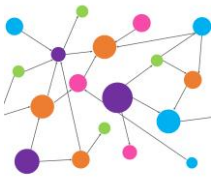


Insertion et exploitation de données hétérogènes dans un graphe de connaissance

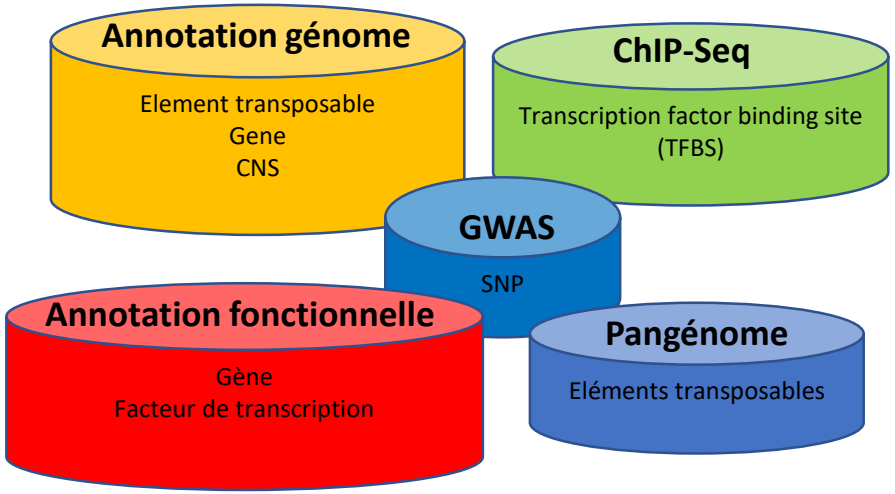
Confais J, Francillonne N, Semery M, Gonnet I

Paris, 2023-09-15

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

Des données hétérogènes, des bases indépendantes et pas de liens



ReMap2022

tair

JASPAR2022

PlantRegMap

JGI Phytozome 12
THE PLANT GENOMICS RESOURCE

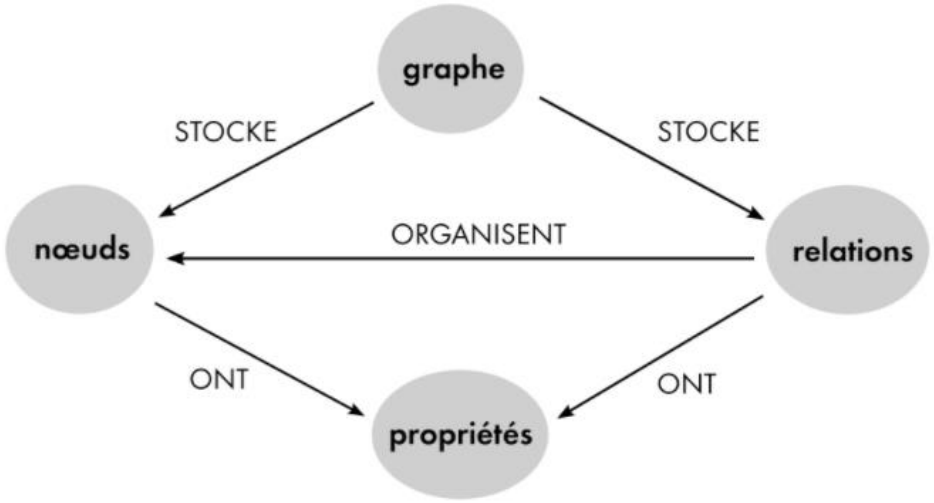
TFBS & TF : PlantRegMap Db (http://planttfdb.cbi.pku.edu.cn/download.php) & Heyndrix et al 2014 (Plant Cell):
TAIR V10 repository : https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_gff3
GWAS: Nordborg study (Genome-wide association study 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)

=> Comment articuler ces données pour répondre à notre question ?

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



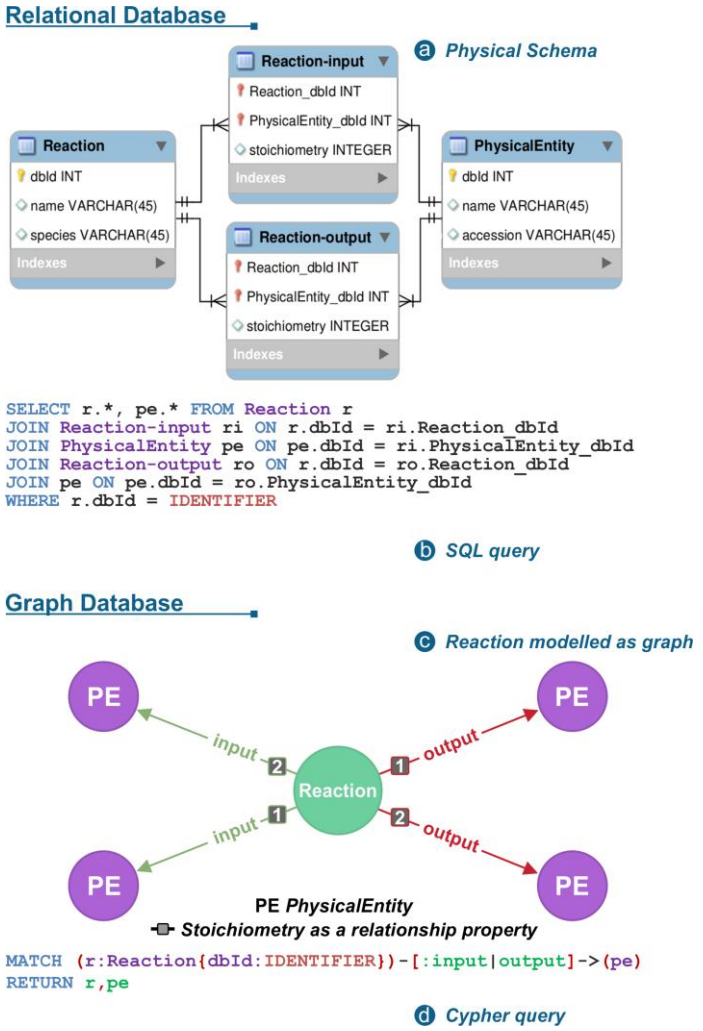
- Création de la base graphe avec Neo4J :
  - Modèle dynamique - CRUD + DML (Data Manipulation Language)
  - Méta-modèle généré à la volée à partir des données
  - Cypher simplifie la prise en main

Intérêts de la base graphe

Reactome graph database : Efficient access to complex pathway data

- Efficace pour les données très connectées (évite les multiples jointures des bases de données relationnelles)
- Type de base plus porté sur l'interrogation que le stockage massif → BDGs pour l'analyse

93% de gain sur le temps de requête



Fabregat et al., Plos Computational Biology (2018)

Types de données

Source	Format des données	Données biologiques
Phytozome	gff, txt	annotation de gènes
	txt	homologie gènes intra *B.distachyon*
	txt	orthologie entre *A.thaliana* et *B.distachyon*
PlantRegMap	gff, txt	groupes d'orthologie (17 & 157 espèces) ; TFBS prédit (motif, motifCE, FunTFBS) ; CNS ; TF
URGI	gff3, tsv, classif	TEs pangénomes
Gordon *et al.*	pdf, csv	climat, classe de floraison, cluster, accessions

⇒ Données d'A.thaliana : Gènes, TEs, CNS, SNPs, TFBS (ChIP-Seq et prédit), TF, Stress, Trait

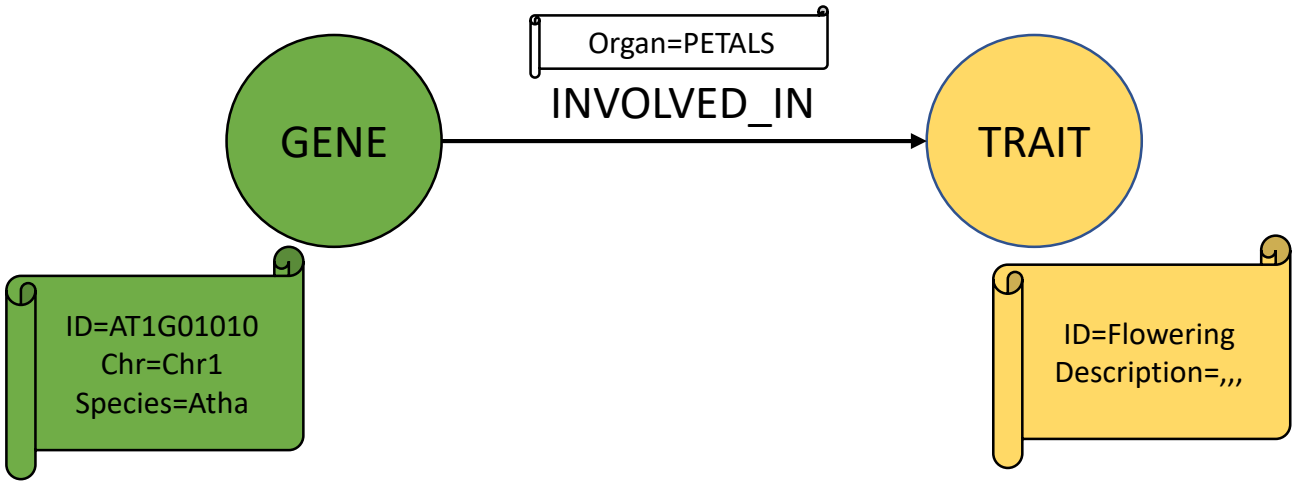
La base graphe avec nos données ?

(Entité)-[relation]-(Entité)

Nœud

Relation

Propriété



Processus d'intégration

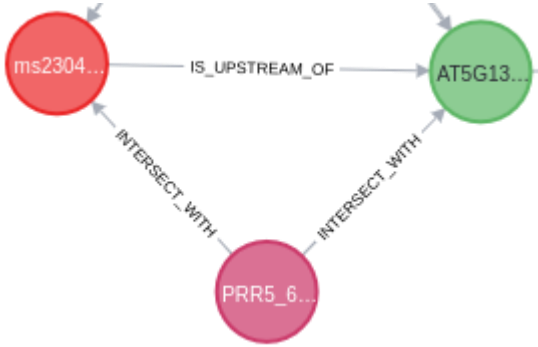
Des données et des formats très hétérogènes

données d'entrée

scripts *ad hoc*.

