Les éléments transposables, de leur annotation à leur intégration en graphes de connaissance

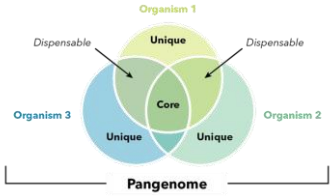
Johann Confais
URGI – Unité de Recherches en Génomique-Info, INRAE Versailles

Journée thématique - Annotation, Intelligence Artificielle et Text-mining
PEPI IBIS le 28/11/2022

Les éléments transposables, de leur annotation à leur intégration en graphes de connaissance

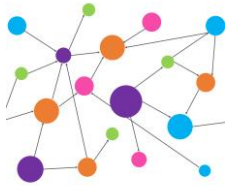
REPET
Annotations TE génomes

1 - Stratégies d'annotation des TEs avec REPET



PanREPET
Construction TE pangénome

2 - Lier des annotation indépendantes pour construire le pangénome des TE

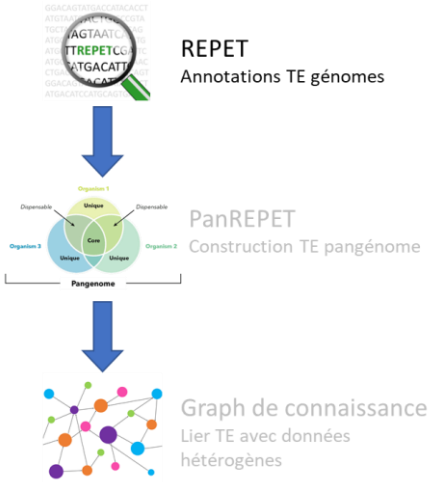


Graph de connaissance
Lier TE avec données hétérogènes

3 - Intégrer les données d'annotation dans un graph de connaissance pour les lier avec d'autres données hétérogènes

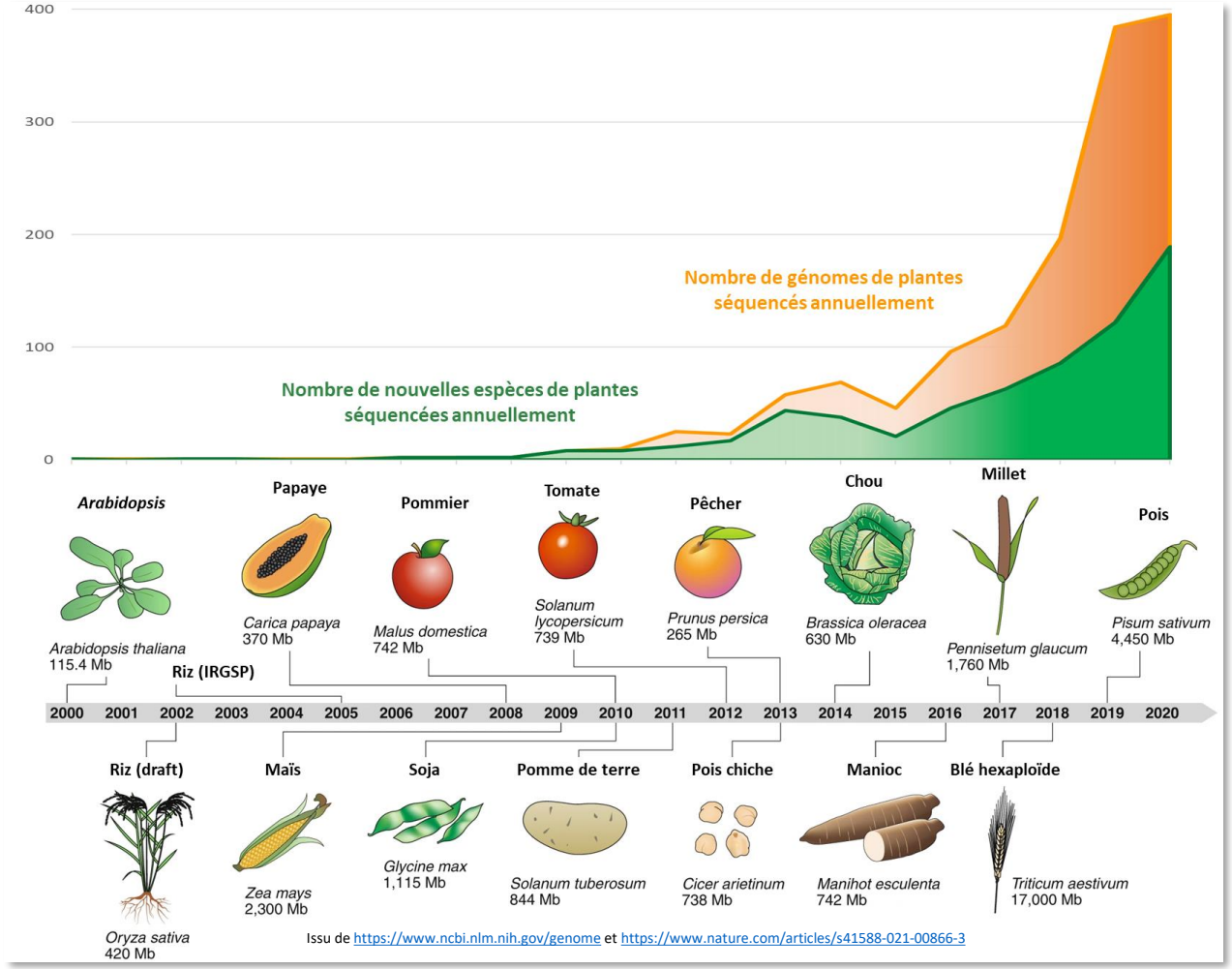
INRAE

L’annotation des TEs dans les génomes avec REPET



Croissance exponentielle du nombre de génomes disponibles

besoin d'annotation de plus en plus efficace

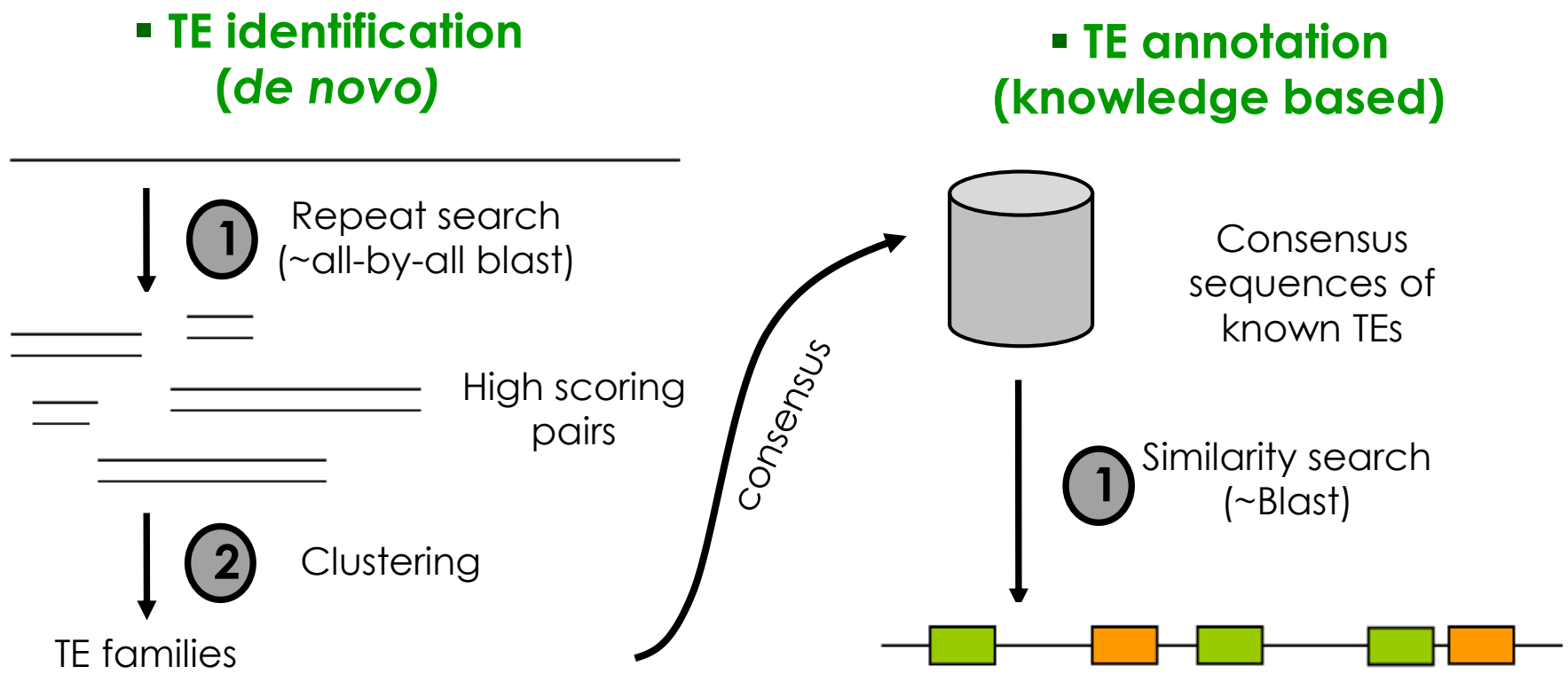


Issu de https://www.ncbi.nlm.nih.gov/genome et https://www.nature.com/articles/s41588-021-00866-3

- Génomes de grande taille
- Nombreux génomes disponibles
- Plusieurs génomes pour une espèce

