

FsoV

INRAE

ao

BASF  
We create chemistry

FLORIMOND  
DESPREZ

KWS

Lidea

Limagrain  
de la terre à la vie

RA-G-T

SECOBRA  
RECHERCHES

syngenta

UCA

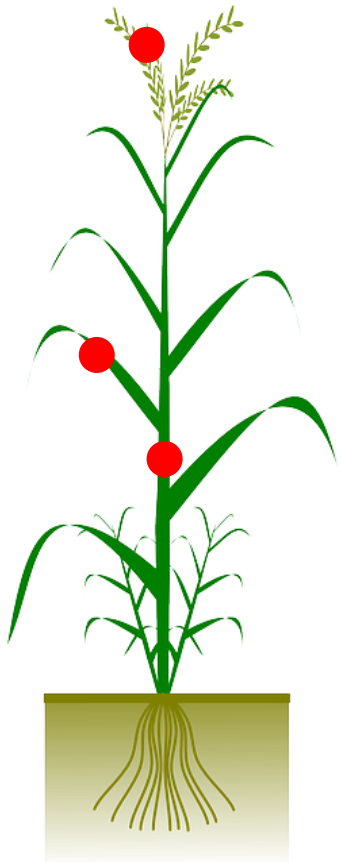
MÉSOCENTRE  
UNIVERSITÉ  
Clermont Auvergne

# DéBAT - Analyse de Diversité d'un Panel de Pré-breeding de Blé tendre par une Approche Transcriptomique

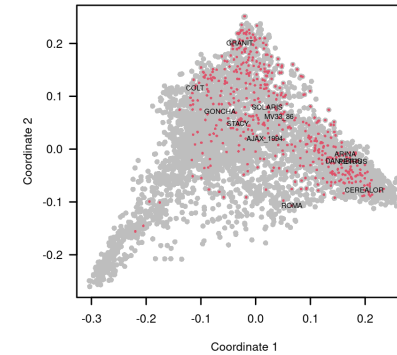
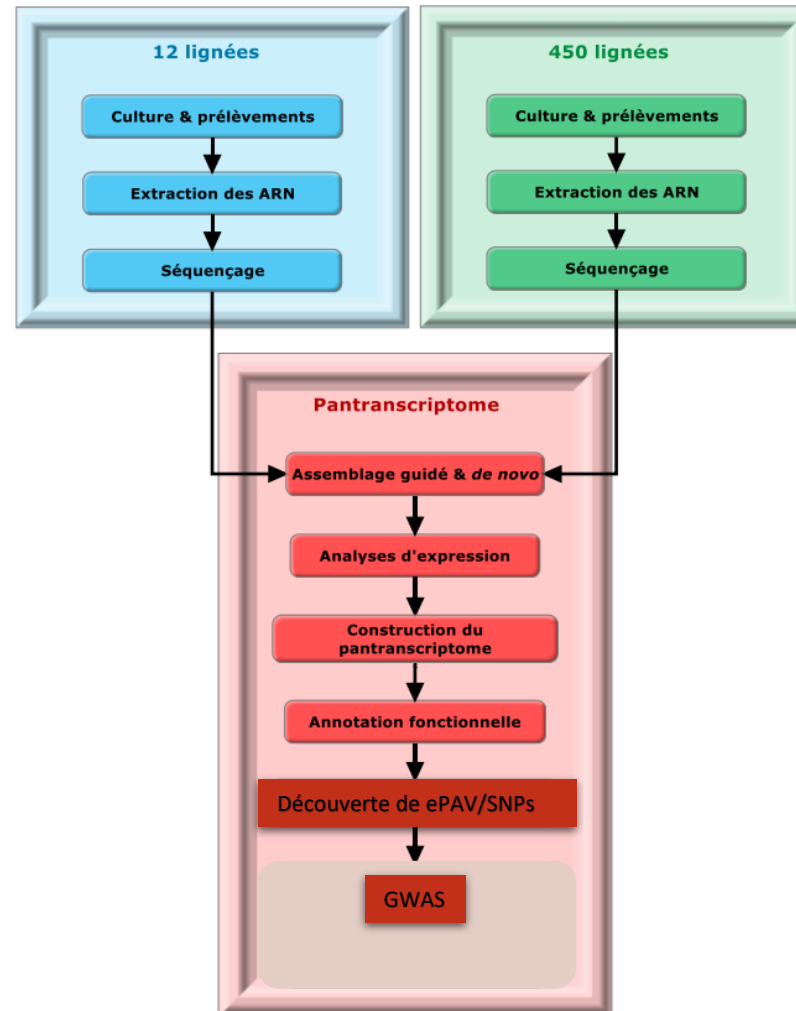
RIMBERT Hélène / BOUCHET Sophie