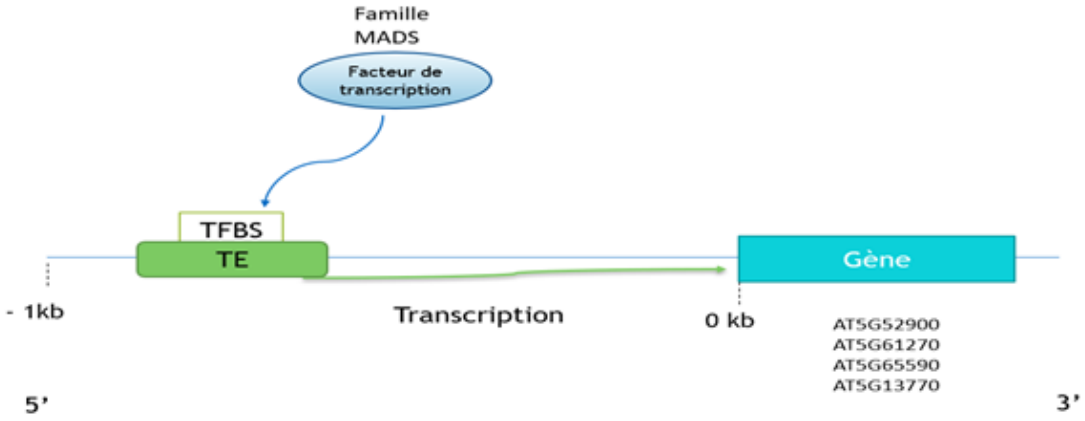
fichiers prêts à être intégrés

Intégration via Neo4J

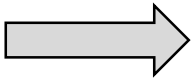


Cas de relations de distance

Données d'annotation fichiers .gff



snakemake



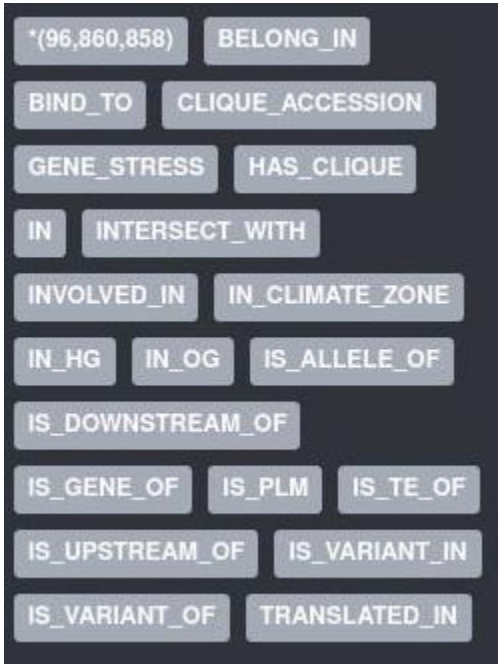
Pybedtools
- Closest
  - Upstream
  - Downstream
- Intersect

relations de distance

Volumétrie

Nœuds

Relation

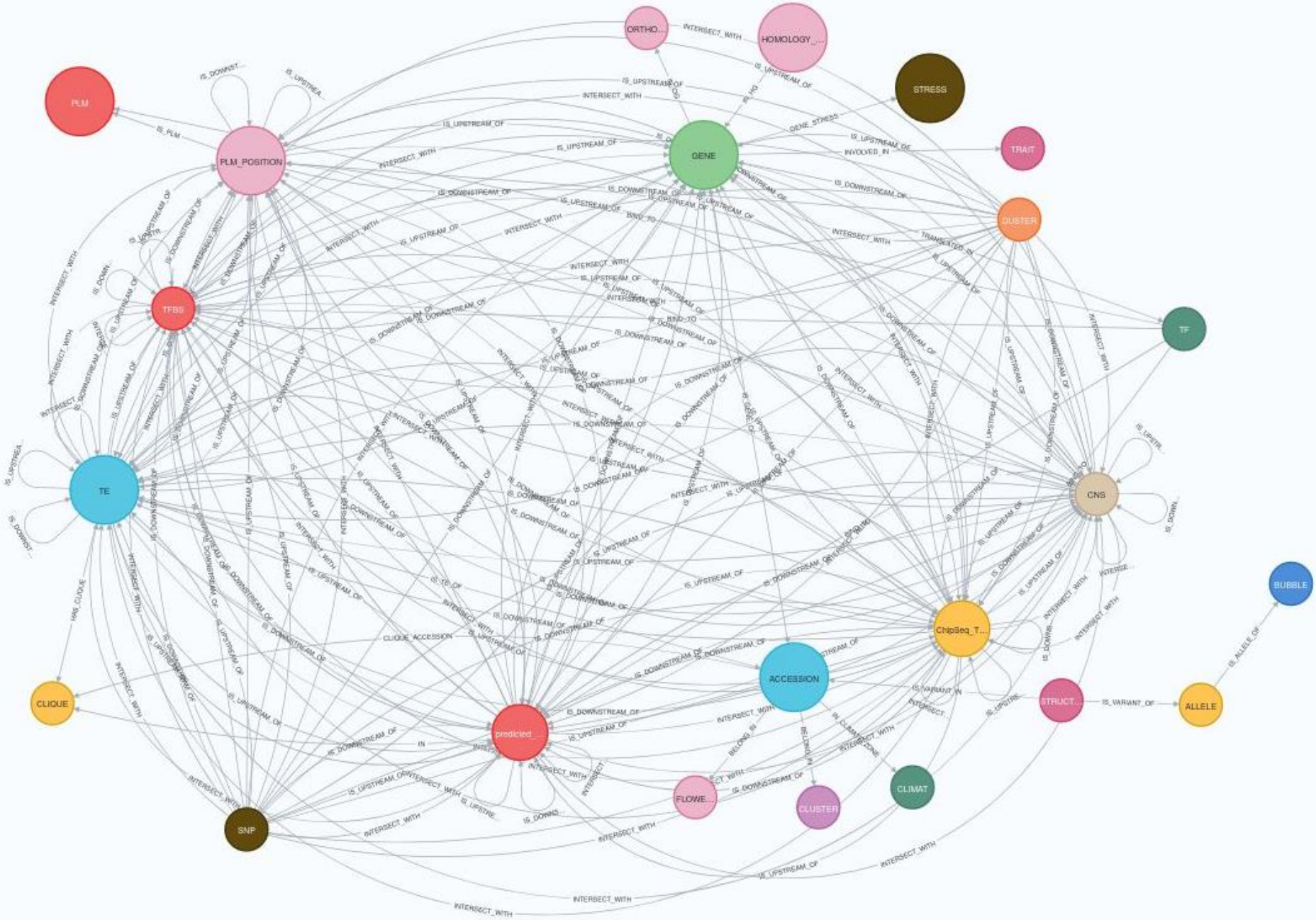


Quelques remarques !
- Plus d'une centaine de propriétés décrivant les nœuds ou relations
- 54 accessions de B.distachyon avec 54 annotations en TE et en gènes

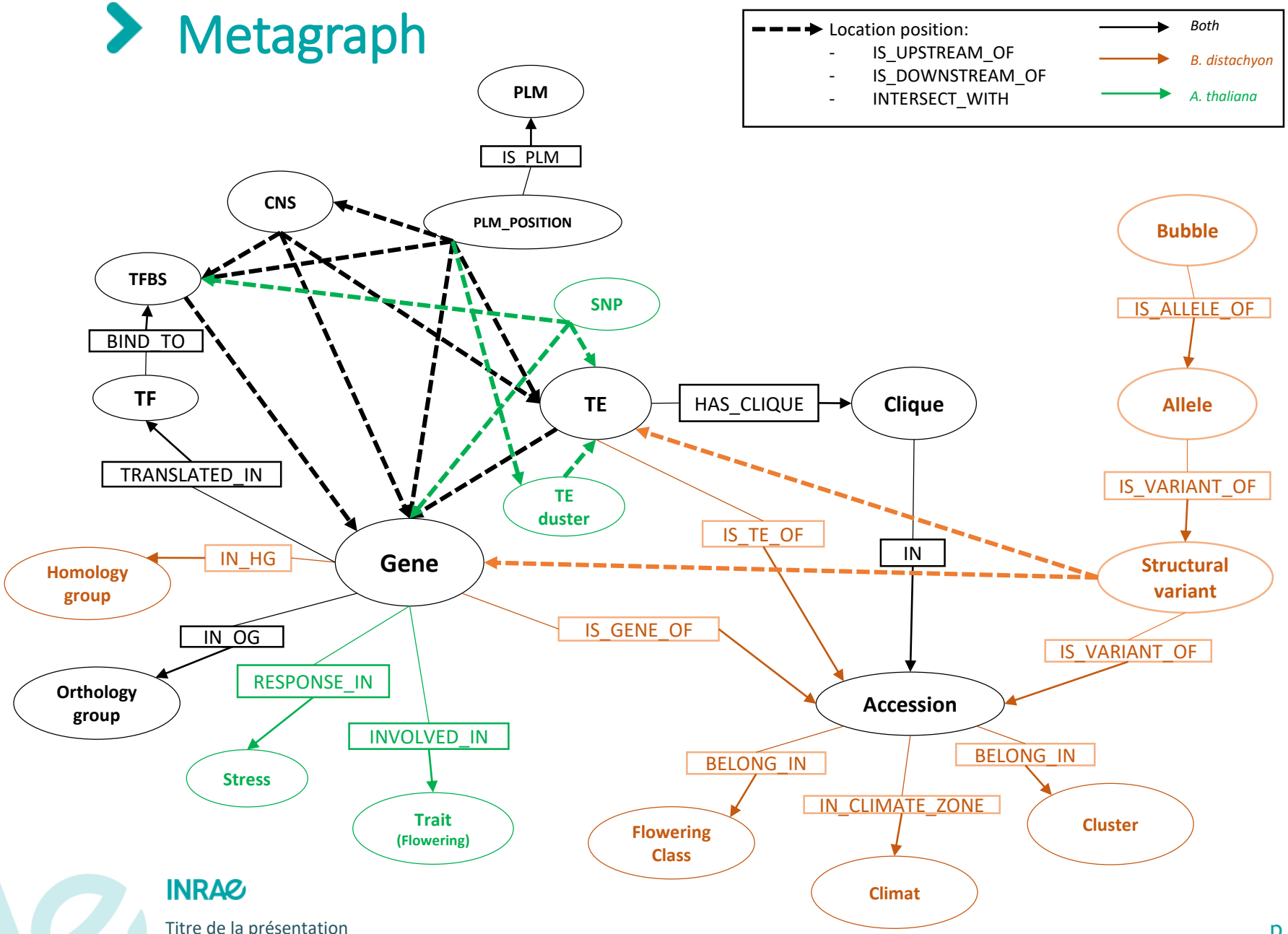
Metagraph

ATTENTION!