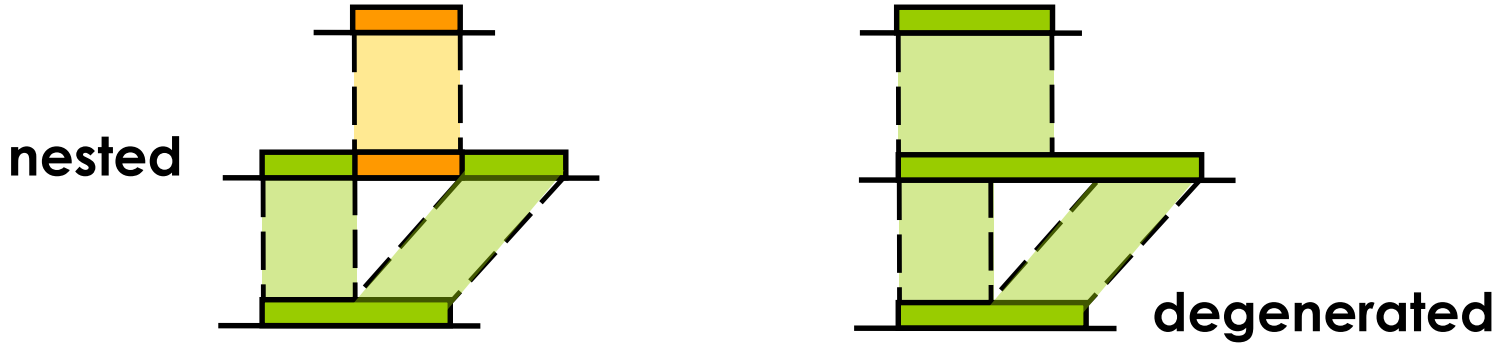
=> Différentes solutions techniques

Principe global des méthodes d'annotation

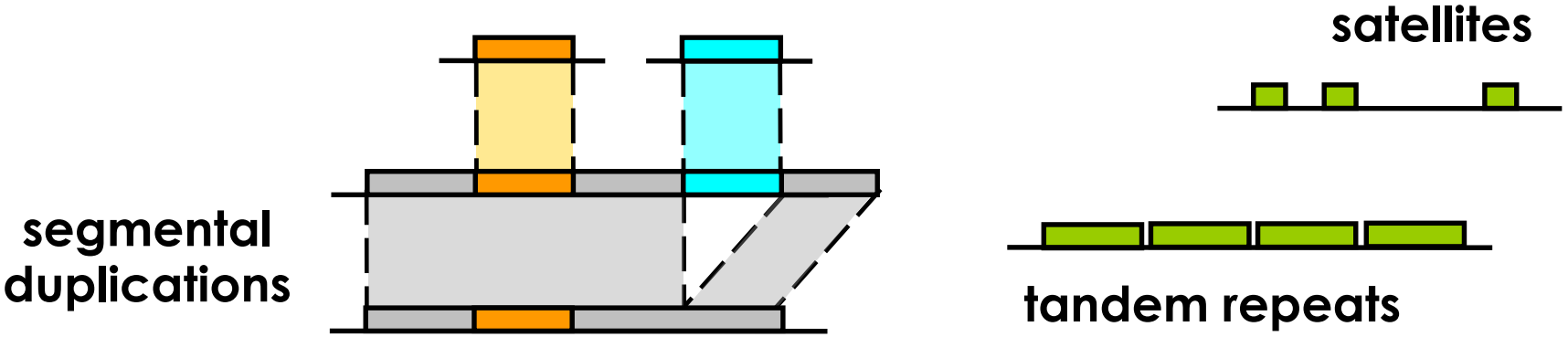


TE annotation challenges

... nested or degenerated elements



... other genomic repeats



INRAE

The REPET "TE denovo" pipeline

Repeat search (~all-by-all blast) 1

```
ATTGCTCGTA
|| ||    ||
ATCGC---TA
```

BLASTER
(Quesneville *et al.* 2003)

```
GTGCTA
|| |||
GTCCTA
```

High Scoring Pairs

Structural search (LTR) 1'

LTRHarvest
(Ellinghaus *et al.* 2008)

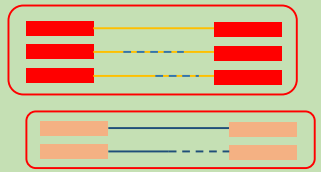
RECON
(Bao and Eddy 2002)
GROUPER
(Quesneville *et al.* 2003)
PILER
(Edgar and Myers 2005)

Clustering 2

```
GTGCTATGCTA
GTCCTA—CTT-
GTGCTATGCTA
GTCCTA—CTT-
GTGCTATGCTA
TGCCTA--GTA
```

```
GTGCTA
GTCCTA
```

2'



MCL
(Van Dongen 2000)
Blastclust (*NCBI*)

TE families

MAP (Huang 1994)
MAFFT (Katoh *et al* 2002)

3 Multiple alignments

```
GTGCTATGCTA
GTCCTA—CTT-
GTCCAAA-CTA
TGCCTA--GTA
```

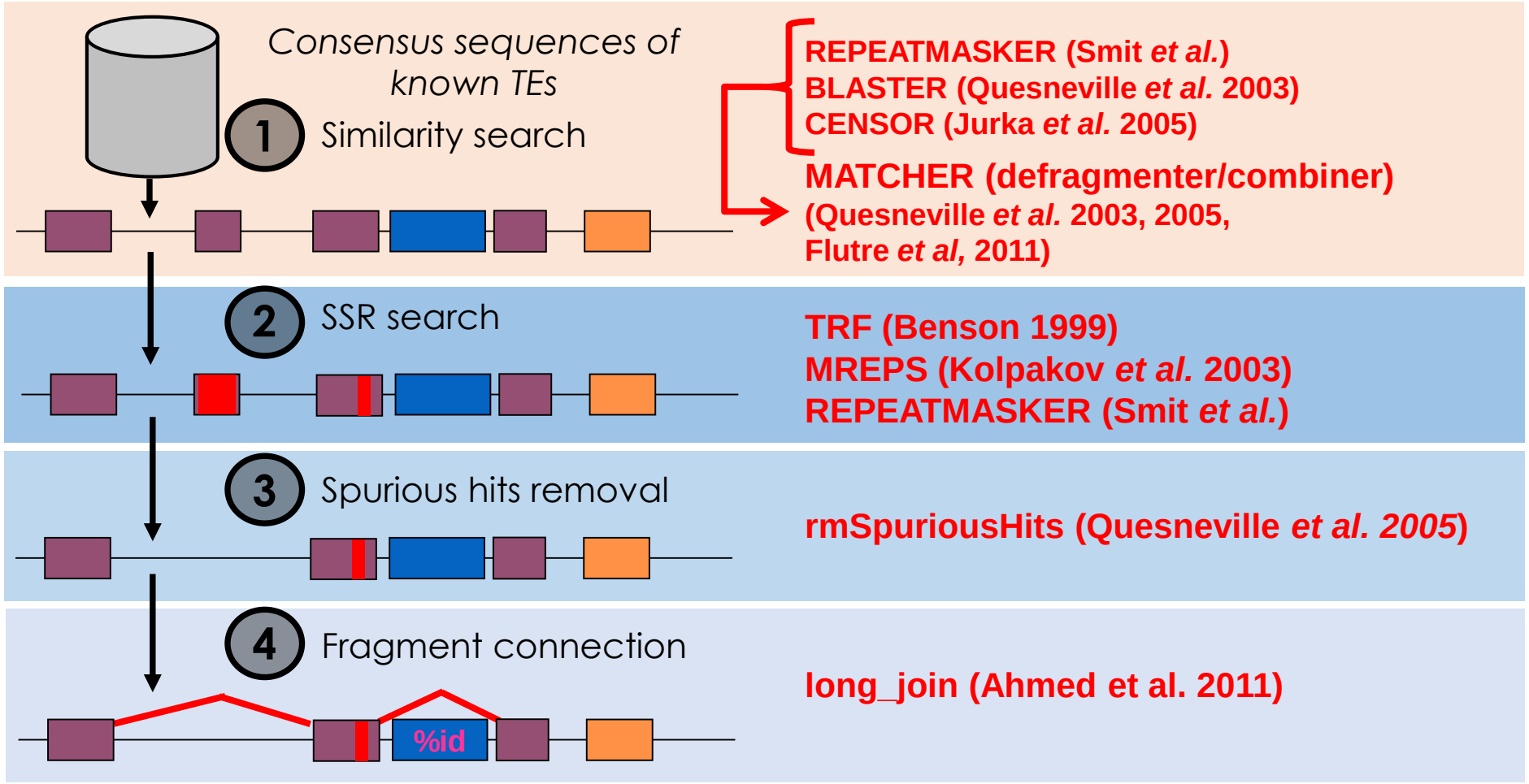
Multiple alignments of repeats

Consensus `GTGCTATGCTA`

PASTEC (TEclassifier/reconciler)