# Catalogue de transcrits représentatif de la diversité

## Le projet

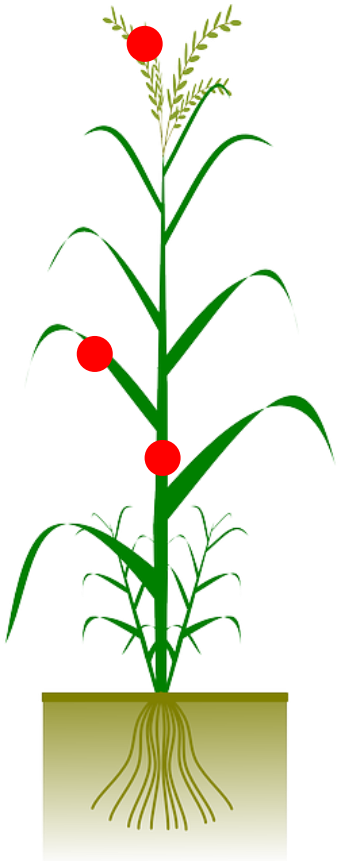


PETRUS  
MV33 86  
CEREALOR  
ROMA ARINAGRANIT  
DANUBIUS  
COLT AJAX  
SOLARIS GONCHA  
STACY



# Catalogue de transcrits représentatif de la diversité

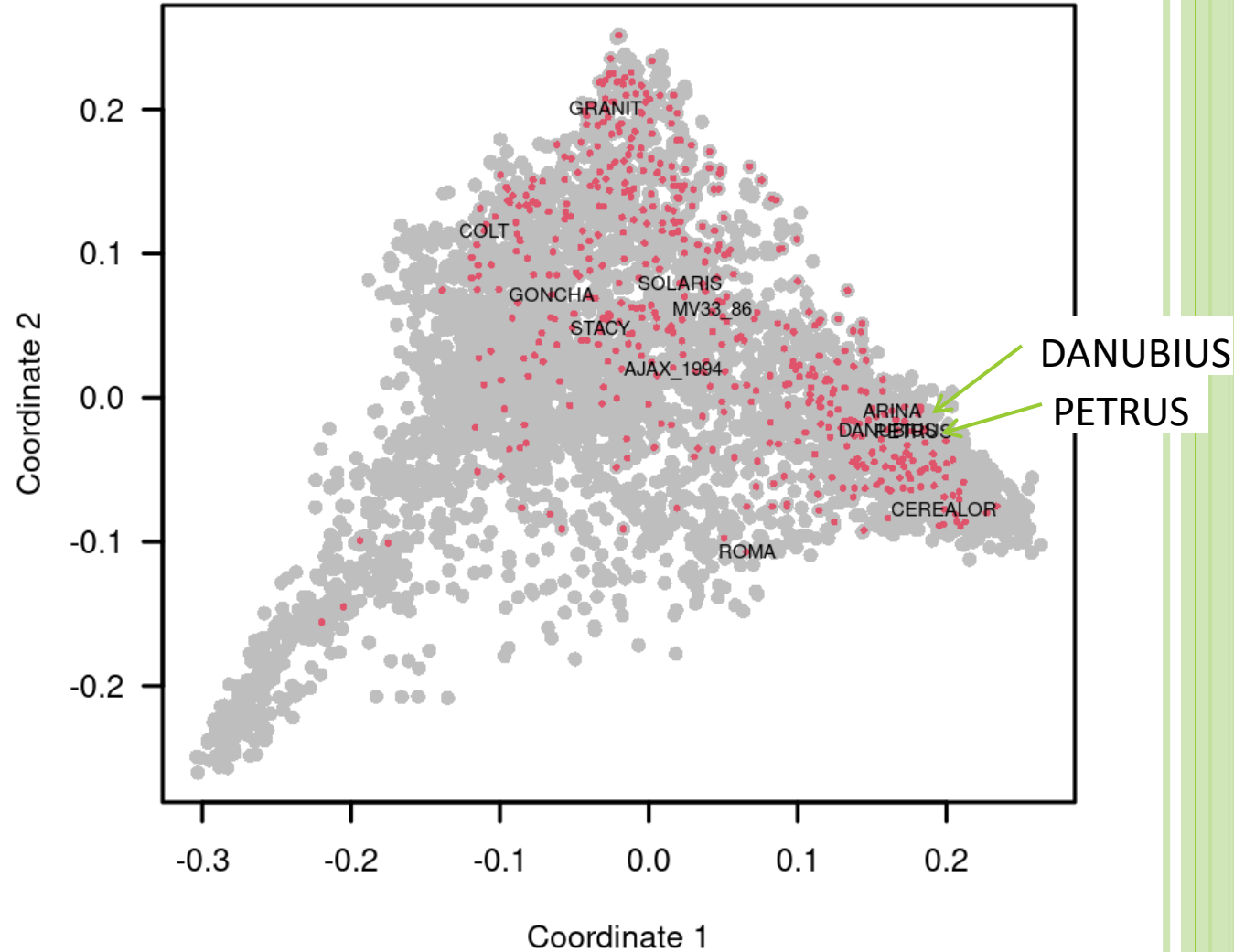
## Matériel végétal



**Volet 1** : 12 accessions

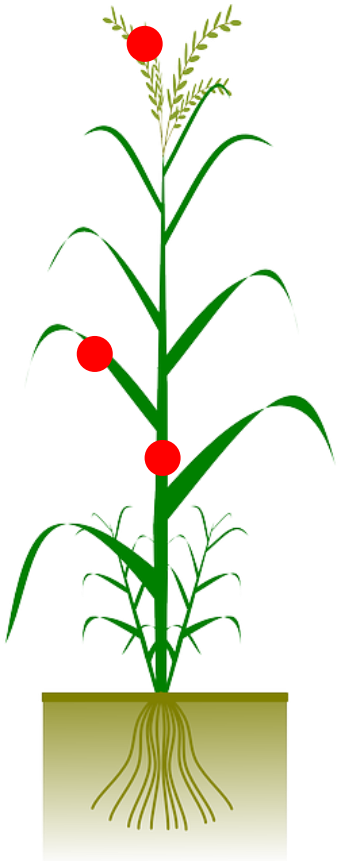
PETRUS  
MV33\_86  
CEREALOR  
ROMA ARINAGRANIT  
DANUBIUS  
COLT AJAX  
SOLARIS GONCHA  
STACY