ça part dans tous les sens


ORTHO...
HOMOLOGY_...
INTERSECT_WITH
PLM
IS_PLM
PLM_POSITION
STRESS
GENE_STRESS
GENE
TRAIT
INVOLVED_IN
CLUSTER
TRANSCRIBED_IN
TFBS
BIND_TO
TF
CNS
TE
HAS_CLIQUE
CLIQUE
CLIQUE_ACCESSION
ACCESSION
ChipSeq_T...
BUBBLE
IS_ALLELE_OF
ALLELE
STRUCT...
IS_VARIANT_OF
predicted_...
FLOWE...
CLUSTER
CLIMAT
SNP
IS_UPSTREAM_OF
IS_DOWNSTREAM_OF

Metagraph


Location position:
- IS_UPSTREAM_OF
- IS_DOWNSTREAM_OF
- INTERSECT_WITH
Both
B. distachyon
A. thaliana
PLM
IS_PLM
CNS
PLM_POSITION
TFBS
SNP
BIND_TO
TF
TE
HAS_CLIQUE
Clique
TRANSLATED_IN
TE duster
Bubble
IS_ALLELE_OF
Allele
IS_VARIANT_OF
Structural variant
IS_TE_OF
IN
IN_HG
Homology group
Gene
IS_GENE_OF
IS_VARIANT_OF
IN_OG
Orthology group
RESPONSE_IN
Accession
INVOLVED_IN
Stress
Trait (Flowering)
BELONG_IN
BELONG_IN
IN_CLIMATE_ZONE
Flowering Class
Cluster
Climat

Pleins de données mais pour faire quoi ?

3 cas d'utilisation de la base

- 1er cas: Quelles seraient les séquences régulatrices ancestrales impliquées dans la floraison

- 2e cas: Visualisation de TEs pouvant avoir un impact sur l'adaptation aux stress de chaleur et de luminosité

- 3e cas: Associer les familles de TE à des motifs fonctionnels

Cas d'utilisation

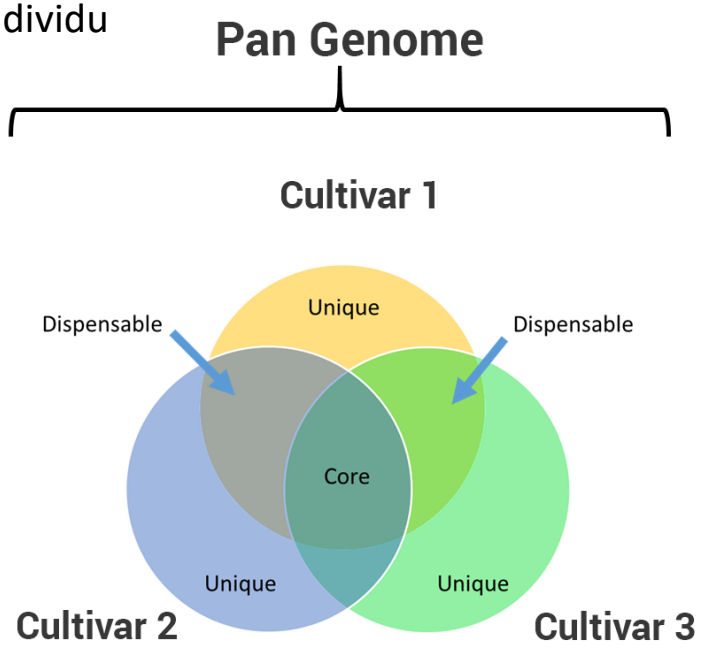
Étude du pangénome des éléments transposables (TEs) chez *Arabidopsis thaliana*

- *Arabidopsis thaliana* est une espèce modèle
- 4 accessions séquencées et annotées en éléments transposable
- Absence/Présence d'éléments transposables chez chaque individu
- Définition TE en clique :

TE qui partage une même position

dans différents génomes :

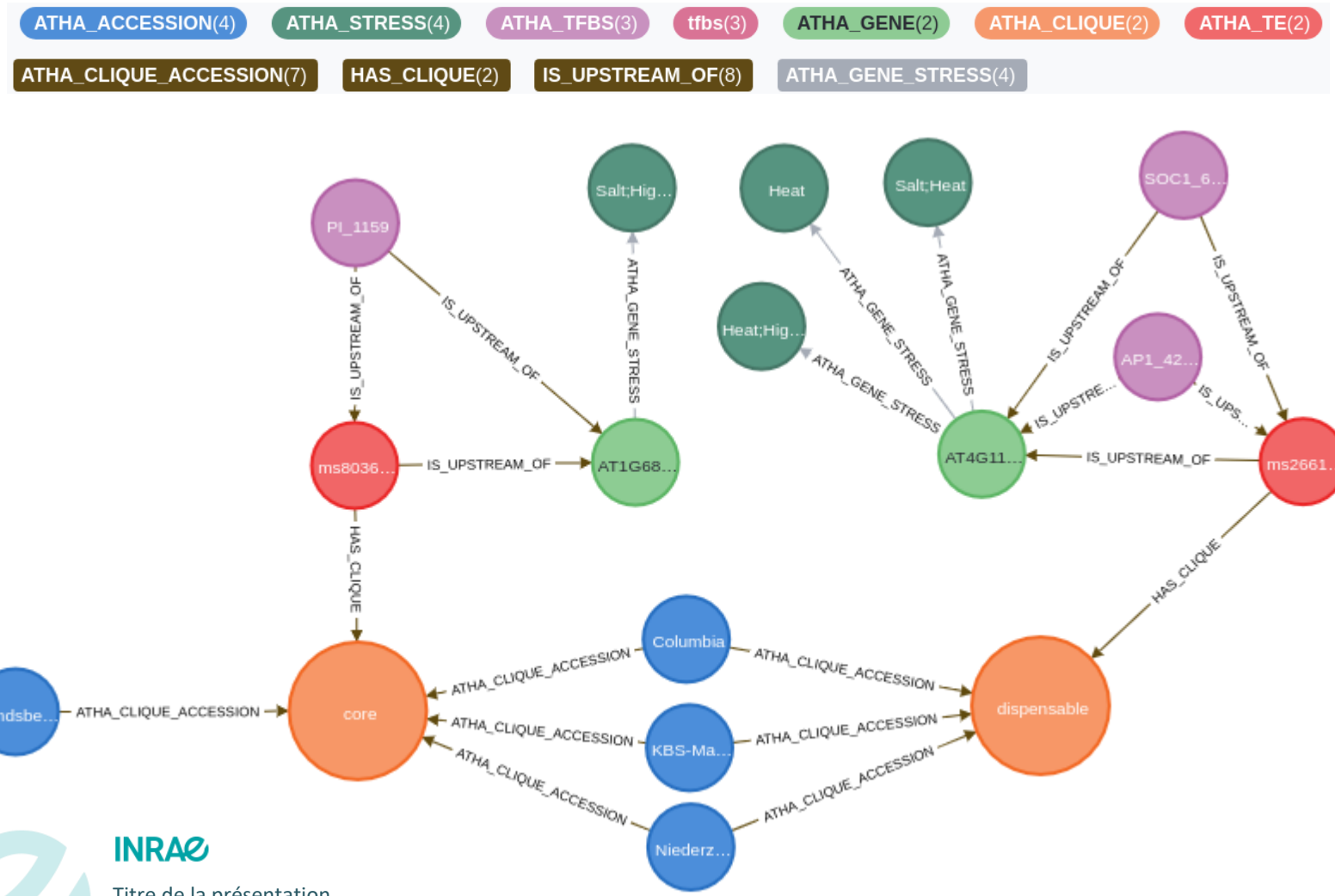
- core = présente chez toutes les accessions
- dispensable = partagées par certaines accessions



=> Peut-on inférer aux éléments transposables un impact sur certains phénotypes exprimés dans certaines variétés/accessions ?

