, Isabelle Blanc-Lenfle, Marc Bras, Baptiste Brault, Nathalie Choisne, Sandra Derozier, Sophie Durand, Benoît Hilselberger, Claire Hoede, Olivier Inizan, Véronique Jamilloux, Aminah Keliet, Erik Kimmel, Jonathan Kreplak, Nicolas Lapalu, Isabelle Luyten, Nacer Mohellibi, Valérie Moli-Rasolofo, Emmanuel Permal, Cyril Pommier, Sébastien Reboux, Delphine Steinbach, Dorothé Valdenaire, Daphné Verdelet, Matthias Zytnicki

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet
- Defense jury: Khashayar Pakdaman, Olivier Panaud, Pierre Rouzé, Thomas Wicker
- Thesis advisory commitee: Sébastien Aubourg, Hugues Roest Crollius
- URGI: Michael Alaux, Françoise Alfama, Joëlle Amselem, Sandie Arnoux, Nazneen Badroudine, Isabelle Blanc-Lenfle, Marc Bras, Baptiste Brault, Nathalie Choisne, Sandra Derozier, Sophie Durand, Benoît Hilselberger, Claire Hoede, Olivier Inizan, Véronique Jamilloux, Aminah Keliet, Erik Kimmel, Jonathan Kreplak, Nicolas Lapalu, Isabelle Luyten, Nacer Mohellibi, Valérie Moli-Rasolofo, Emmanuel Permal, Cyril Pommier, Sébastien Reboux, Delphine Steinbach, Dorothé Valdenaire, Daphné Verdelet, Matthias Zytnicki
- GDEC: Aurélien Bernard, Frédéric Choulet, Franck Giacomoni, Nicolas Guilhot, Philippe Leroy

Colleagues and collaborators

- PhD directors: Hadi Quesneville and Catherine Feuillet
- Defense jury: Khashayar Pakdaman, Olivier Panaud, Pierre Rouzé, Thomas Wicker
- Thesis advisory commitee: Sébastien Aubourg, Hugues Roest Crollius
- URGI: Michael Alaux, Françoise Alfama, Joëlle Amselem, Sandie Arnoux, Nazneen Badroudine, Isabelle Blanc-Lenfle, Marc Bras, Baptiste Brault, Nathalie Choisne, Sandra Derozier, Sophie Durand, Benoît Hilselberger, Claire Hoede, Olivier Inizan, Véronique Jamilloux, Aminah Keliet, Erik Kimmel, Jonathan Kreplak, Nicolas Lapalu, Isabelle Luyten, Nacer Mohellibi, Valérie Moli-Rasolofo, Emmanuel Permal, Cyril Pommier, Sébastien Reboux, Delphine Steinbach, Dorothé Valdenaire, Daphné Verdelet, Matthias Zytnicki
- GDEC: Aurélien Bernard, Frédéric Choulet, Franck Giacomoni, Nicolas Guilhot, Philippe Leroy
- IJM: Anna-Sophie Fiston-Lavier, Elodie Duprat

Institutions

- my research institute: INRA
- my graduate school: Frontières du Vivant
- my university: Université Paris Diderot

Thank you for your attention!