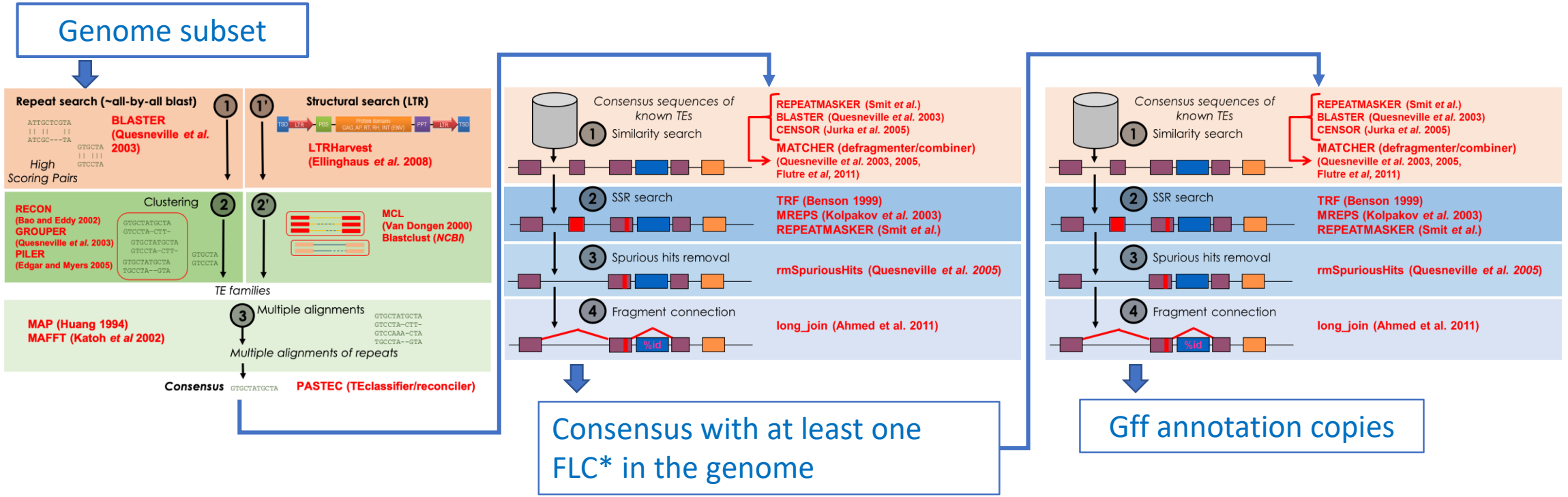
INRAE

The REPET "TE annot" pipeline



Annotation des génomes de grande taille

Stratégie d'annotation

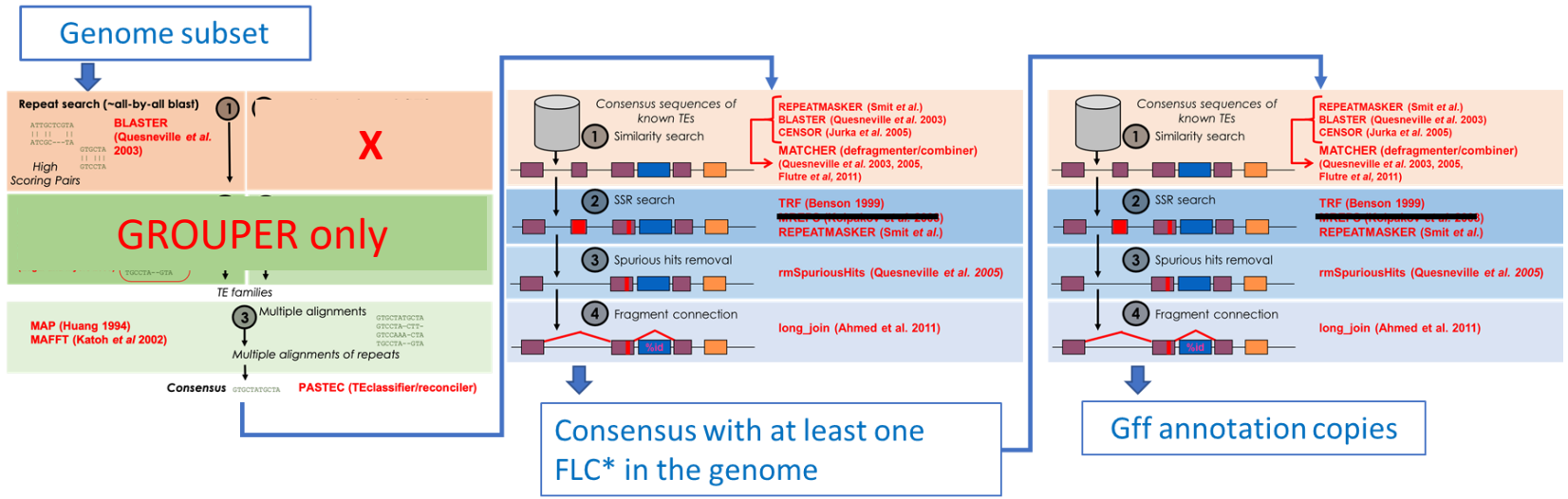


*FLC = Full Length Copy (fragmented and unfragmented annotation aligned over more than 95% of the consensus TE sequence)

V. Jamilloux, *et al. 2017*, "De Novo Annotation of Transposable Elements: Tackling the Fat Genome Issue," *IEEE*, doi: 10.1109/JPROC.2016.2590833.

Annotation en série de plusieurs génomes

Lancements automatisés des annotations de REPET en série : REPET_factory



REPET_factory

*FLC = Full Length Copy (fragmented and unfragmented annotation aligned over more than 95% of the consensus TE sequence)

- Formate le génome d'entrée
- Lance automatiquement toutes les étapes

➔ **REPET_factory : 1 commande pour une annotation complète de batches de génomes**

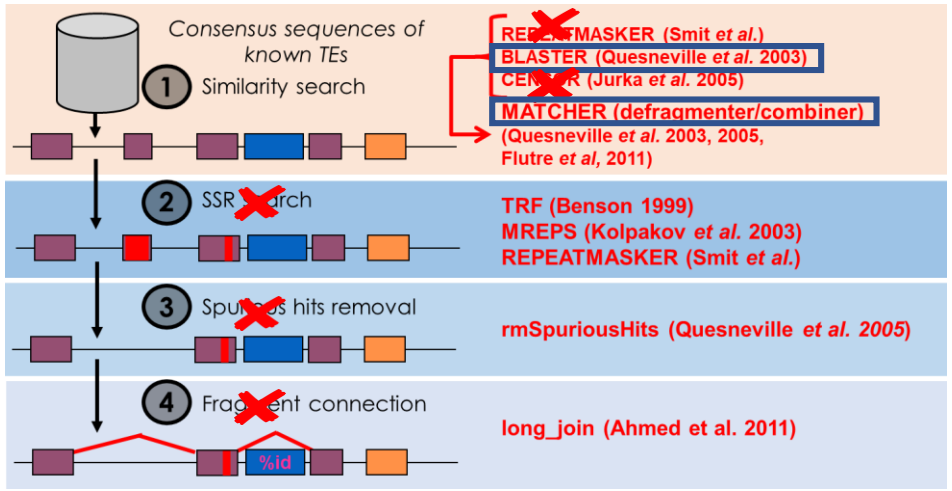
Mariène Wan

Annotation en série de plusieurs génomes

Lancements via snakemake des principales étapes de TEannot

Somia Saidi

=> Développement initié lors de l'atelier du hackathon inter CATI à Sète en octobre 2021.



- Temps d'exécution rapide

⇒ Permet de traiter des grands jeux de données

- Permet d'identifier les copies pleine longueur (a priori jeunes)

=> Réponds au besoin d'approche pangénomique

➔ TEannot en snakemake : annotation rapide pour les besoins en pangénomique ou masquer rapidement le TE

REPET en résumé

Distribution actuelle

https://urgi.versailles.inra.fr/Tools/REPET

https://hub.docker.com/r/urgi/docker_vre_aio
=> simplification de la distribution de REPET

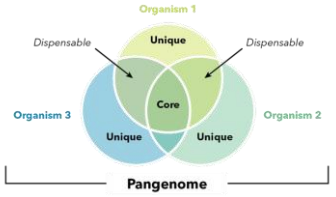
TE Finder :
https://cloud.sylabs.io/library/hquesneville/default/te_finder
Distribution de certain outils de la suite REPET dans une image Singularity/Apptainer

Ce qu'il faut retenir

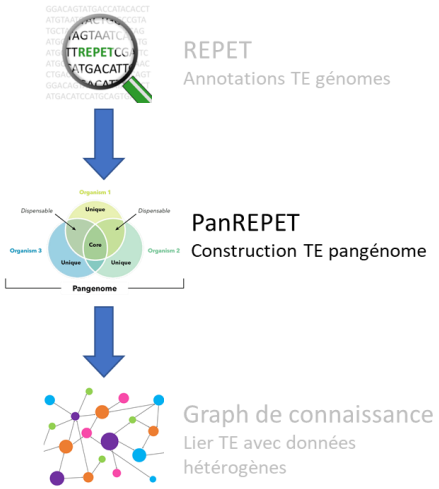
REPET est un outil polyvalent

- Qui peut être ajusté selon les besoins :
  - Rapide ou sensible
  - Pour des petits ou grands génomes
  - Pour annoter plusieurs génome en série
  - Masquer les TEs avant une annotation en gène
- Plus facile à installer
- Plus facile à utiliser
- Un service d'annotation est disponible sur la plateforme

PLANT BIOINFORMATICS FACILITY

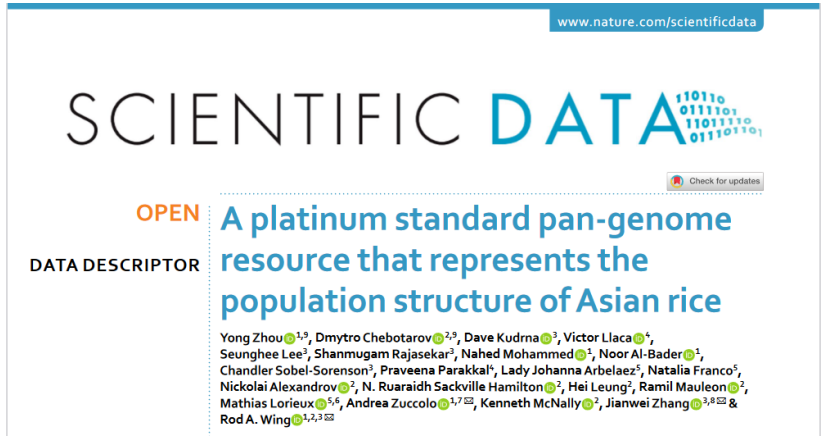


La construction du pan-génome des TEs avec PanREPET



Pan-génome des éléments transposable

Qu’est ce que le pangénome des éléments transposables ?



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

A platinum standard pan-genome resource that represents the population structure of Asian rice

Yong Zhou[1,9], Dmytro Chebotarov[2,3], Dave Kudrna[3], Victor Llaca[4], Seunghee Lee[3], Shanmugam Rajasekar[3], Nahed Mohammed[1], Noor Al-Bader[1], Chandler Sobel-Sorenson[3], Praveena Parakkal[4], Lady Johanna Arbelaez[5], Natalia Franco[5], Nickolai Alexandrov[2], N. Ruaraidh Sackville Hamilton[2], Hei Leung[2], Ramil Mauleon[2], Mathias Lorieux[5,6], Andrea Zuccolo[1,7], Kenneth McNally[2], Jianwei Zhang[3,8] & Rod A. Wing[1,2,3]

16 génomes de riz qualité platinum

Transposable element (TE) prediction

To determine the pan-transposable element content of cultivated Asian rice, we analyzed the 12 new reference genomes, presented here, along with the MH 63, ZS 97, N 22 PacBio reference genomes. In addition, we also included a reanalysis of the IRGSP RefSeq-1.0, as it is conventionally considered the standard rice genome for which all comparisons are conducted.

Pan-génome des éléments transposable

Table 4 Abundance of the major TE classes in the 16 *Oryza sativa* genomes.