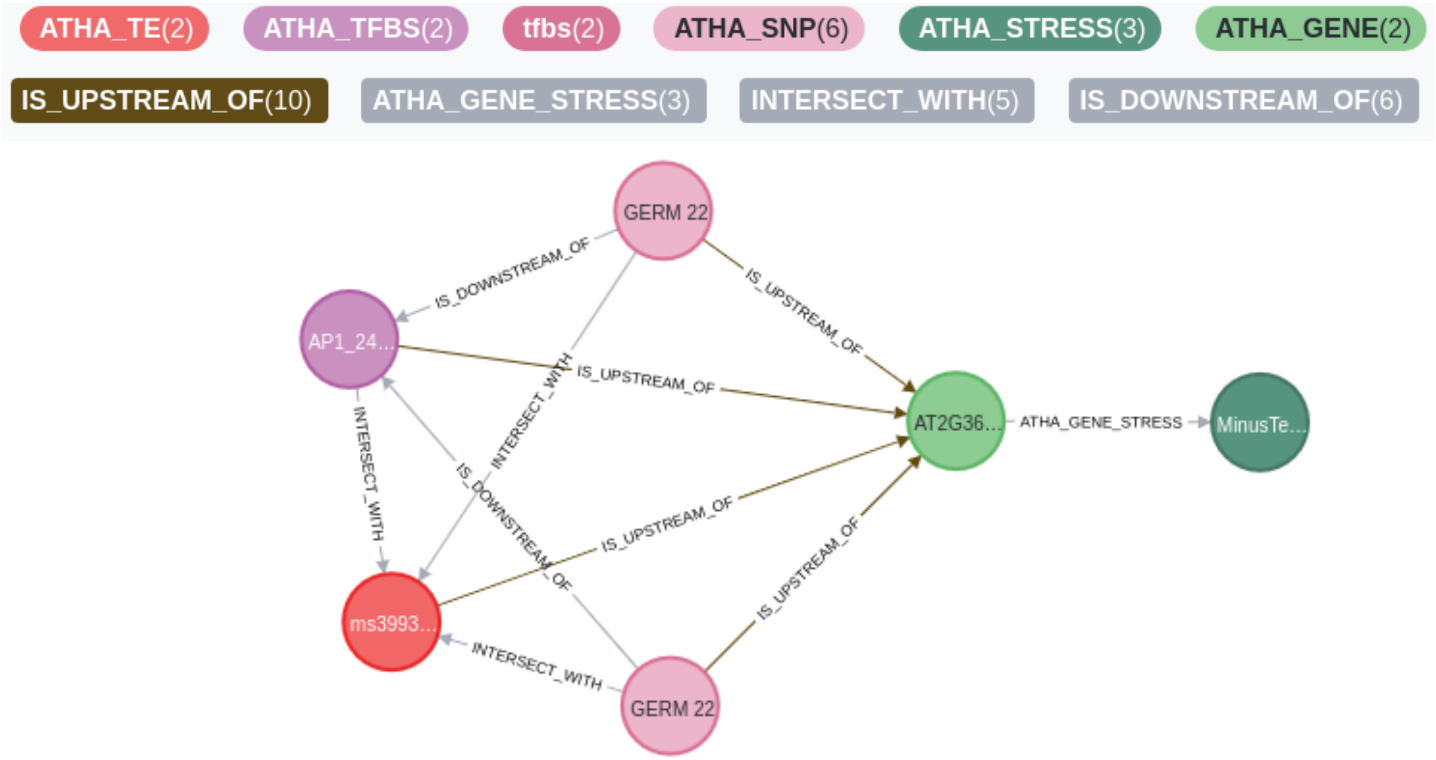
Exemple de requête dans la base :

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



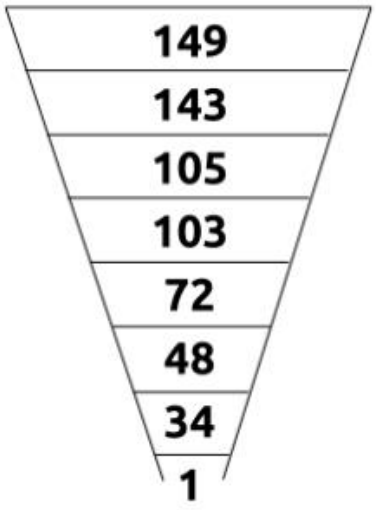
Exemple de requête dans la base :

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
```

1er cas: Quelles seraient les séquences régulatrices ancestrales impliquées dans la floraison



Conditions cumulées

Gènes d'*A.thaliana* impliqués dans la floraison

\+ CNS en amont des gènes d'*A.thaliana*

\+ CNS qui chevauche un TFBS validé par analyse ChIP-Seq

\+ ChIP-Seq TFBS se liant avec un TF

\+ Gène d'*A.thaliana* orthologue à un gène de *B.distachyon*

\+ CNS en amont des gènes de *B.distachyon*

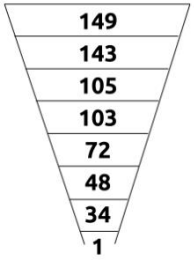
+CNS de *B.distachyon* chevauche un TFBS prédit se liant à un TF

+TF d'*A.thaliana* et de *B.distachyon* de la la même famille

Puis affichage des TFBS prédits d'*A.thaliana* et des groupes d'homologie et d'orthologie de *B.distachyon*

Use case

Quelles seraient les séquences régulatrices ancestrales impliquées dans la floraison ?



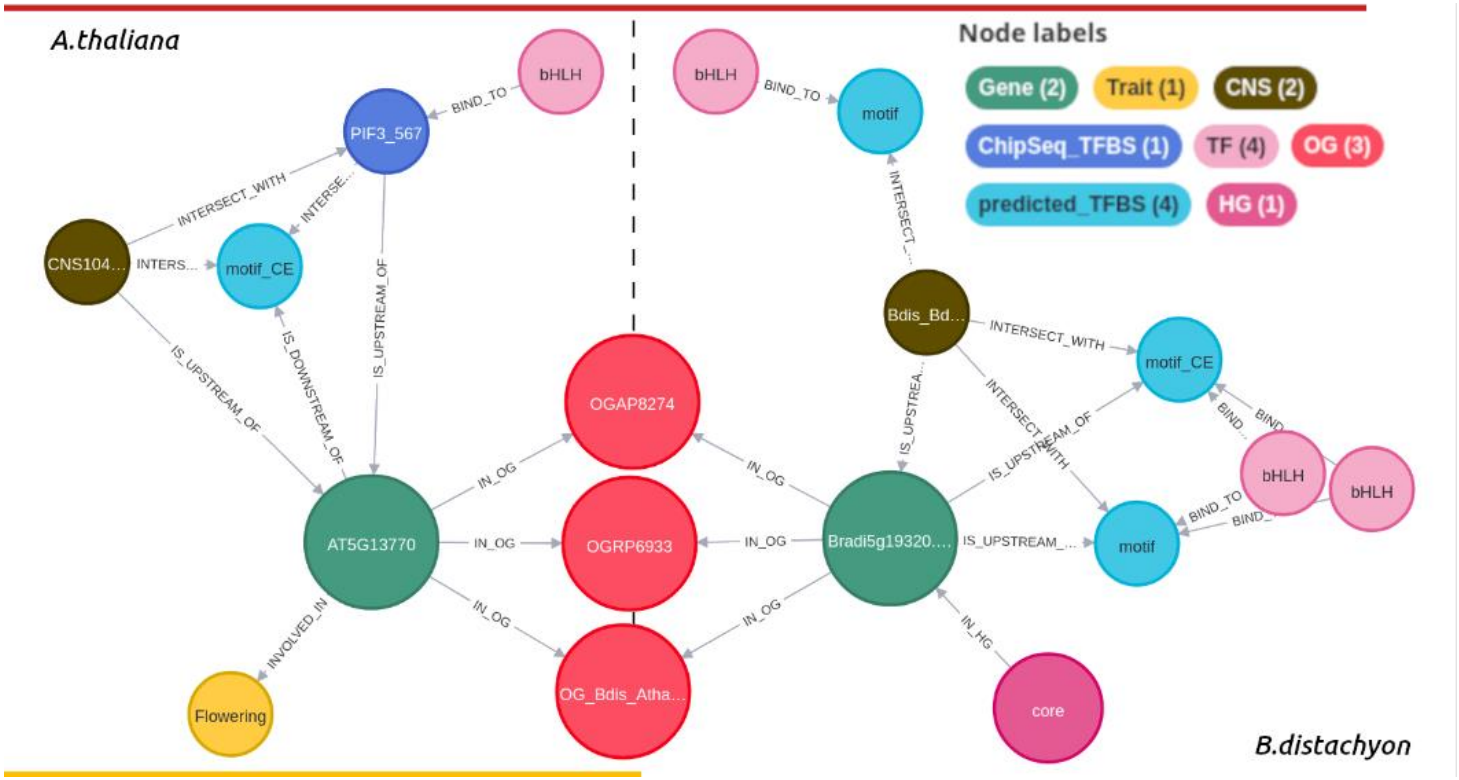
Conditions cumulées

Gènes d'*A.thaliana* impliqués dans la floraison
+ CNS en amont des gènes d'*A.thaliana*
+ CNS qui chevauche un TFBS validé par analyse ChIP-Seq
+ ChIP-Seq TFBS se liant avec un TF
+ Gène d'*A.thaliana* orthologue à un gène de *B.distachyon*
+ CNS en amont des gènes de *B.distachyon*
+CNS de *B.distachyon* chevauche un TFBS prédit se liant à un TF
+TF d'*A.thaliana* et de *B.distachyon* de la la même famille

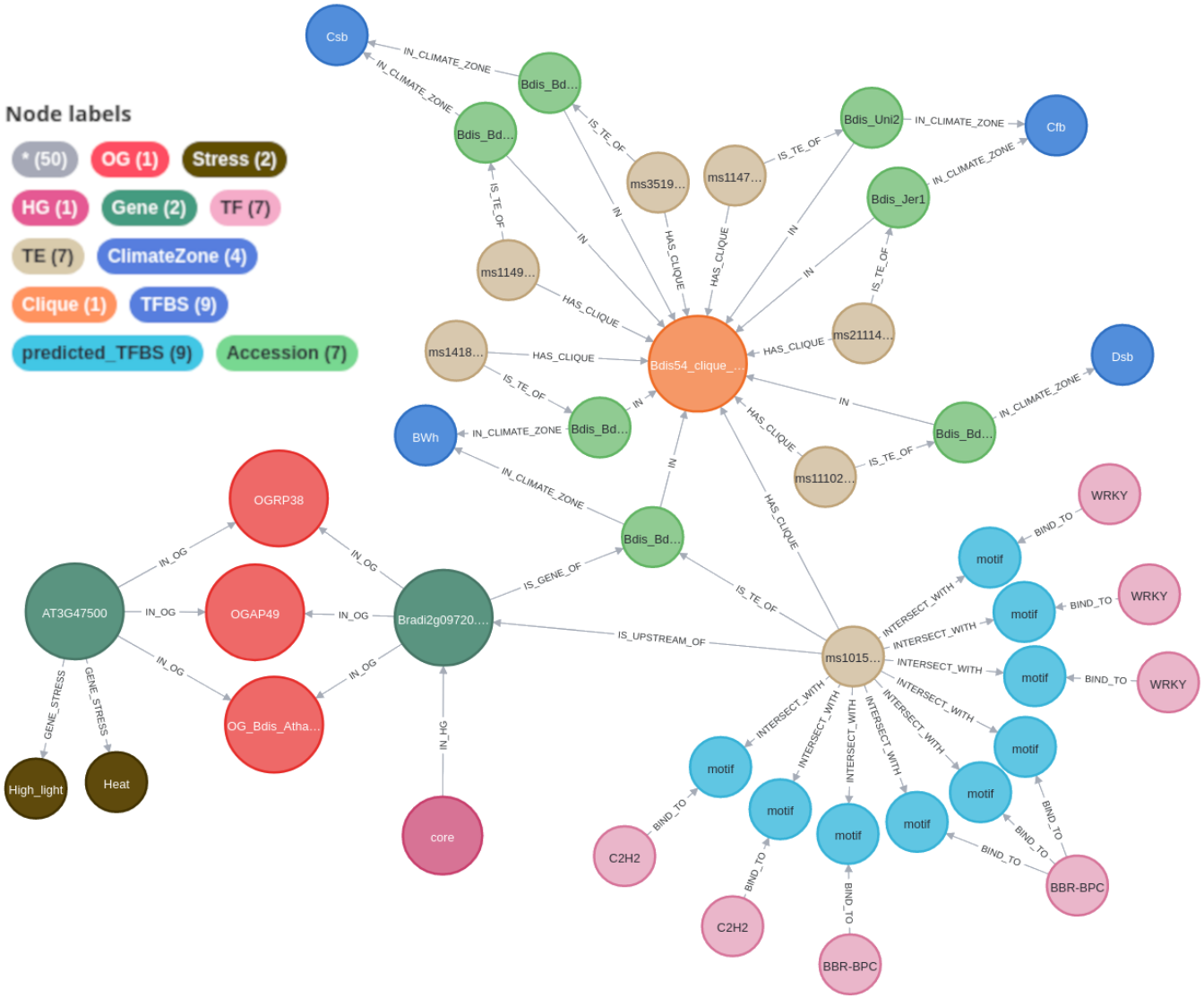
Puis affichage des TFBS prédits d'*A.thaliana* et des groupes d'homologie et d'orthologie de *B.distachyon*