**Volet 2** : 450 accessions  
->pas d'Asiatiques



# Catalogue de transcrits représentatif de la diversité

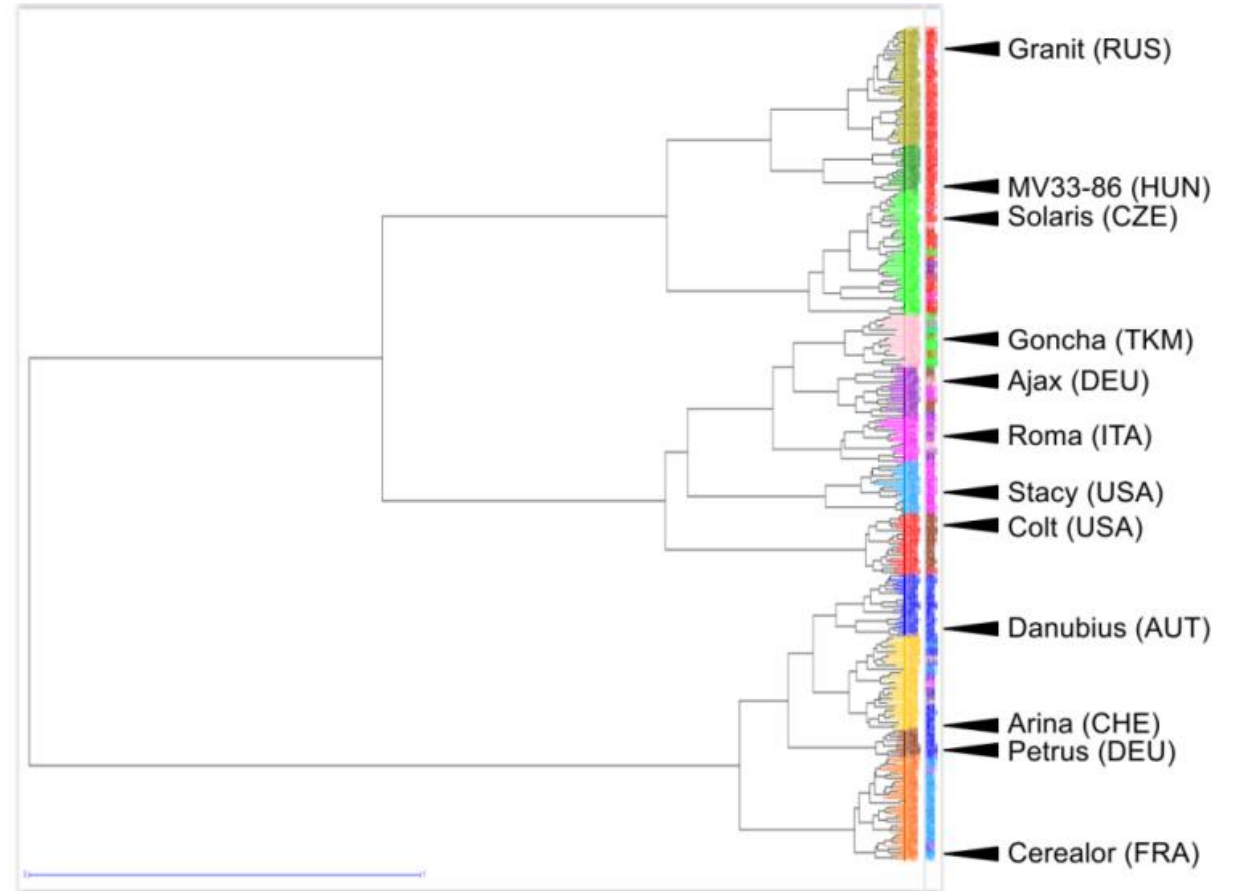
## Matériel végétal



### Volet 1 : 12 accessions

PETRUS  
MV33 86  
CEREALOR  
ROMA ARINAGRANIT  
DANUBIUS  
COLT AJAX  
SOLARIS GONCHA  
STACY

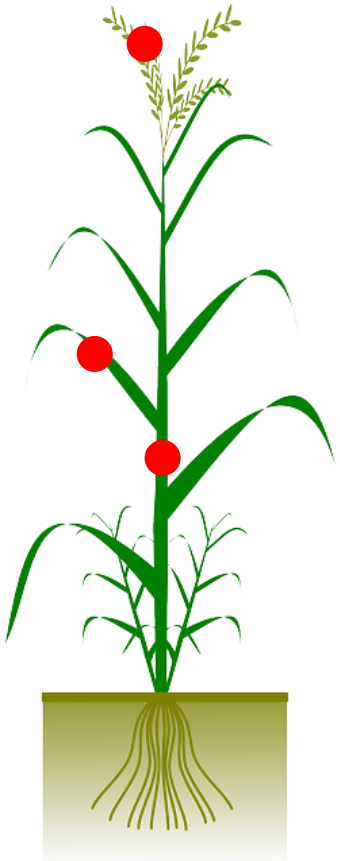
### Volet 2 : 450 accessions ->pas d'Asiatiques



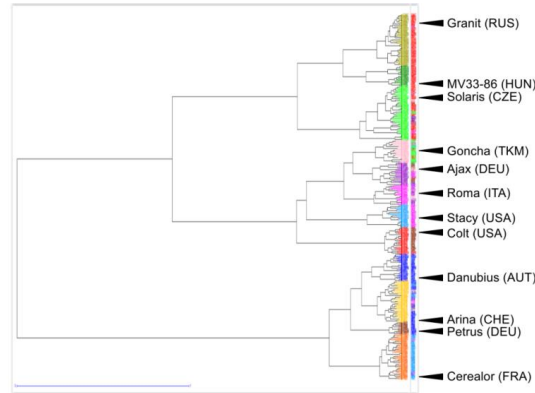
*Dendrogramme de Ward montrant les relations phylogénétiques entre les 450 lignées du panel BWP3.  
D'après Balfourier et al (Balfourier et al. 2019)*

# Catalogue de transcrits représentatif de la diversité

## Séquençage



### Volet 1 : 12 accessions



Séquençage forte profondeur:

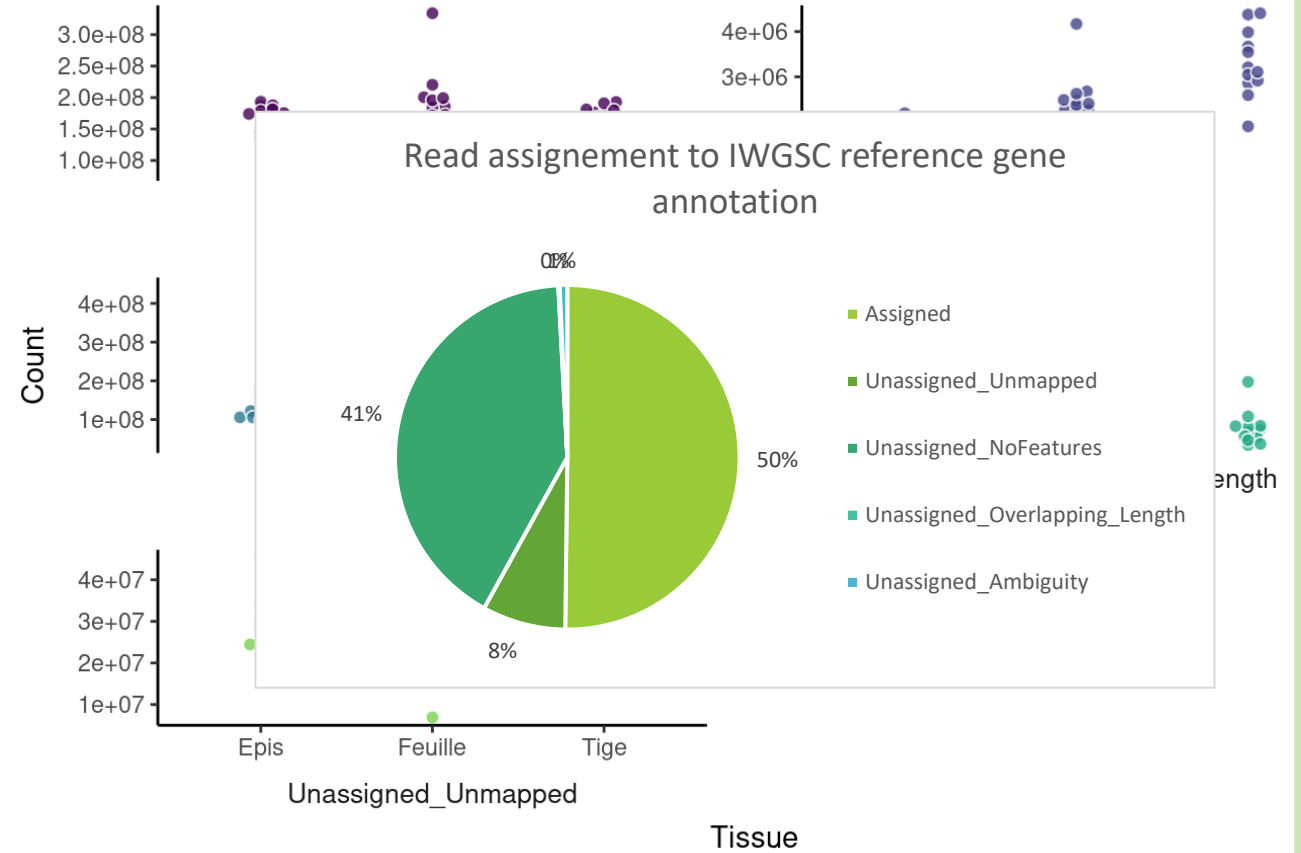
~ **100M** lectures par échantillon

**75% uniques** sur le génome (STAR)

**51% assignées à des gènes**

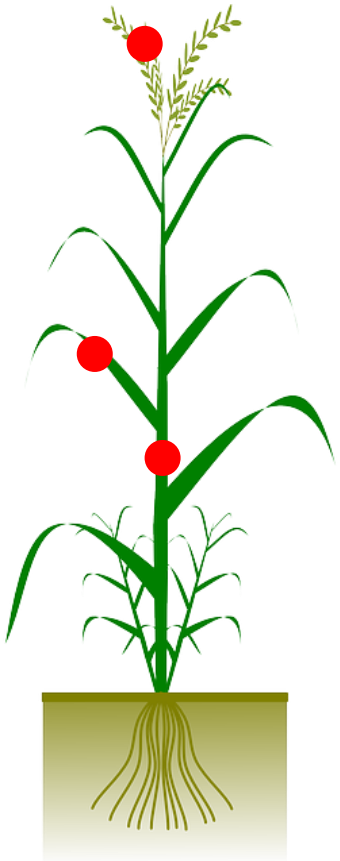
Matrice de comptage (FeatureCount)