From: A platinum standard pan-genome resource that represents the population structure of Asian rice

Variety Name	Total	LTR-RT	LINEs	SINEs	DNA_TEs	Unclassified
NIPPONBARE	46.07	23.55	1.52	0.41	16.18	4.41
CHAO MEO::IRGC 80273-1	46.25	24.00	1.46	0.40	15.59	4.80
Azucena	47.07	24.48	1.47	0.40	15.82	4.89
KETAN NANGKA::IRGC 19961-2	46.99	24.87	1.47	0.40	15.72	4.53
ARC 10497::IRGC 12485-1	46.95	24.74	1.48	0.40	15.68	4.65
PR 106::IRGC 53418-1	47.95	26.82	1.41	0.39	15.05	4.28
Minghui 63	47.97	26.61	1.44	0.4	15.3	4.22
IR 64	47.87	26.82	1.42	0.40	14.97	4.26
Zhenshan 97	47.95	26.79	1.42	0.39	15.19	4.16
LIMA::IRGC 81487-1	48.04	26.87	1.40	0.39	15.01	4.37
KHAO YAI GUANG::IRGC 65972-1	48.27	27.27	1.40	0.39	14.87	4.34
GOBOL SAIL (BALAM)::IRGC 26624-2	48.15	26.99	1.40	0.39	14.99	4.38
LIU XU::IRGC 109232-1	46.92	27.06	1.26	0.32	14.31	3.97
LARHA MUGAD::IRGC 52339-1	48.05	26.74	1.41	0.39	15.09	4.42
N 22::IRGC 19379-1	47.79	25.95	1.44	0.39	15.20	4.81
NATEL BORO::IRGC 34749-1	47.33	25.75	1.42	0.40	15.12	4.64

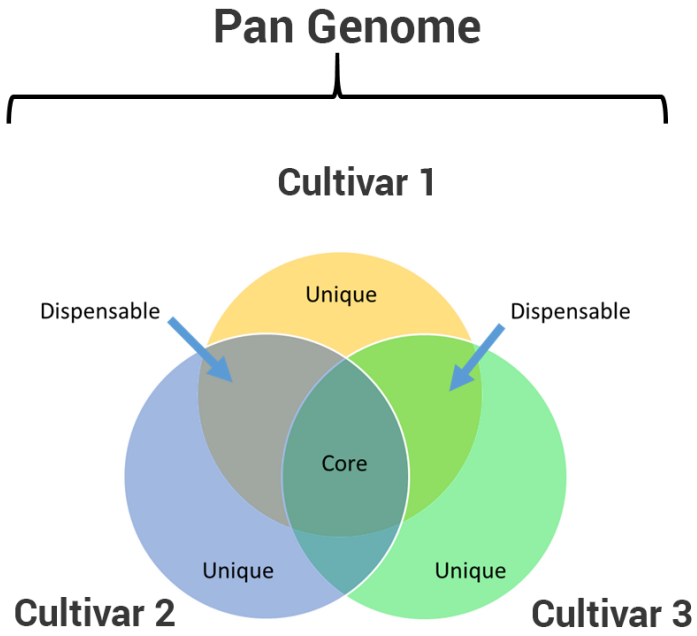
On peut observer que le contenu en TE varie d'un individu à l'autre mais on ne sait pas quelles sont les copies partagées ou pas entre les individus.

Pan-génome des éléments transposable

définition

Une copie de d'élément transposable qui partage une même position dans différents génomes peut avoir 2 statut :

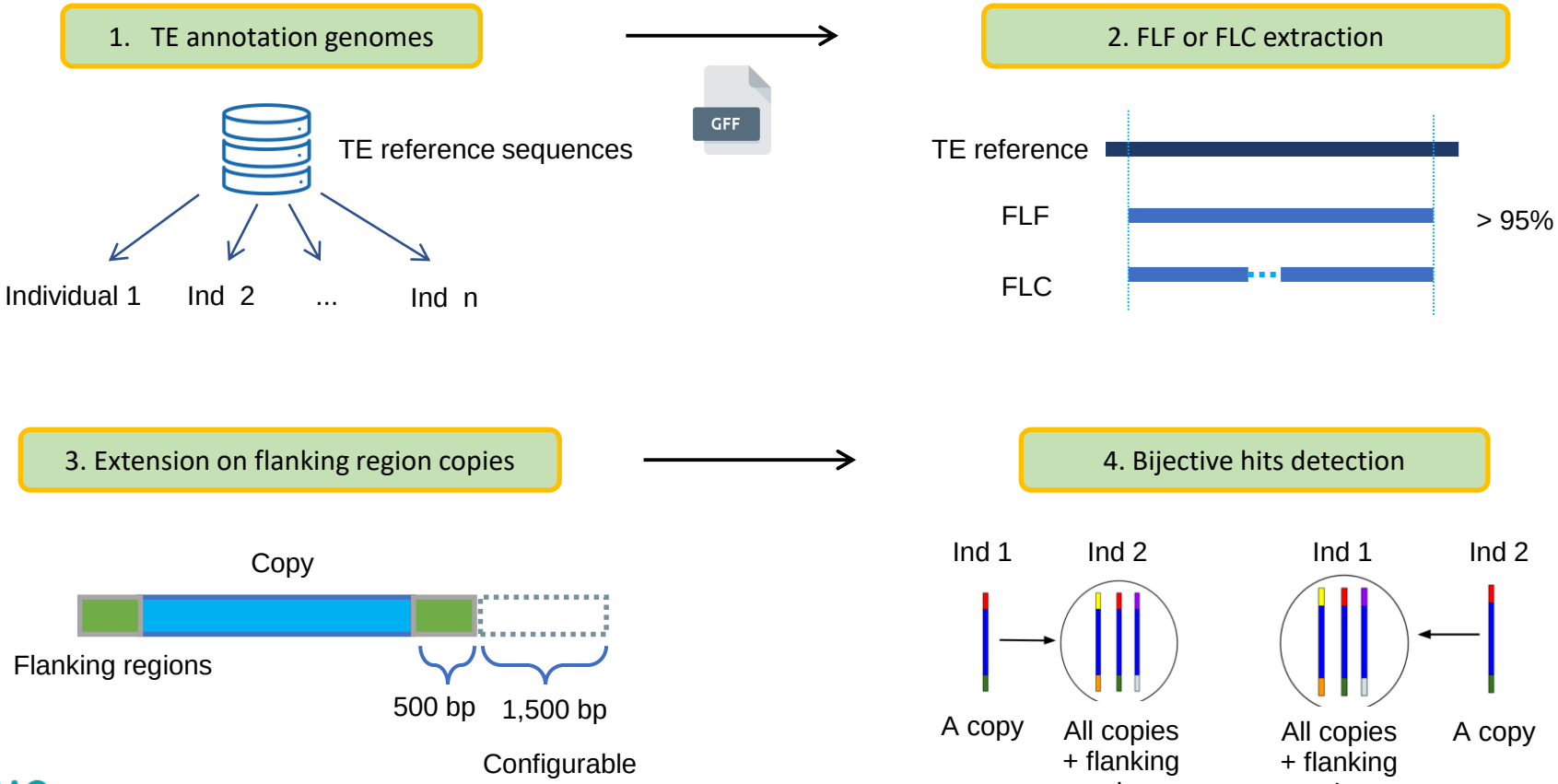
- core = présente chez toutes les accessions
- dispensable = partagées par certaines accessions



Développement de PanREPET

Pipeline en snakemake pour décrire le pan-génome des éléments transposables

Somia Saidi

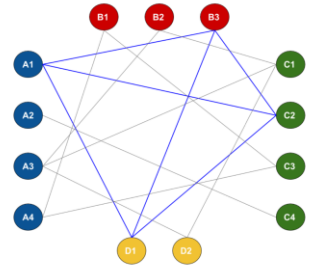


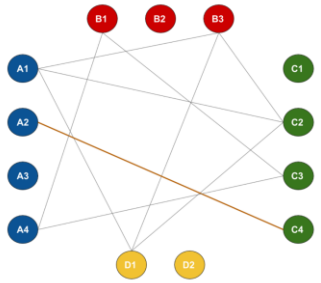
Développement de PanREPET

Pipeline en snakemake pour décrire le pan-génome des éléments transposables

Somia Saidi

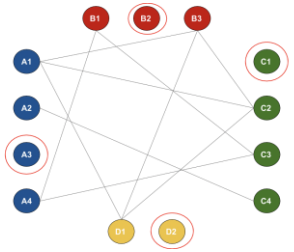
5. Core and Dispensable compartments

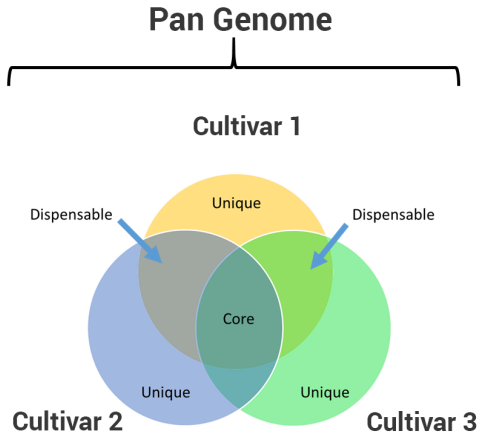




A line represents a **bijective link**
(NetworkX)

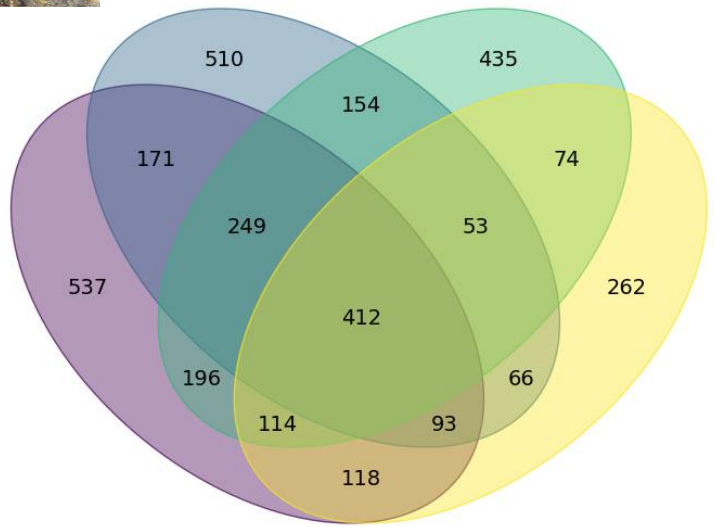
6. Unique copies





Résultats obtenus chez Arabidopsis thaliana ou Oryza sativa

4 génomes d'Arabidopsis



Distribution of copies in pangenome compartments

- 6,352 FLF copies
- 3,444 TE insertions
  - 12% core (412)
  - 37% dispensable (1,288)
  - 51% unique (1,744)
- Col0 and KBSMac74 shared the higher number of copies in dispensable compartments (size 2)