```
MATCH a=(g1:Gene{specie:"Atha"})--(:Trait{TRAIT_NAME:"Flowering"}),
b=(g1)-[:IS_UPSTREAM_OF]-(cns1:CNS),
c=(cns1)-[:INTERSECT_WITH]-(tfbs1:ChipSeq_TFBS),
d=(tfbs1)--(tf1:TF),
e=(tfbs1)-[:INTERSECT_WITH]-(pTFBS1:predicted_TFBS)-[:INTERSECT_WITH]-(cns1),
f=(pTFBS1)-[:IS_DOWNSTREAM_OF]-(g1),
g=(g1)--(:OG{name_orthogroup:"OG_Bdis_Atha"})--(g2:Gene{specie:"Bdis"}),
h=(g2)-[:IS_UPSTREAM_OF]-(cns2:CNS),
i=(cns2)-[:INTERSECT_WITH]-(tfbs2:predicted_TFBS),
j=(tfbs2)-[:BIND_TO]-(tf2:TF), k=(g2)--(:HG), l=(g1)--(:OG)--(g2)
WHERE tf1.family=tf2.family
RETURN a,b,c,d,e,f,g,h,i,j,k,l
```

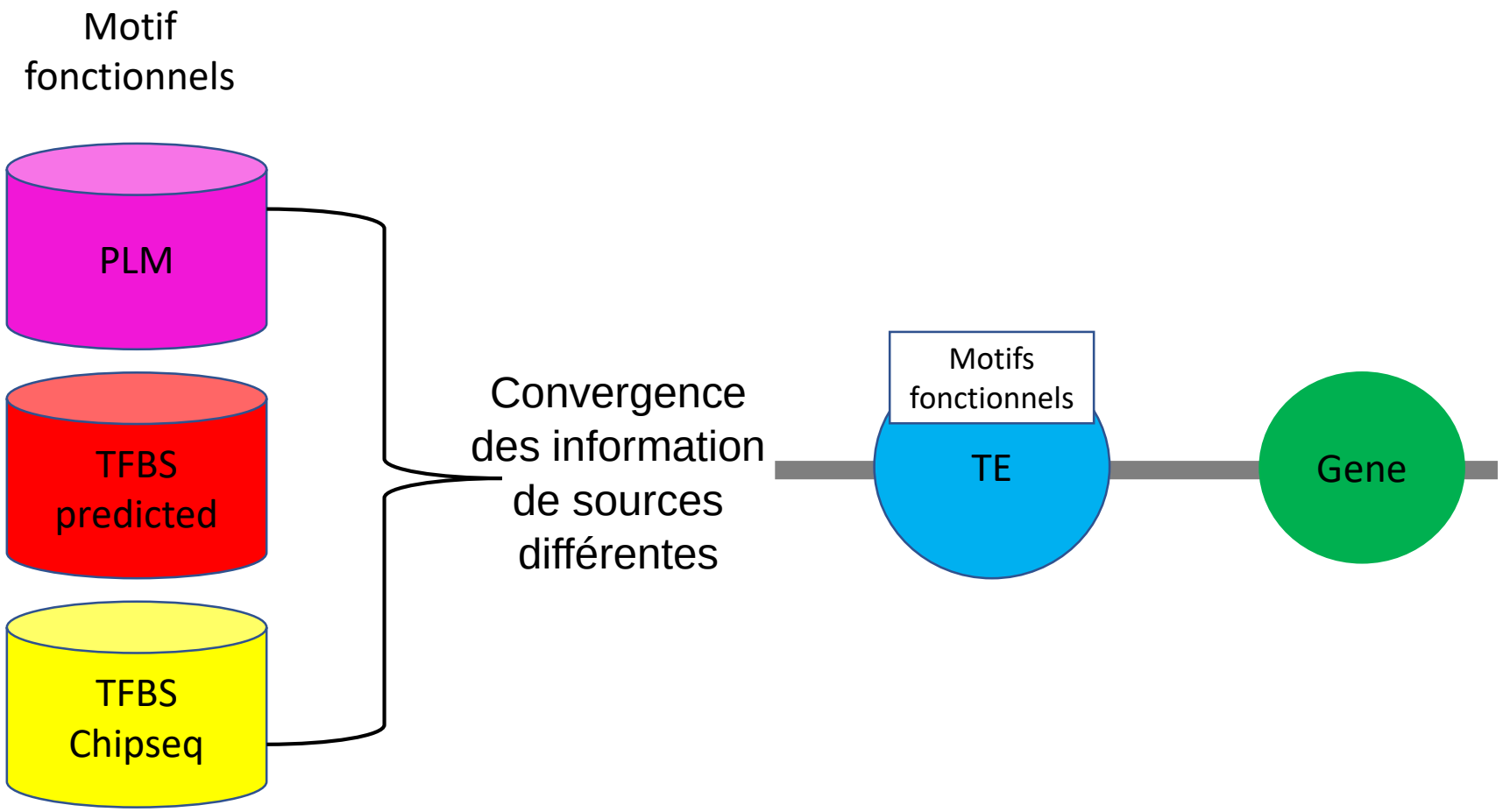
Use case


A.thaliana
B.distachyon
Node labels
Gene (2)
Trait (1)
CNS (2)
ChipSeq_TFBS (1)
TF (4)
OG (3)
predicted_TFBS (4)
HG (1)
bHLH
PIF3_567
BIND_TO
INTERSECT_WITH
CNS104...
motif_CE
IS_UPSTREAM_OF
IS_DOWNSTREAM_OF
AT5G13770
IN_OG
OGAP8274
OGRP6933
OG_Bdis_Atha...
INVOLVED_IN
Flowering
motif
Bdis_Bd...
Bradi5g19320....
IS_UPSTREAM_...
IN_HG
core

2e cas: Visualisation de TEs pouvant avoir un impact sur l'adaptation aux stress de chaleur et de luminosité



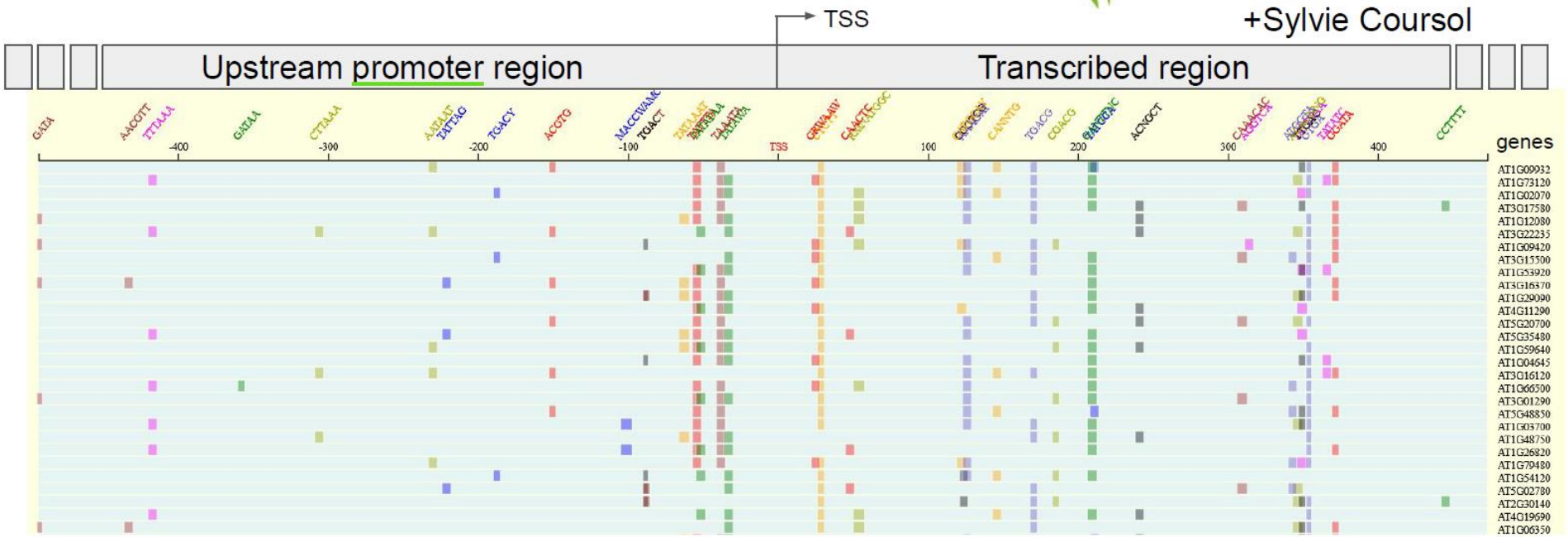
3e cas: Associer les familles de TE à des motifs fonctionnels