### Read Assignment to IWGSC reference gene annotation featureCount on sub-samples sequencing data (50% of mapped reads)

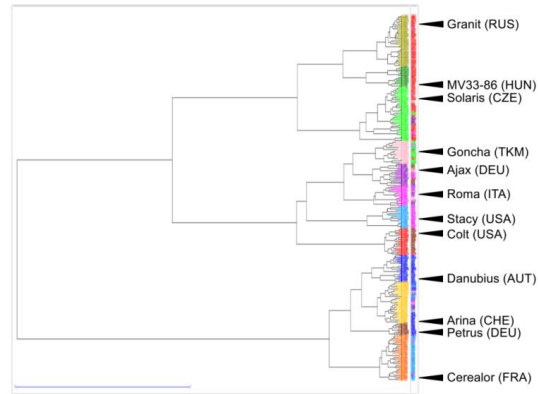


# Catalogue de transcrits représentatif de la diversité

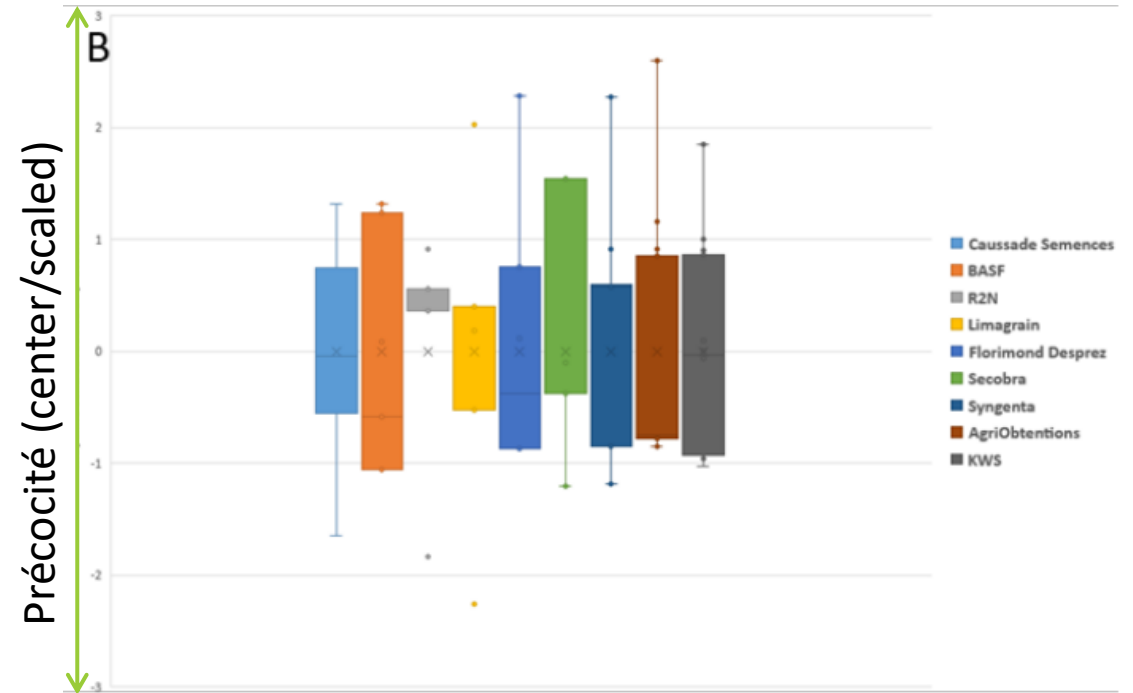
## Séquençage



### Volet 2 : 450 accessions

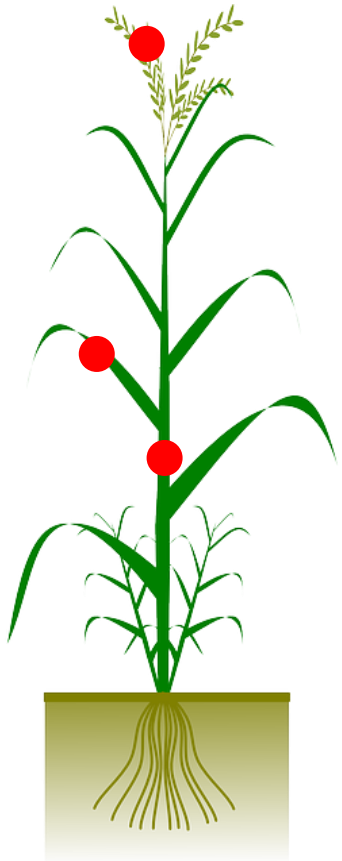