14 génomes de riz

- 222,268 FLF copies
- 56,182 TE insertions
  - 4% core (2,279)
  - 57% dispensable (31,929)
  - 39% unique (21,974)

Résultats obtenus chez Brachypodium distachyon

54 *Brachypodium distachyon* genomes from Phytozome(1)

- Data generated by the same method: **homogeneous**
- 92x median genome coverage, 100 bp paired-end Illumina short reads
- The mean assembled genome size is 268 Mbp, very close to the 272 Mbp reference genome size
- Scaffold L50 of 1 Mbp for the best assembly and an average of 75 kb

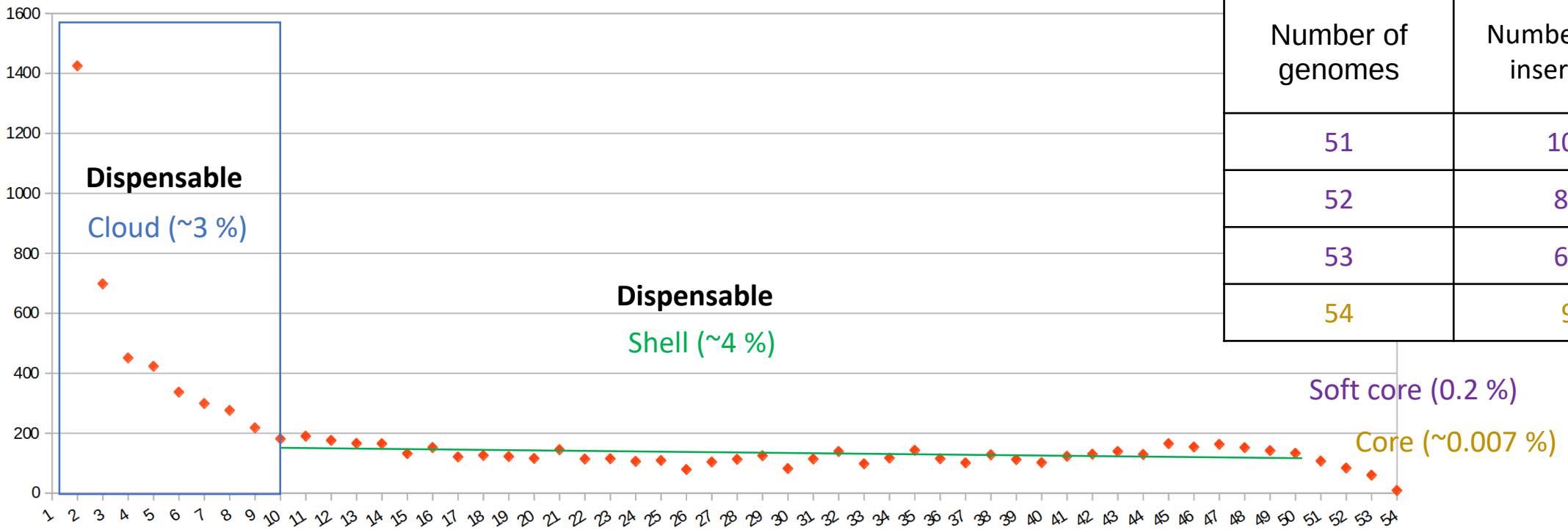
Whole-genome *de novo* assembled

(1) Gordon, Sean P et al. “Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure.” Nature communications vol. 8,1 2184. 19 Dec. 2017

Résultats obtenus chez Brachypodium distachyon

- 286,278 FLC
- 128,220 TE insertions
  - 92 % unique copies (118,244)

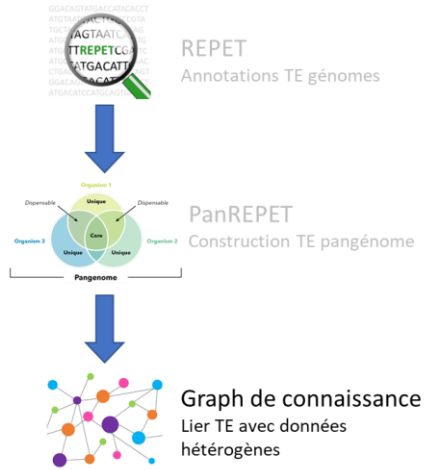
Number of TE insertions



Number of genomes	Number of TE insertions
51	107
52	84
53	60
54	9

Unique copies are not represented on the graph

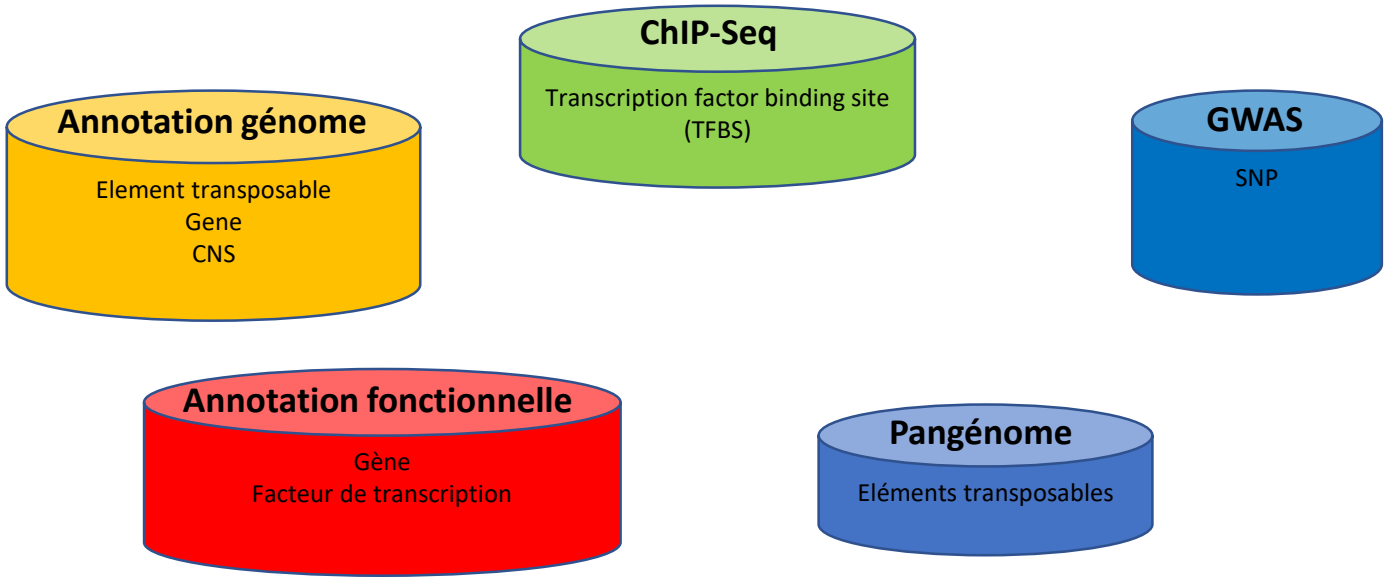
L’intégration des données TE dans un graph de connaissance



Lier les données de pan-génomes des ETs avec d'autres données

Comment inférer l'impact fonctionnel des ET sur le génome hôte ?

Nicolas Francillonne

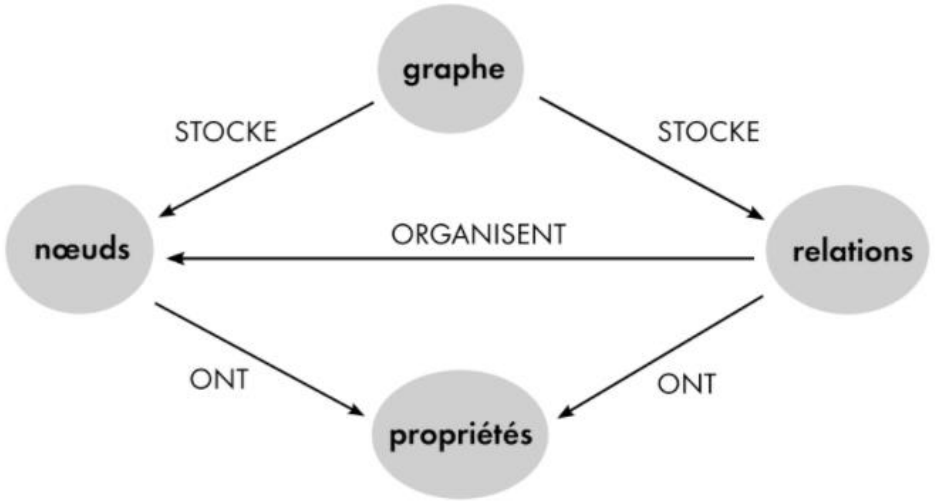


TFBS & TF : PlantRegMap Db (http://planttfdb.cbi.pku.edu.cn/download.php) & Heyndrix et al 2014 (Plant Cell):
TAIR V10 repository : https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_gff3
GWAS: Nordborg study (Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines; Atwell et al. - Nature 2010)
REPETDB https://urgi.versailles.inrae.fr/repetdb
CNS : Van de Velde et al 2014 (Plant Cell)

C’est quoi une base graphe

- Modélisation flexible qui s’adapte à l’hétérogénéité des données disponibles
- Création de relation entre entités qui portent un sens biologique

=> Choix d’un graphe de propriété



- Création de la base graphe avec Neo4J :
  - Modèle dynamique - CRUD + DML (Data Manipulation Language)
  - Méta-modèle généré à la volée à partir des données
  - Cypher simplifie la prise en main

Comment construire notre base ?

