Traces of past transposable element presence in *Brassicaceae* genome dark matter

Hadi Quesneville

*Alphy, 7 February 2019*
TEs are key players of genome evolution

How important are transposons for plant evolution?

Damon Lisch
TE dynamics

Burst of transpositions

ancestral TE sequence

insertions
deletions
substitutions

accumulation of mutations and gradual removal

time

Hadi Quesneville
Scientific questions

TEs participate to the DNA turnover forming the raw material for genetic innovations.

→ How to recognize very old and degenerated repeated sequences up to 40 Myr old.

→ Search whether they played key evolutionary roles.
Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*.

**Florian Maumus** & **Hadi Quesneville**

**Cross-species TE annotation**

- **Arabidopsis**
- **Brassica**
- **Arabis**
- **Capsella**
- **Thellungiella**

**Identity distribution in 1% bins**

**Relative frequency**

- **At_Col**
- **At_other**
- **A. lyrata**
- **B. rapa**
- **C. rubella**
- **T. parvula**
- **T. halophila**
- **A. alpina**

**Cross-species TE annotation library**

- **At_Col**
- **At_other**
- **A. lyrata**
- **A. alpina**
- **B. rapa**
- **T. parvula**
- **T. halophila**
- **C. rubella**

**molecular clock**

- 25-40 mya
- 10-14 mya
- 3-6 mya

**Brassicaceae family**

- **Brassicaceae library**

**Cross-species TE annotation**

- **At_Col**
- **At_other**
- **A. lyrata**
- **A. alpina**
- **B. rapa**
- **T. parvula**
- **T. halophila**
- **C. rubella**

**36.8**

**3.7**

**25.8**

**6**

**0.5**

**Hadi Quesneville**
Deep Investigation of *Arabidopsis thaliana* Junk DNA Reveals a Continuum between Repetitive Elements and Genomic Dark Matter

Florian Maumus*, Hadi Quesneville*

UR1164 URGI - Research Unit in Genomics-Info, INRA, Versailles, France

---

**RepeatScout**

**RepeatModeler**

**TEedenovo**

**Repbase** (Buisine et al.)

---

**Bundle library**

**TEannot**

---

**Complete bundle annotation**

---

Fusion of two ancestral chromosomes

---

*URGI* SCIENCE & IMPACT
• Use TE copies instead of TE consensus for annotation

• Use annotation from many related species to take advantage of « cross-species annotation strategy »

→ Need efficient algorithm for massive comparisons
Duster algorithm

$k$-mer with regularly spaced holes

TE sequences (Model)  Hash indexe

Sorted list of tuple of (diag, seq, pos)

(-1, 3,50)
(1, 3,10)
(1, 3,20)
(1, 3,40)
(10, 3,10)
(15, 2,100)
(15, 2,110)
(30, 1,20)
(30, 1,30)
(30, 1,40)

Hadi Quesneville
Combining annotations

- **Arabidopsis**
  - 25-40 mya
  - 10-14 mya
  - 3-6 mya
  - **Arabidopsis thaliana**
  - **Arabidopsis lyrata**
- **Arabis**
- **Brassica**
- **Eutrema**
- **Capsella rubella**
- **Arabis alpina**
- **Brassica rapa**
- **Thellungiella parvula**

**Brassicaceae library**

**Duster**

**TAIR10 TEs**

**Brassicaceae**

- 52.33
- 12.45
- 0.35
- 40.44
- 51.95
- 39.69
- 12.61

Hadi Quesneville
Location to closest gene
Base composition
Epigenetic status

Conservation → Orthologs

OrthoMCL

Homologous genes (21689)

Select clusters with 4 proteins
1 protein / species

Orthologous genes (6921, 32%)

Extend +500bp upstream

Select clusters with TE annotation overlap

Orthologous genes with TE overlap (6265, 29%)

50%?
Age of insertion

Number of copies

Age

Duster-Specific
Brassicaceae-Specific
TEsBuisine-Specific

Arabidopsis thaliana
Arabidopsis lyrata
Capsella rubella
Thefungiella parvula

Hadi Quesneville
Conclusions

• *Duster* more sensitive and faster than *BLAST* and *MegaBLAST* (*not shown in this presentation*)

• Up to ~50% of *A. thaliana* genome derives from TEs.
  - Official annotation ~20%

• Found old and degenerated TE fragments

• Upstream of genes
Thanks

The « Genome Analysis » team

Joëlle Amselem
Florian Maumus
Françoise Alfama
Nathalie Choisne
Johann Confais
Véronique Jamilloux
Hervé-Antoine Kerjean
Vikas Sharma
Agnès Baud
Mariène Wan

Danielle Nouaud
Dominique Anxolabéhère