Séquençage faible profondeur:  
~ **50M** lectures par échantillon

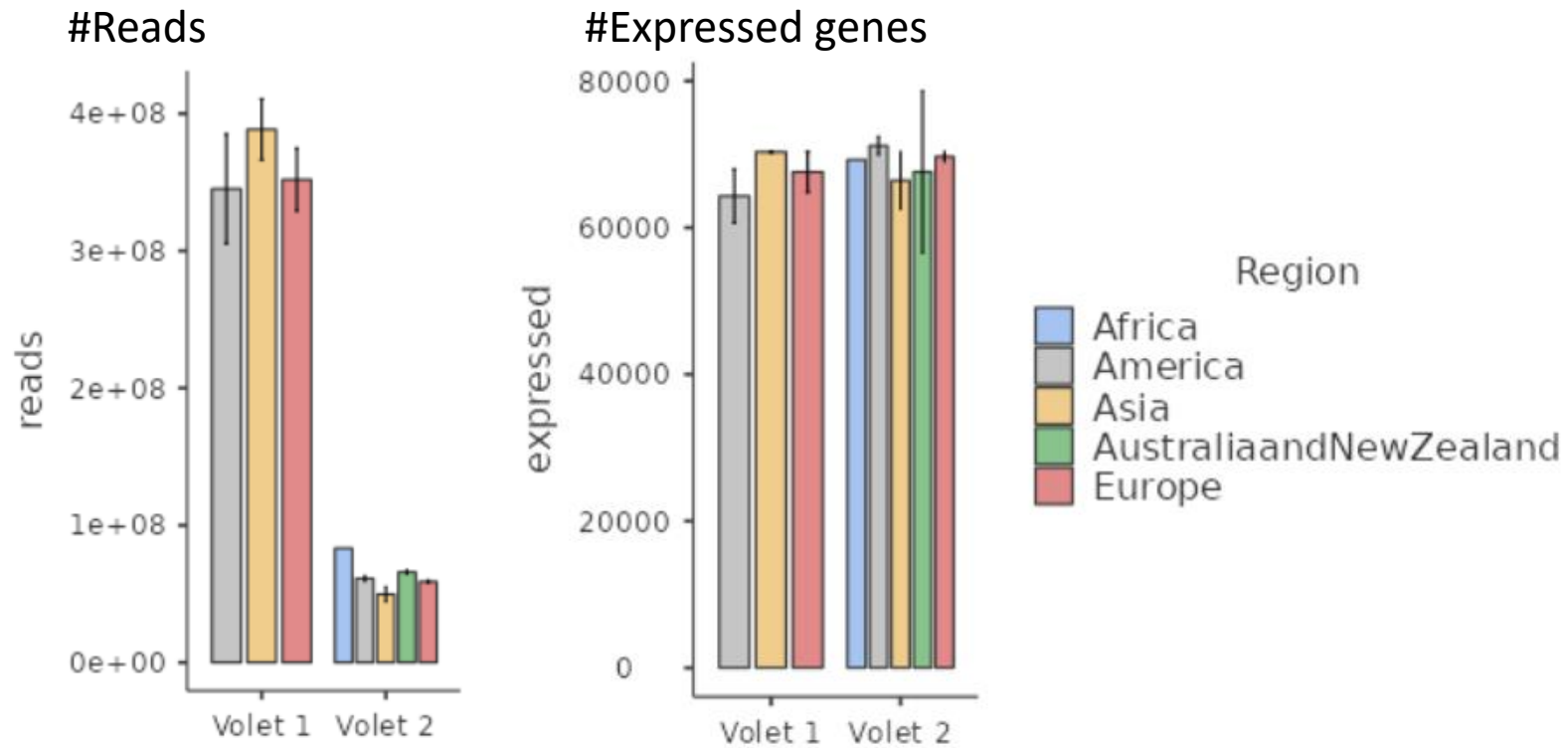


# Catalogue de transcrits représentatif de la diversité

## Séquençage

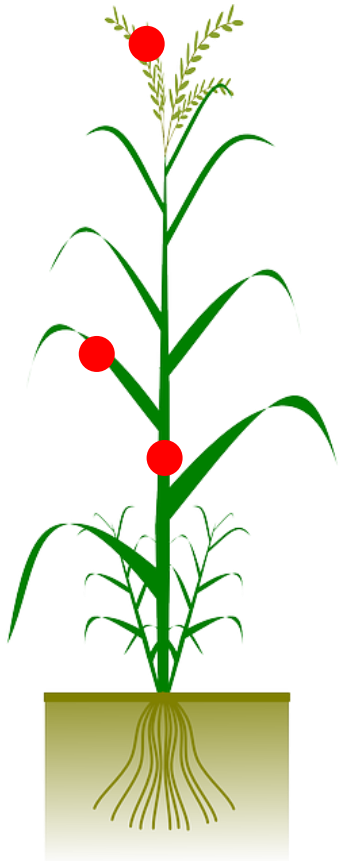


### Volet 1 & 2

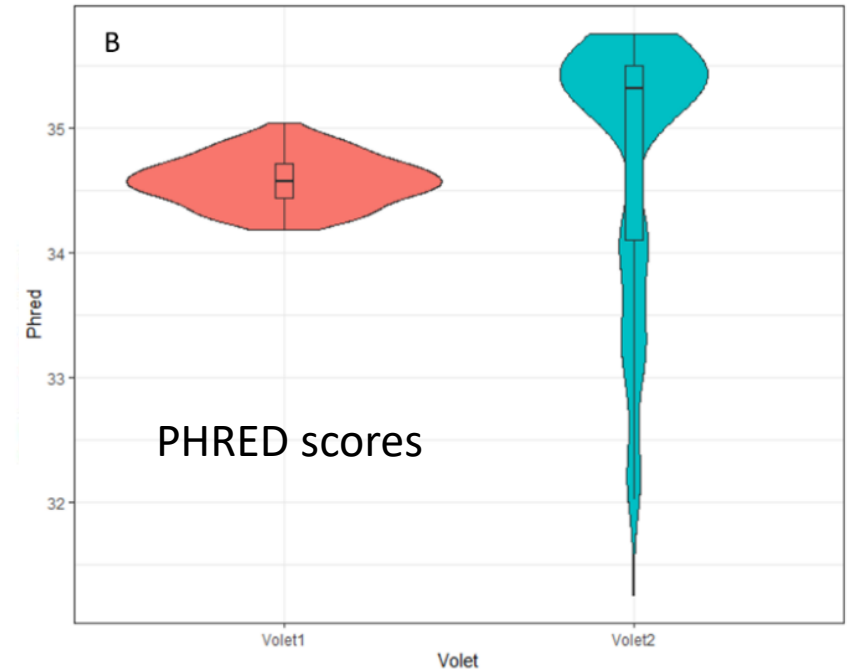
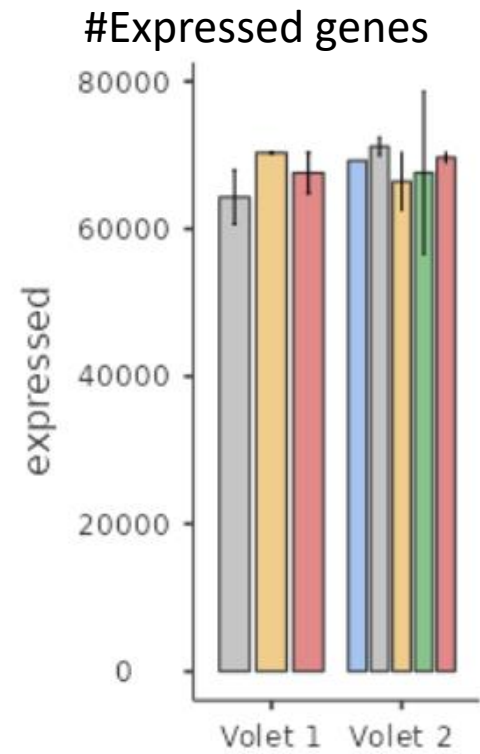
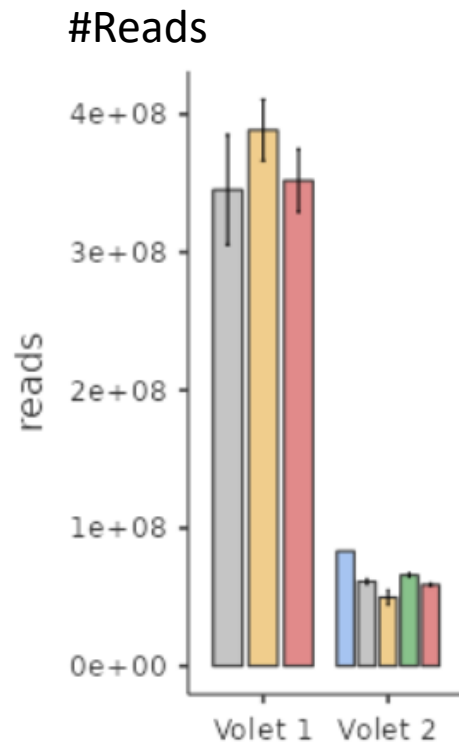