Notre méthodologie

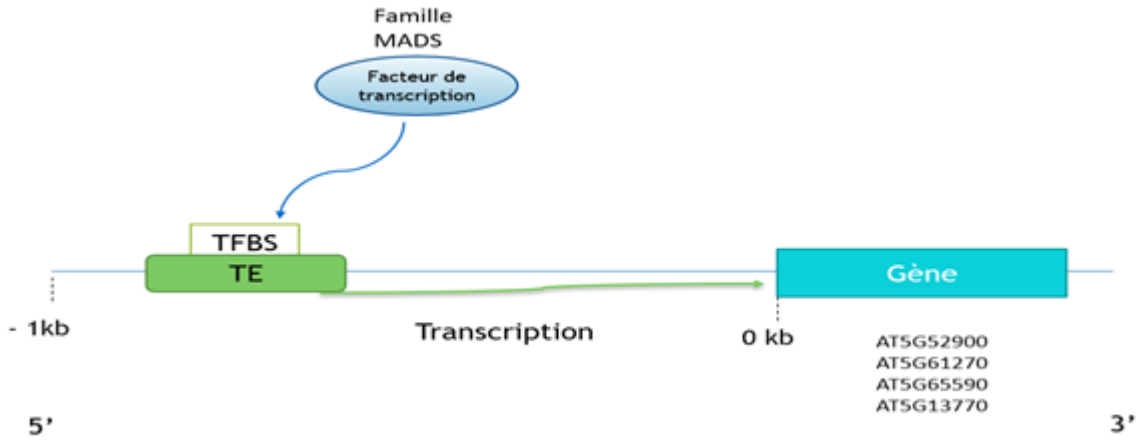
- **Modélisation** des relations entre les données
  - Ontologie
  - Création de termes
- **Pré-compiler** certaines relations pour optimiser la navigation dans le graphe
  - Calcul de la distance entre entités génomiques
- **Développement** de workflow pour extraire, transformer et charger les données (ETL)

Pré-compilation (1)

La question de la proximité d'entités génomiques

L'espace inter-génique chez *A. thaliana* est de 2Kb en moyenne.
=> On cherche des éléments potentiellement régulateurs dans un périmètre d'1 Kb des gènes
=>Pré-compilation = identification de ces relations de proximité :

- Amont
- Aval
- Intersection

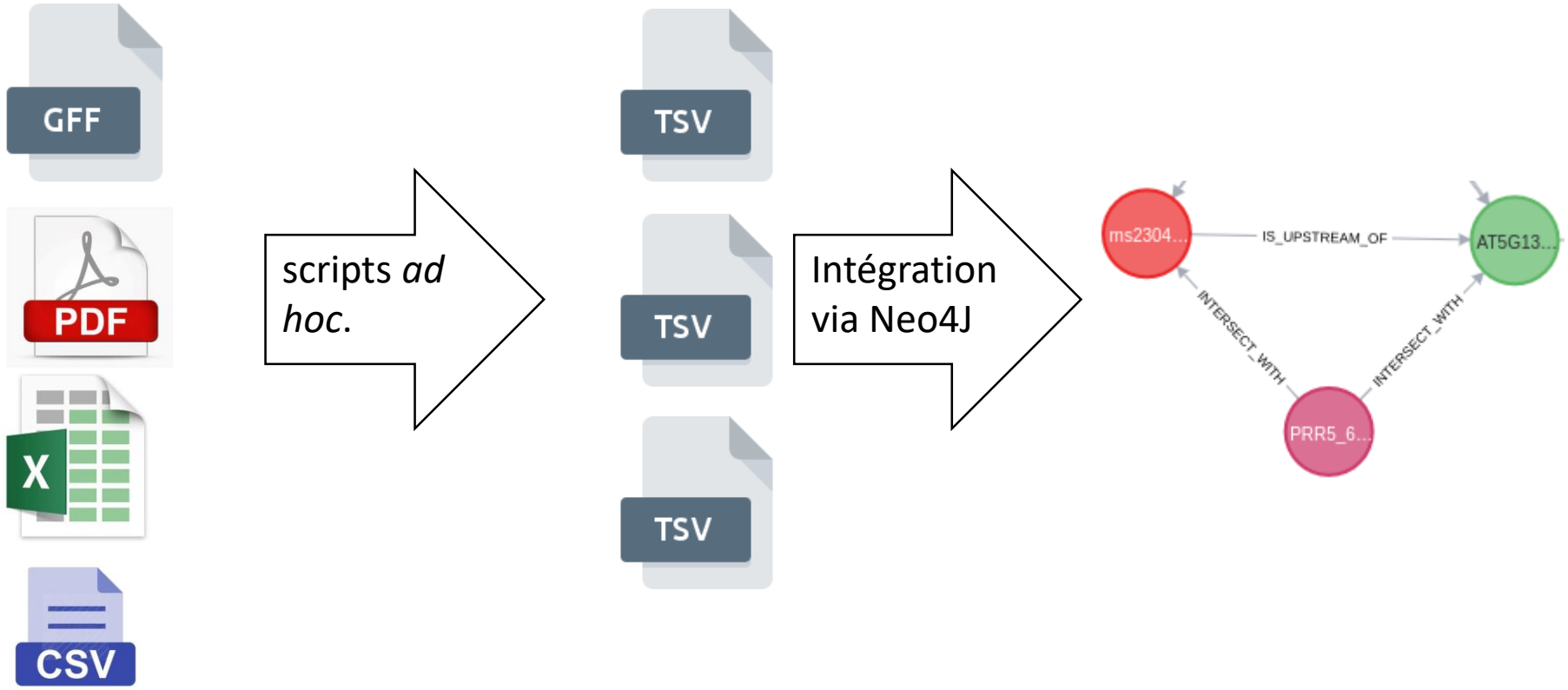


Processus d'intégration

Des données et des formats très hétérogènes

données d'entrée

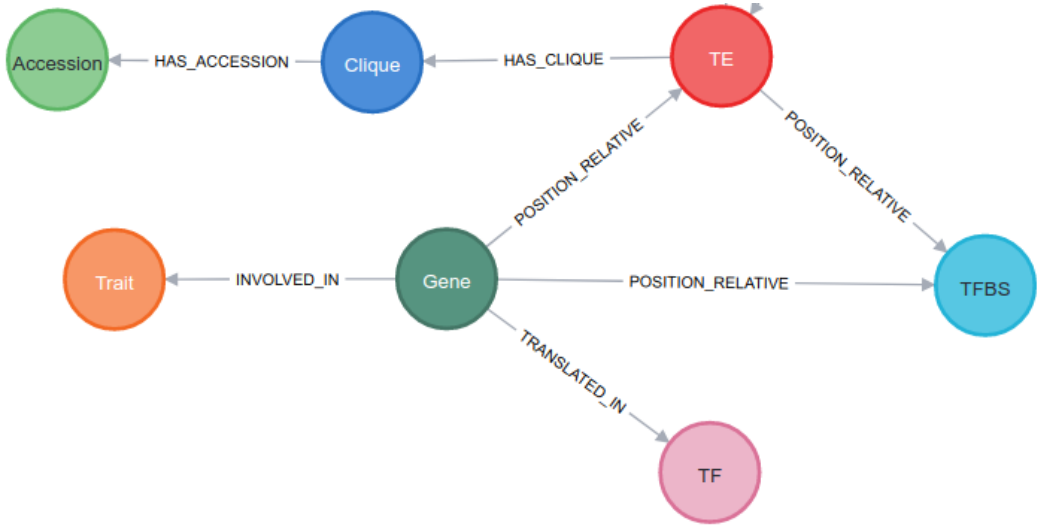
fichiers prêts à être intégrés



Cas d'utilisation

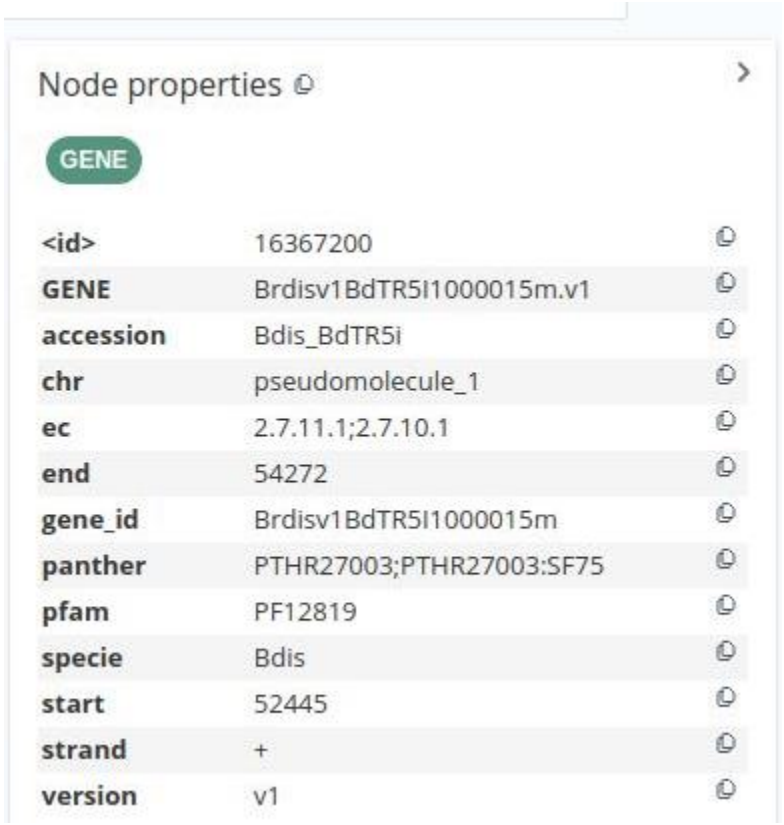
Notre base graphe :

- Nos nœuds :
  - TE, gène, accession, CNS, SNP, stress,TFBS, TF, TRAIT
- Nos relations :
  - POSITION_RELATIVE (IS_UPSTREAM_OF, IS_DOWNSTREAM_OF, INTERSECT_WITH)
  - Involved_in, bind_to, translated_in, CLIQUE_ACCESION, HAS_CLIQUE, GENE_STRESS