3e cas: Associer les familles de TE à des motifs fonctionnels

PLM (Preferentially Located Motif)

IPS2 Institute of Plant Sciences Paris - Saclay
GNet Marie-Laure Martin
Julien Rozière
+Sylvie Coursol

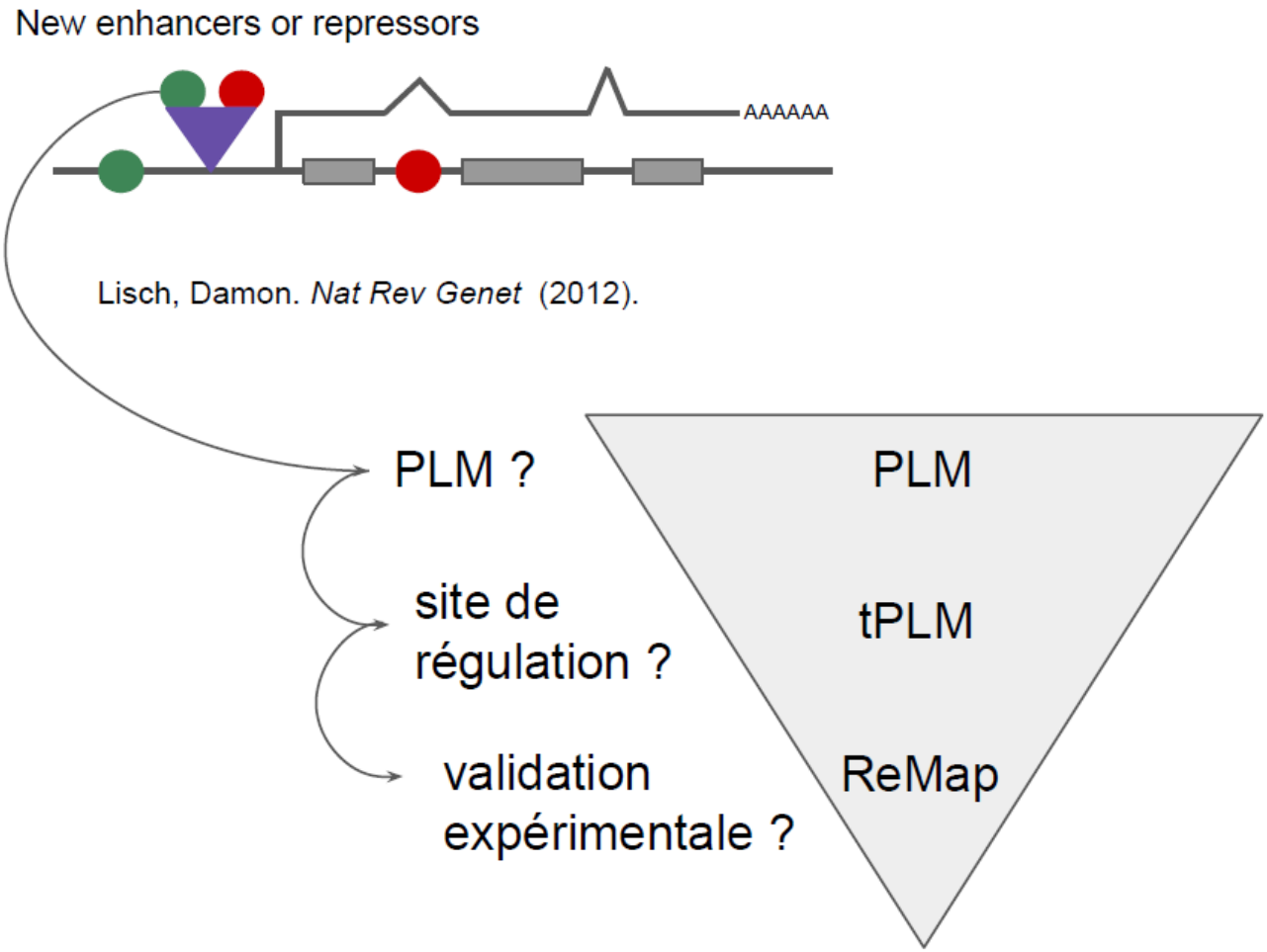


http://plmview.ips2.universite-paris-saclay.fr/

→ POTENTIELS MOTIFS CIS-REGULATEURS

Rozière, J., *et al.* (2022) Frontiers

Requêter les motifs associés aux TEs



Famille de TE associé à des motifs

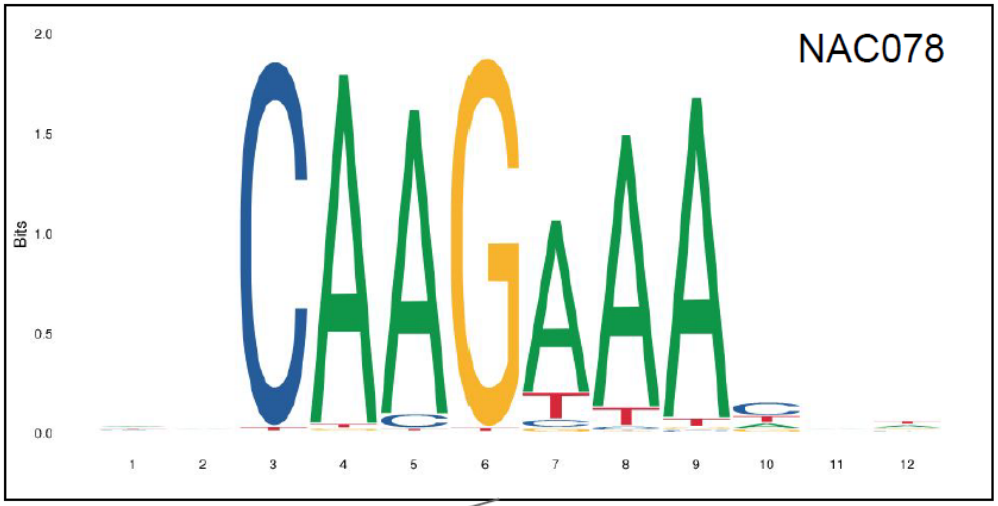
105 copies

ET
MCL30

portent

PLM
“TTCTTTC”

tPLM

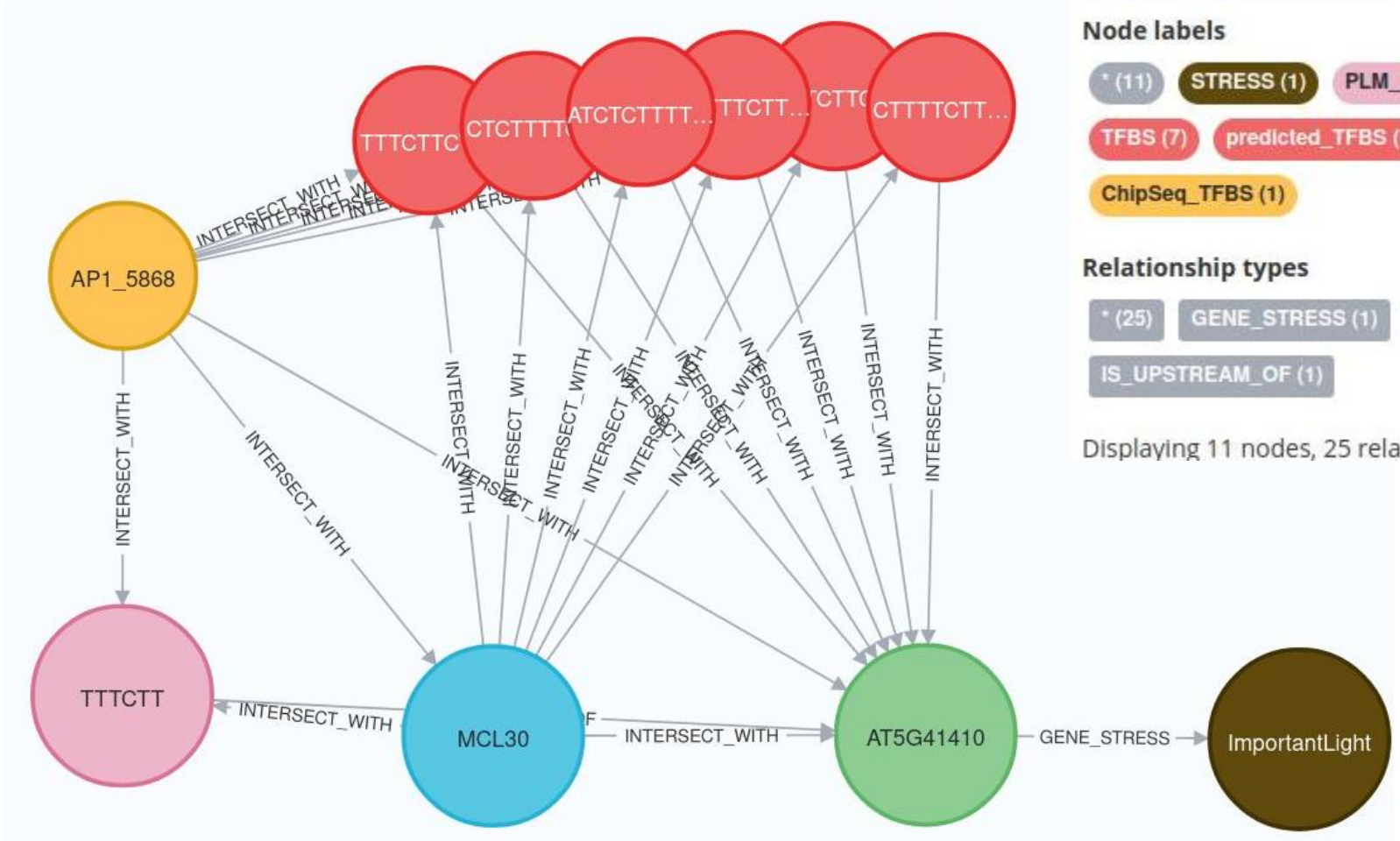


JASPAR 2022

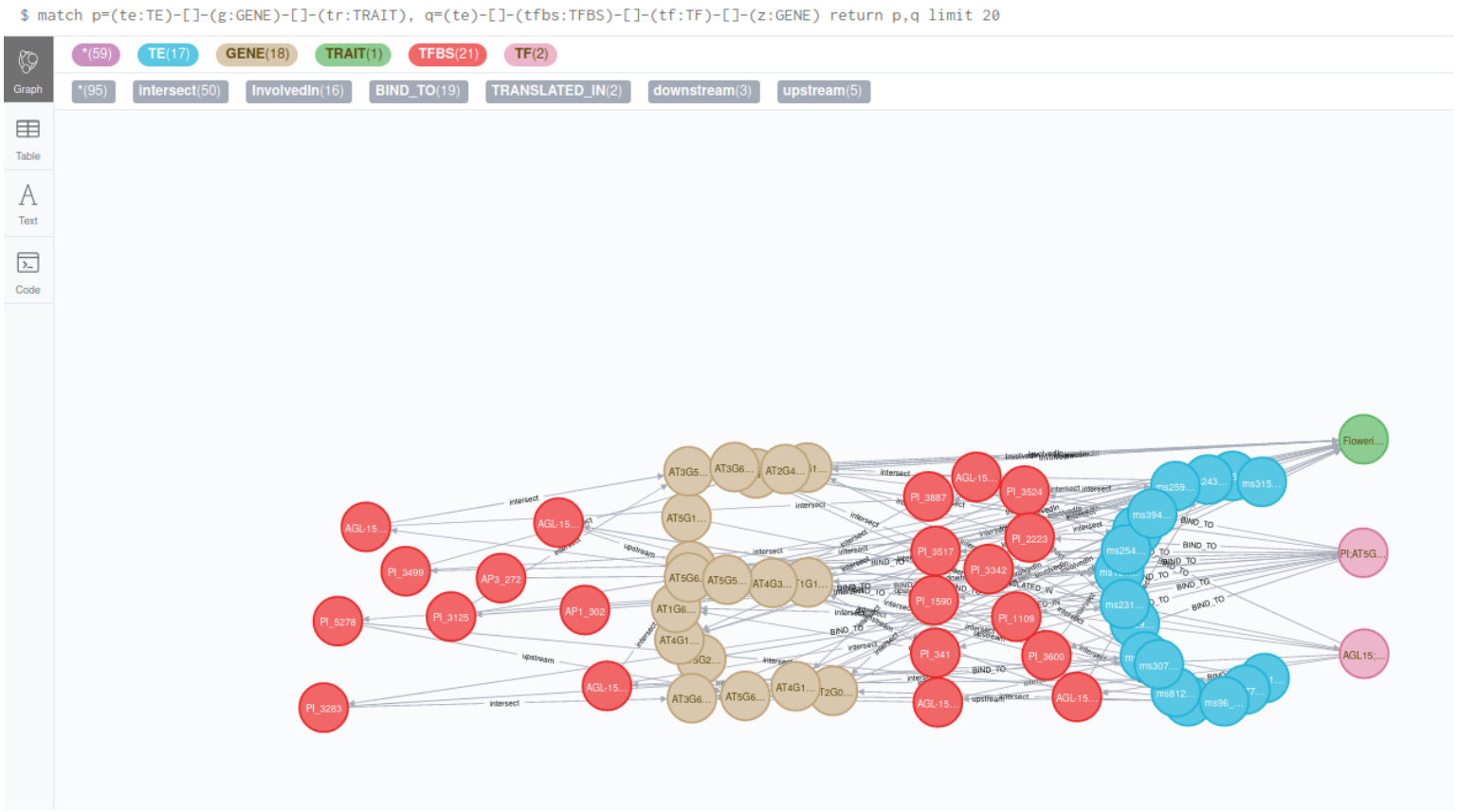
ReMap 2022

regulates flavonoid
biosynthesis under **high light**

Associer les évidences

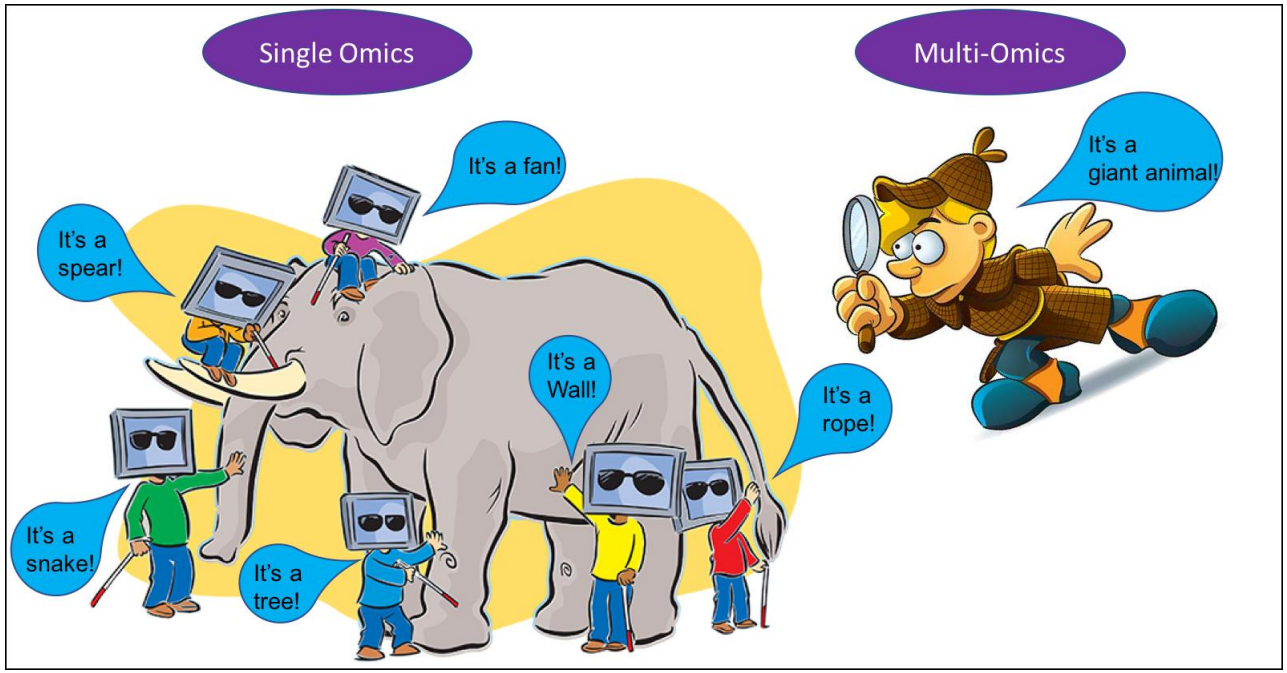


Questions ?


$ 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)

Bilan données intégrées

- Échelle **multi-omique** (annotations de gènes, d'éléments transposables, marqueurs génomiques/génétiques, fonctions, phénotypes/traits, etc.)
- Encore énormément de données à moissonner dans des articles



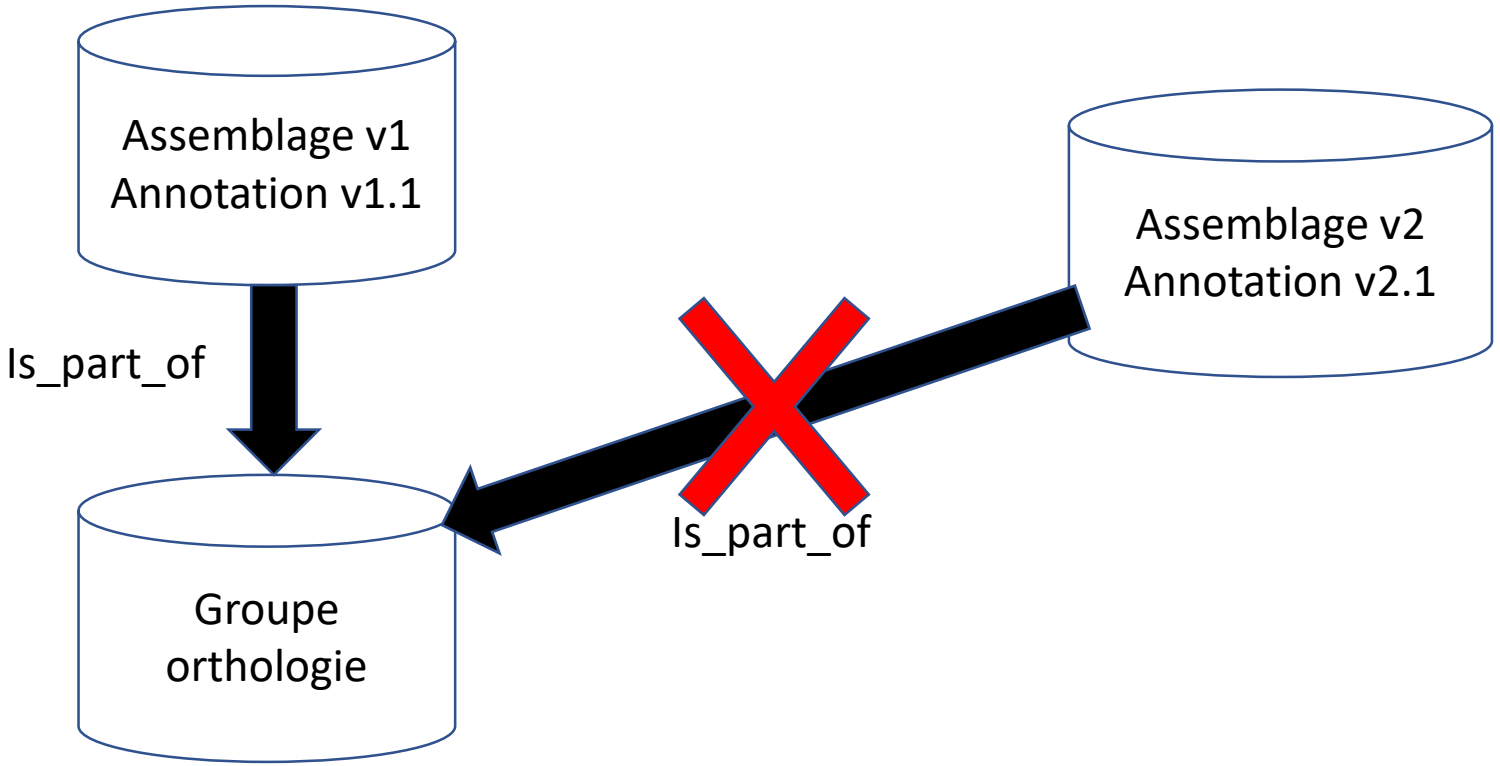
Quelques remarques