# Catalogue de transcrits représentatif de la diversité

## Séquençage



### Volet 1 & 2



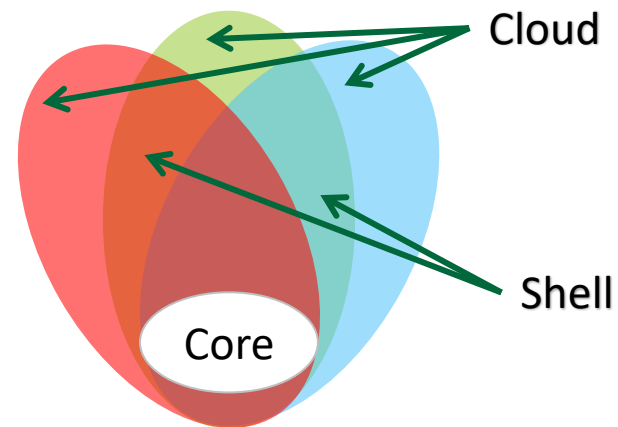


# Pan-transcriptome

Fraction spécifique : partagé par **moins de 10%** des accessions : **“Cloud”**

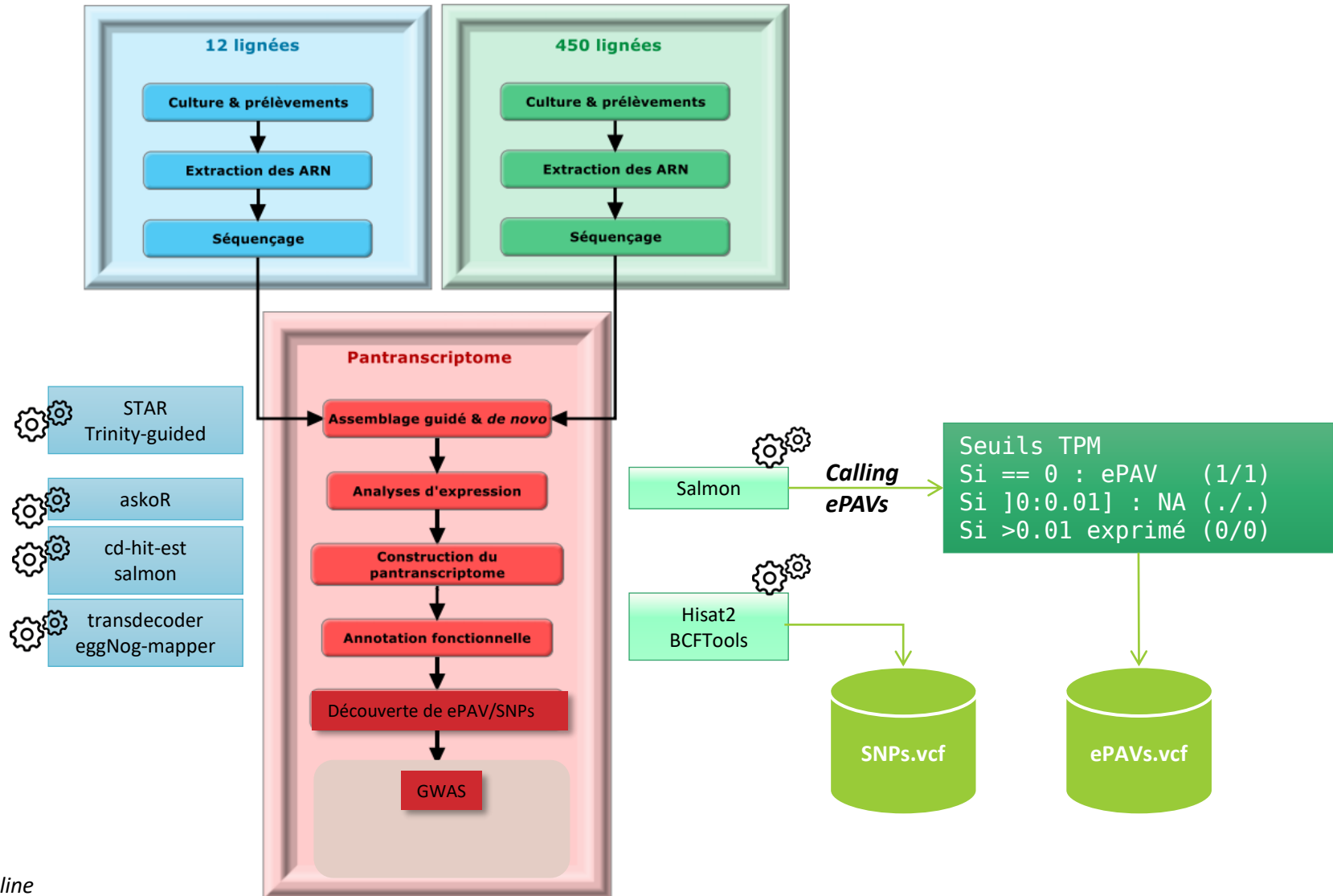
Accessoire : de **10% à 90%** des accessions : **“Shell”**

**“Core”** : plus de **90%** des accessions



# Pan-transcriptome

Bio-informatic pipeline

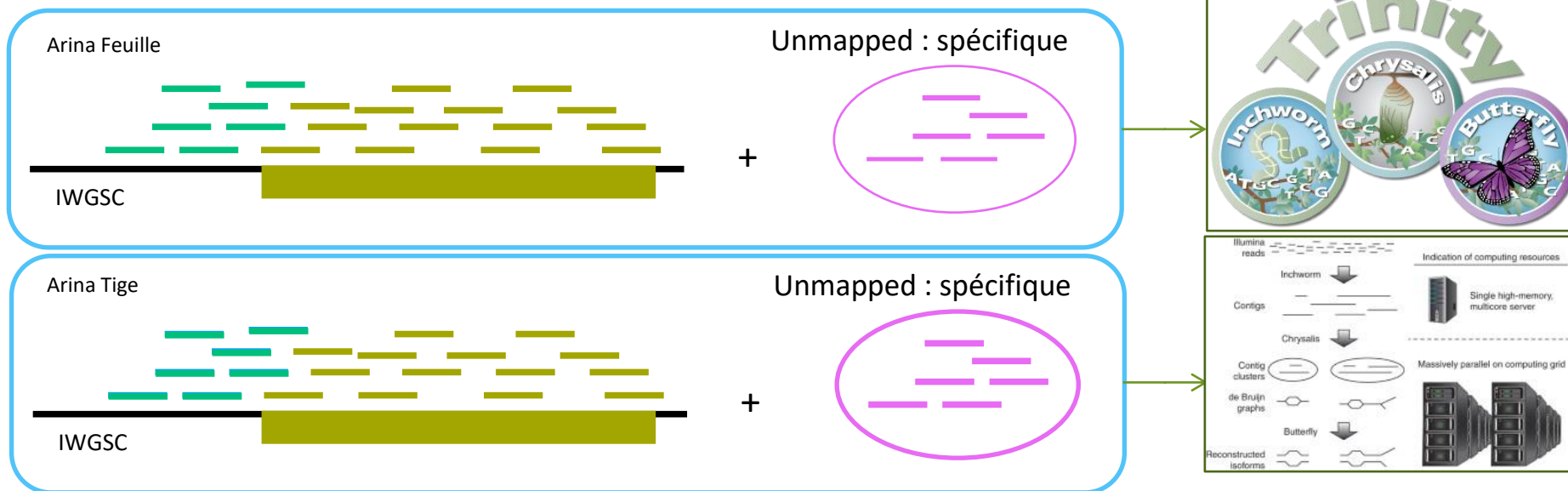


<https://forgemia.inra.fr/fsov-debat/debat-bashpipeline>  
<https://github.com/HeleneRimbert/oscar-triannot>