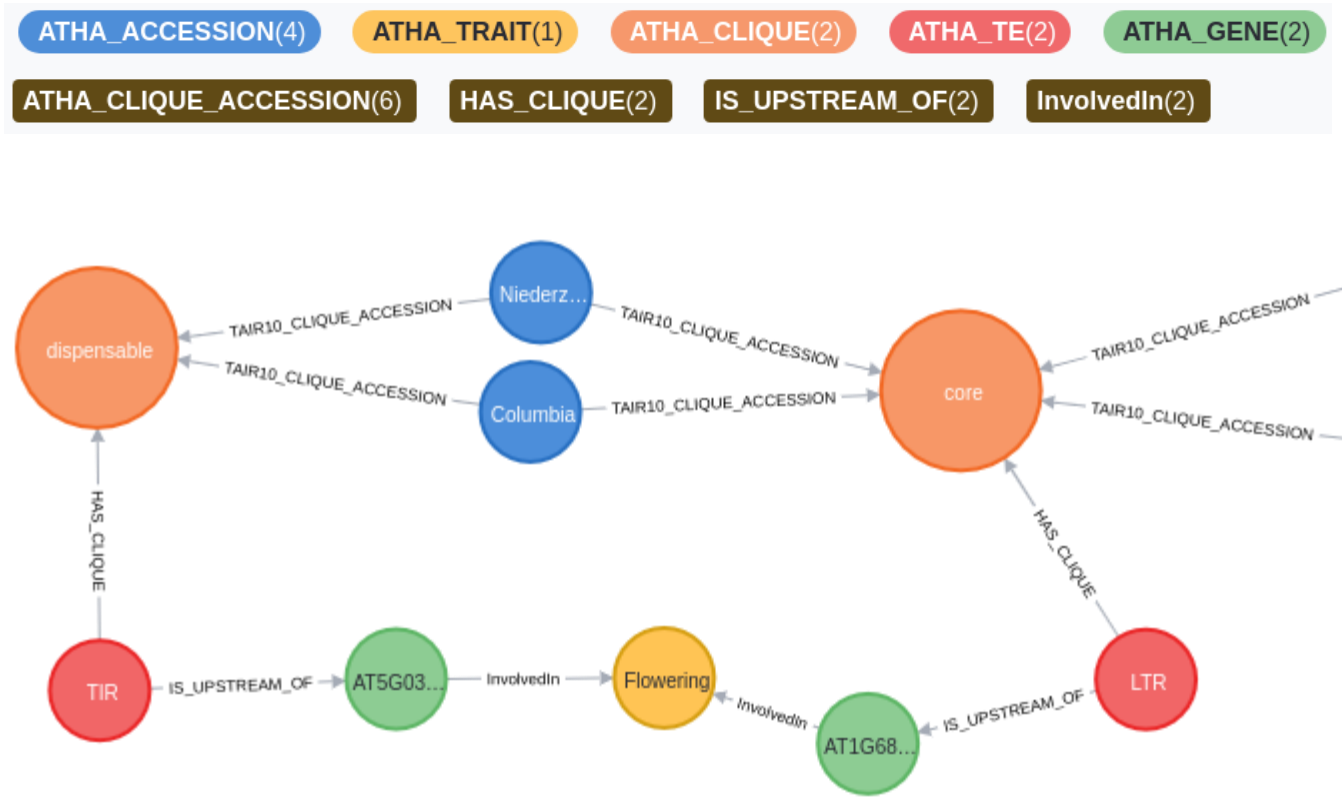
Cas d’utilisation

Information disponible dans un nœud



Exemple de requête dans la base :

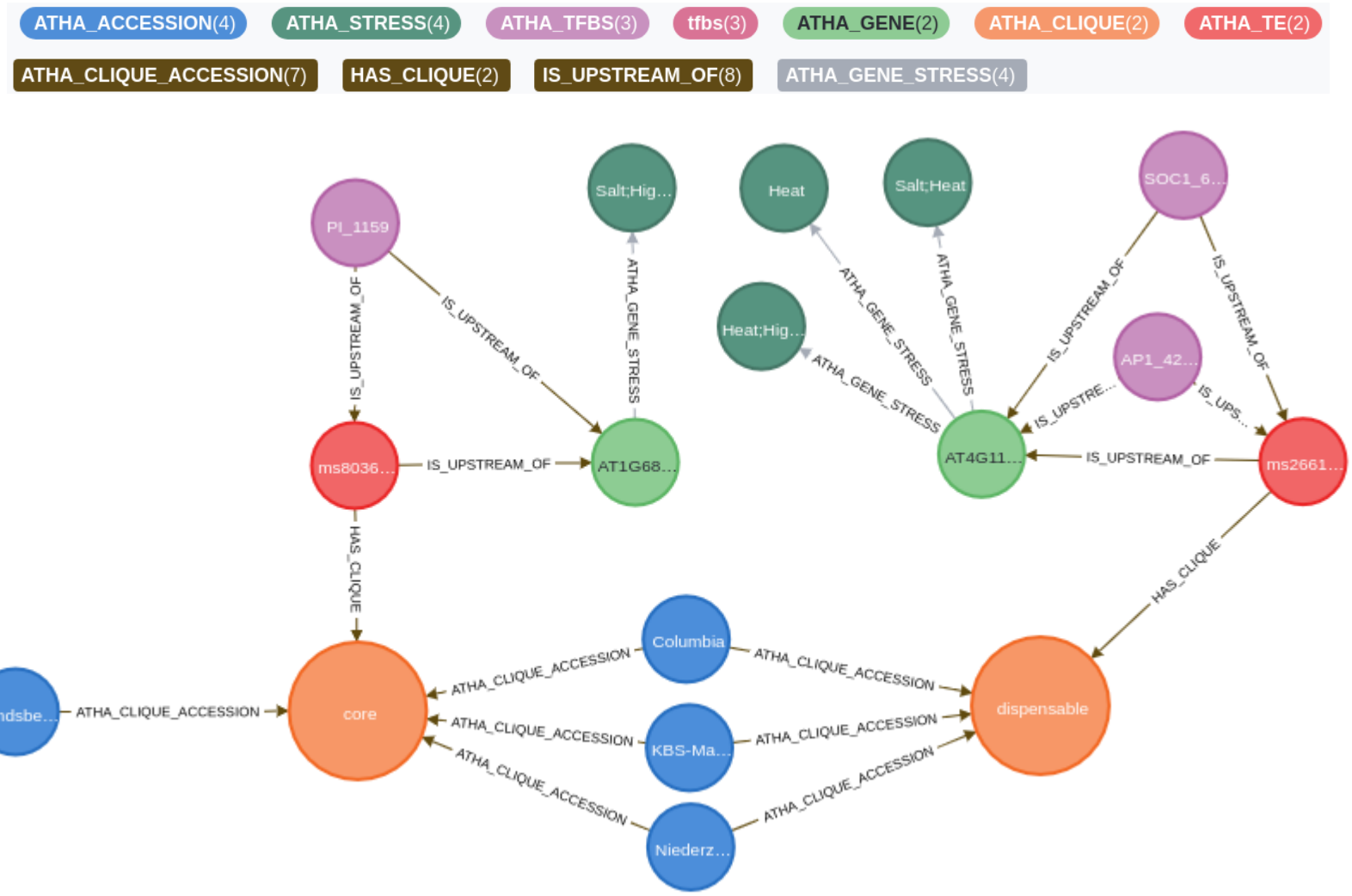
Quels sont les éléments transposables qui pourraient être impliqués dans la régulation fonctionnelle de gènes impliqués dans la floraison ?




```
match p=(a:ATHA_ACCESSION)--(cl:ATHA_CLIQUE)--(te:ATHA_TE)-[:IS_UPSTREAM_OF]-(g:ATHA_GENE)--(:ATHA_TRAIT {TraitName:'Flowering'}) return p
```

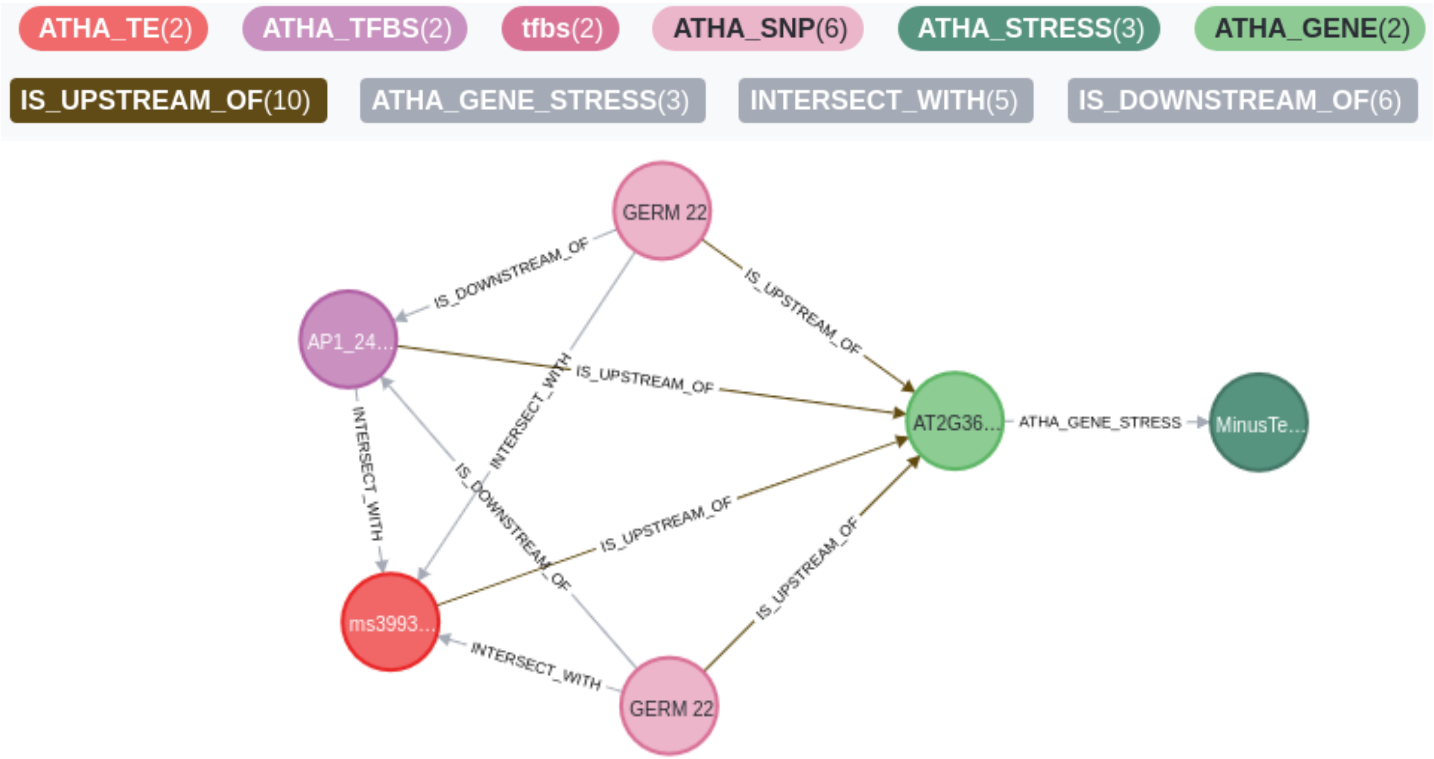
Exemple de requête dans la base :

Quels sont les éléments potentiellement régulateurs (TFBS, TE) en amont de gènes impliqués dans la réponse à un stress ?