- Ontologies **utilisées** :
  - de référence (GO, SO)
  - ou spécifiques (PTO)
- Ontologies « **à réfléchir** » :
  - relation de distance entre entités génomiques
  - Description phénotypique : gènes impliqués dans des stresses
  - Nouveau concept (lié à la pangénomique comme clique)
- **Modélisation** très liée aux entités manipulées, mais aussi et surtout aux questions envisagées
- **Gestion de la provenance** perfectible. Pistes :
  - PAV (Provenance, Authoring and Versioning) : https://pav-ontology.github.io/pav/ (plutôt léger)
  - Provenance DCTerms : https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/provenance/
  - PROV-O : https://www.w3.org/TR/2013/REC-prov-o-20130430/ (overkill)
  - DublinCore/PROV : https://www.w3.org/TR/prov-dc/

Pré-compilation (2)

Matrice de conversion entre version d'annotation

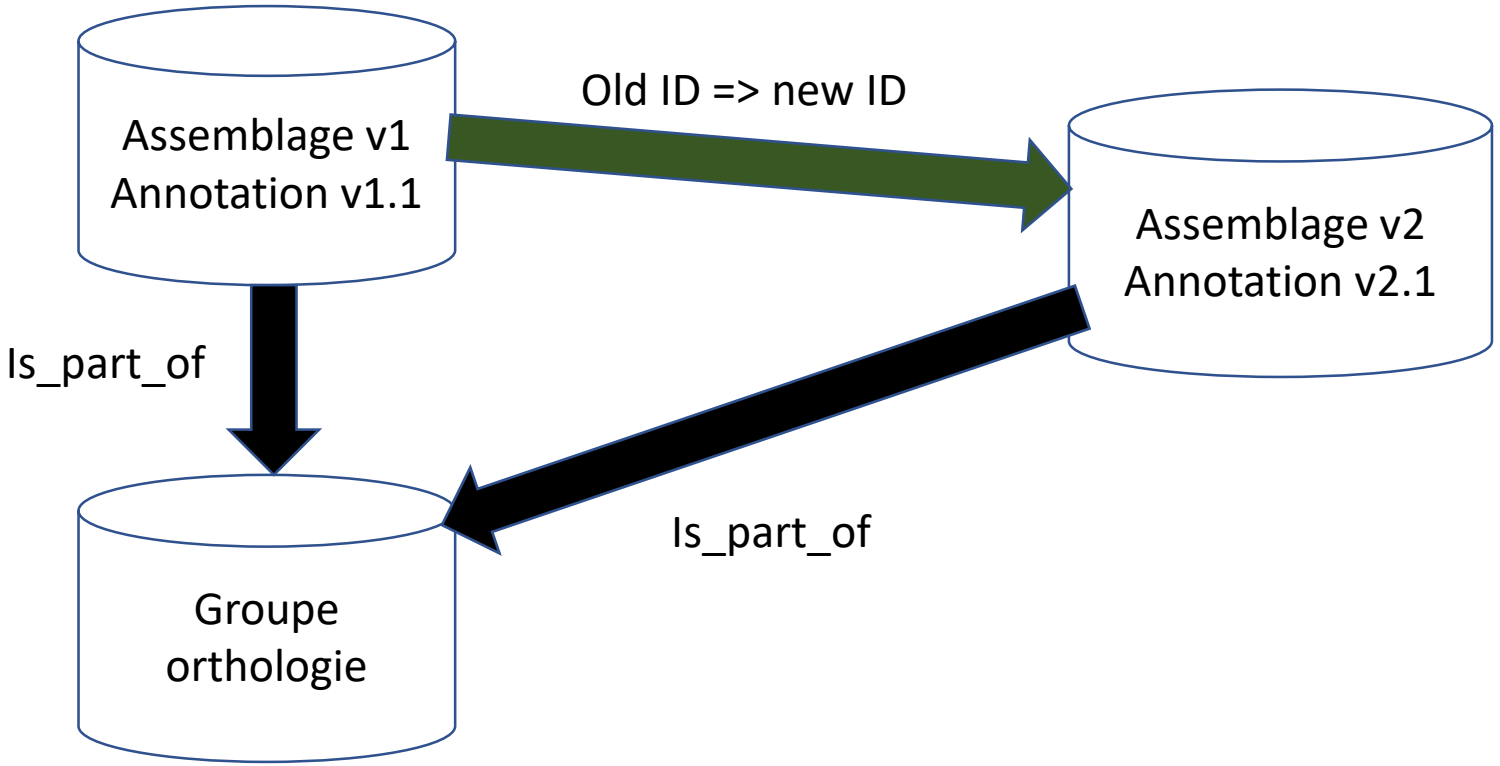
Certaines données ne sont disponibles que pour certaines versions d'assemblage
Nécessité de lier les éléments génomiques sur la même version



Pré-compilation (2)

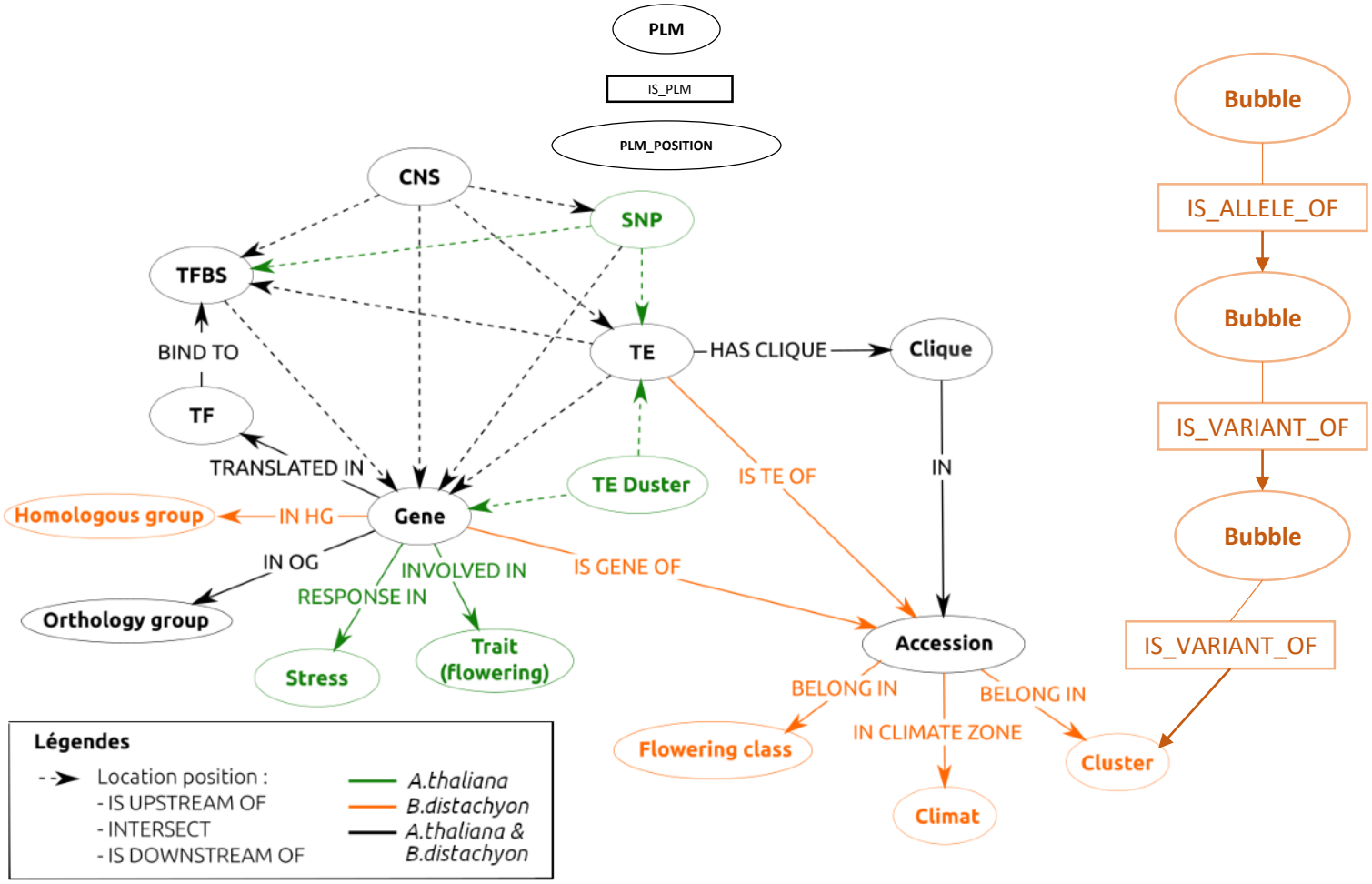
Matrice de conversion entre version d'annotation

Certaines données ne sont disponibles que pour certaines versions d'assemblage