Exemple de requête dans la base :

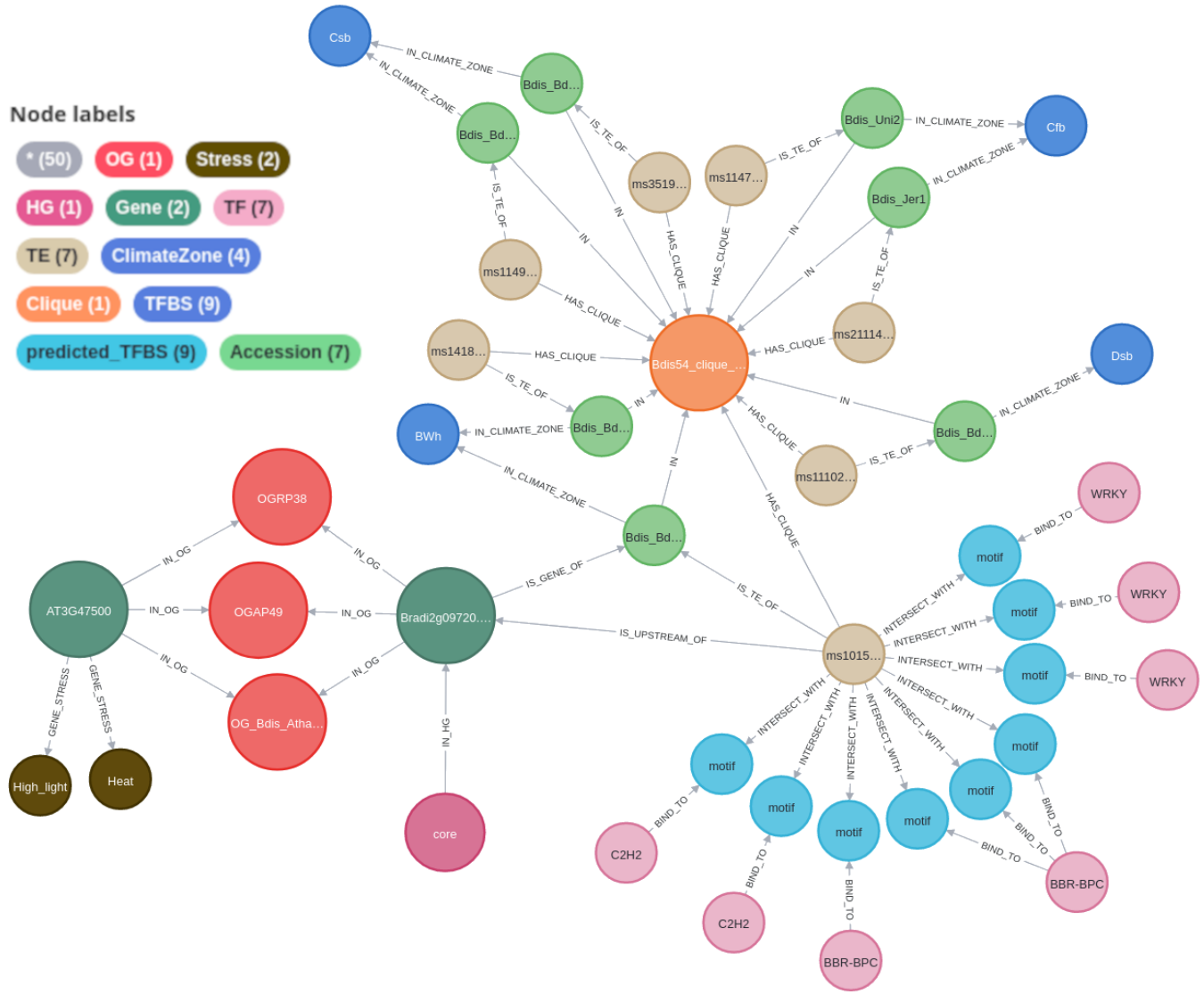
Quels sont les éléments régulateurs (TE, TFBS, marqueur SNP) que l'on peut trouver en amont d'un gène impliqué dans un stress ?




```
match p=(te:ATHA_TE)-[:IS_UPSTREAM_OF]-(g:ATHA_GENE),
tfa=(tf:ATHA_TFBS)-[:IS_UPSTREAM_OF]-(g),
s=(:ATHA_SNP)-[:IS_UPSTREAM_OF]-(g)--(:ATHA_STRESS),
t=(te)-[:INTERSECT_WITH]-(tf)
return p,s,t, tfa
```

INRAE

Visualisation de TEs pouvant avoir un impact sur l'adaptation aux stress de chaleur et de luminosité



Bilan / Perspectives

Intégration dans Neo4J

- Utilisation des données d'orthologie permettant d'inférer la fonction des gènes de B.distachyon
- 30 millions de nœuds et 60 millions de relations en 7-8 minutes

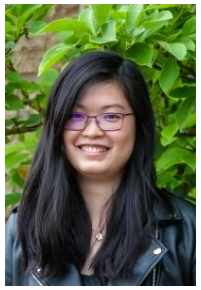
– Prochaine étape ? :

- Insertion de données d'une espèce de plante monocotylédone (*Oryza sativa)*

  => stage M2 en 2023
- Etablir des ponts entre *B. distachyon* et *O. sativa* par le biais des relations d'orthologie
- Inférer des relations chez d'autres espèces (biologie translationnelle)

Remerciements


Somia Saidi


Mariène Wan


Nicolas Francillonne

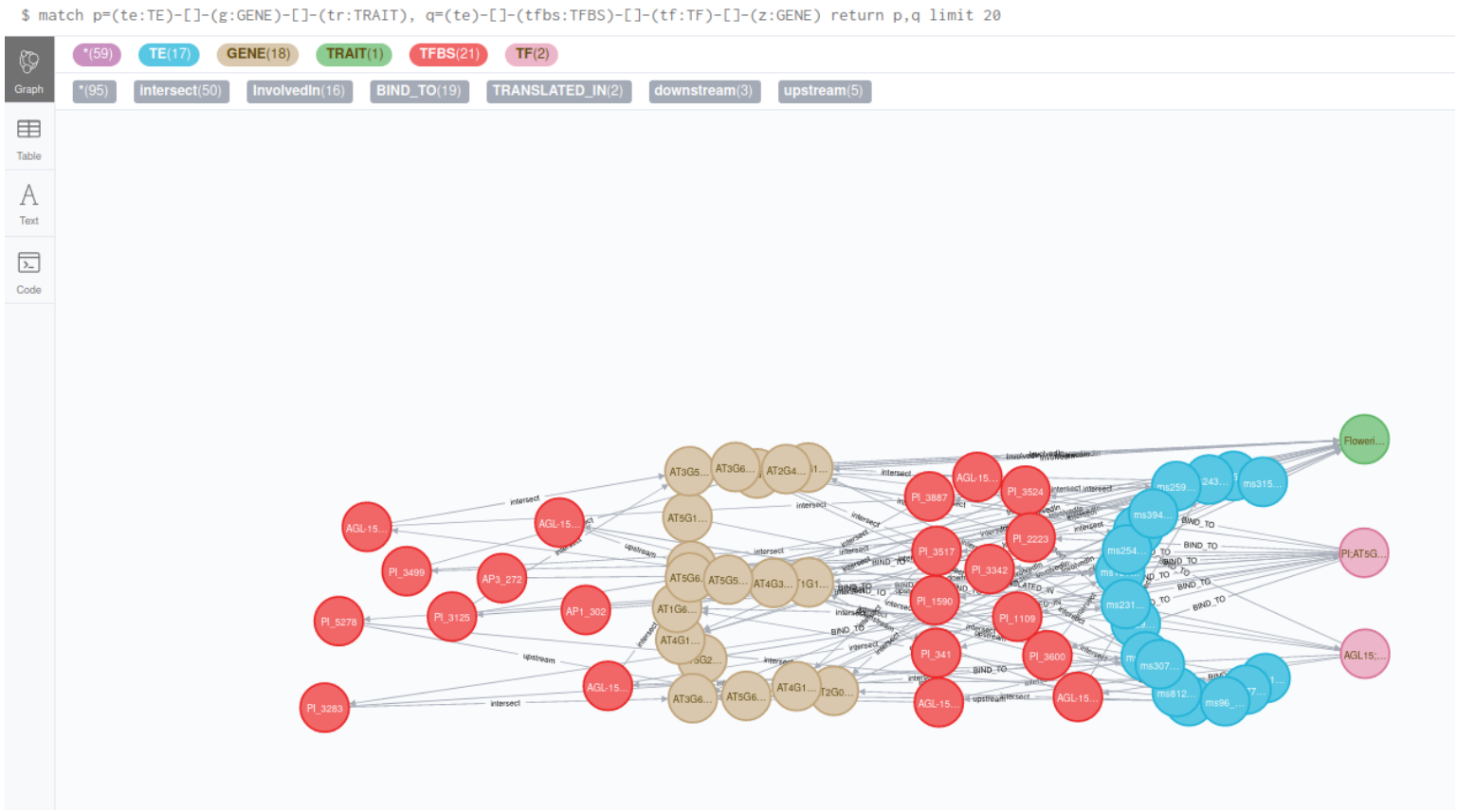

Hadi Quesneville

Étudiants de master 2 :
Mathieu Blaison, Pilar Rodriguez
Yanis Toumert, Maëla Sémery

URGI Team

…. Et tout ceux qui ont participé au développement de REPET

Questions ?


$ match p=(te:TE)-[]-(g:GENE)-[]-(tr:TRAIT), q=(te)-[]-(tfbs:TFBS)-[]-(tf:TF)-[]-(z:GENE) return p,q limit 20
Graph
Table
Text
Code
*(59)
TE(17)
GENE(18)
TRAIT(1)
TFBS(21)
TF(2)
*(95)
intersect(50)
involvedIn(16)
BIND_TO(19)
TRANSLATED_IN(2)
downstream(3)
upstream(5)