Nécessité de lier les éléments génomiques sur la même version



Cas d'utilisation

Brachypodium & Arabidopsis



Recherche du motif conservé d'*A.thaliana* chez *B.distachyon*

- Recherche du motif_CE TFBS d'*A.thaliana* en amont du gène Bradi5g19320 (v3.2)
- ⇒ séquence similaire sur *B.distachyon* entre 22 490 239 - 22 490 251 (- 4 003 avant le gène)
- Recherche de CNS correspondant à ces positions : Bdis_Bd21_CNS_409825 (Bd5 : 22 490 141 - 22 490 306)
- Pas de TFBS prédit sur ces coordonnées
- Récupération et alignement du CNS de *B.distachyon* sur *A.thaliana*

- Pas de lien spécifique entre la classification de climat et la présence de la clique de TE observée

- Beaucoup de cas à explorer même si pas de corrélations observées dans le papier de Gordon

Recherches complémentaires

Fonctions du TF bHLH (Guo *et al.*, 2021)

- Réponse au stress abiotique
- Régulation de la floraison

Fonctions du gène (Zhang *et al.*, 2020 ; Saha *et al.*, 2007)

- Protéine de la superfamille des Pentatricopeptide Repeat (PPR-like)
- Adaptation aux stress et processus développementaux (pollen)
- Famille de protéines retrouvée dans de nombreux organismes