# Pan-transcriptome - Construction volet 1

PETRUS  
MV33 86  
CEREALOR  
ROMA ARINAGRANIT  
DANUBIUS  
COLT AJAX  
SOLARIS GONCHA  
STACY

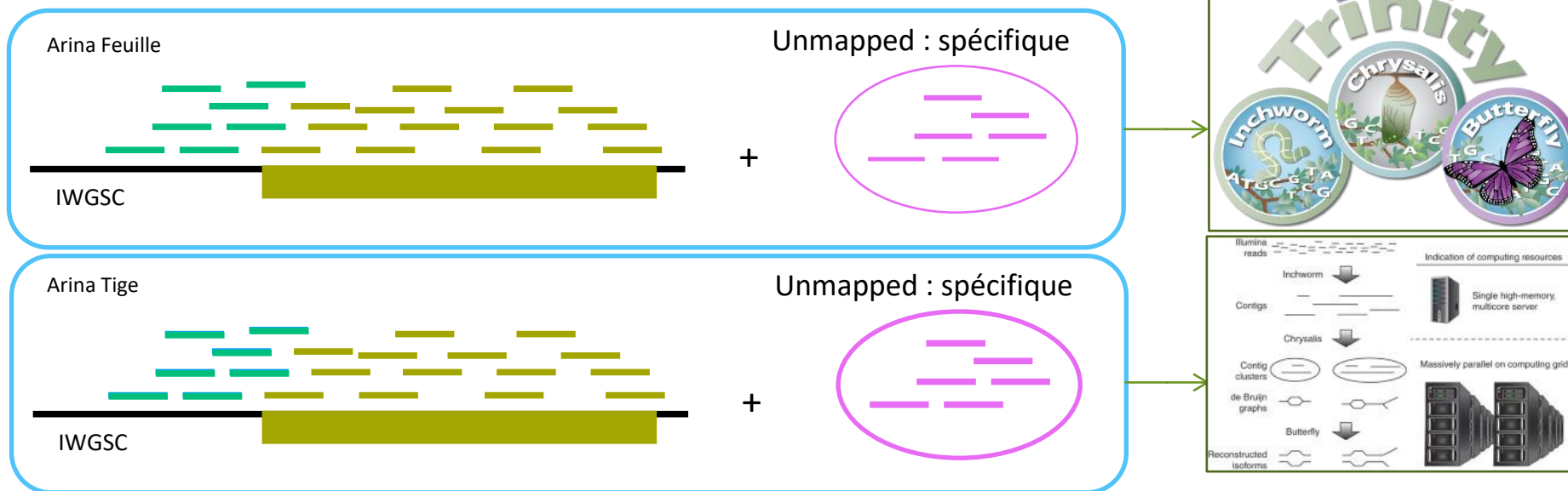
## 36 Assemblages hybride : STAR + Trinity (genome-guided)



# Pan-transcriptome - Construction volet 1

PETRUS  
MV33 86  
CEREALOR  
ROMA ARINAGRANIT  
DANUBIUS  
COLT AJAX  
SOLARIS GONCHA  
STACY

## 36 Assemblages hybride : STAR + Trinity (genome-guided)



...  
x 36  
= 6.7M transcrits/isoformes  
~ 200k / condition

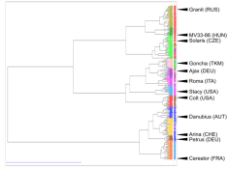
clustering / isoforme  
CD-HIT-EST

3.8M transcrits/isoformes

Re-mapping lectures  
Ancrage CS  
+ Annotation CS v2.1

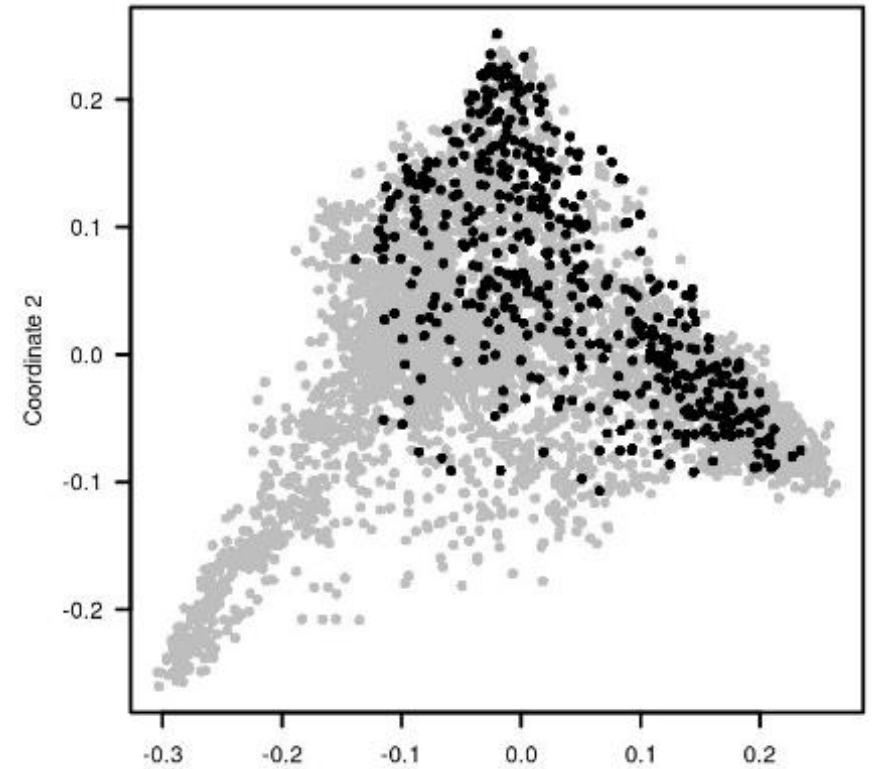
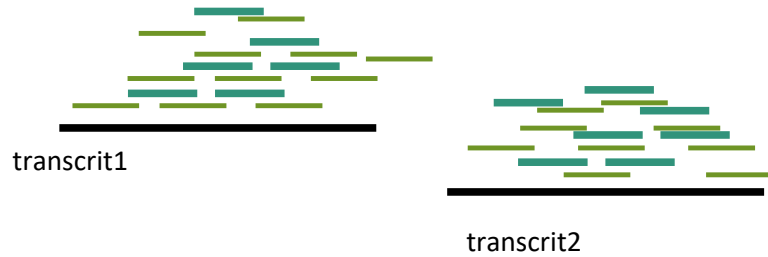
**400k pan-isoformes**

# Pan-transcriptome - Caractérisation volet 2

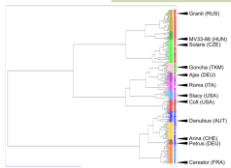


450 RNA-seq : ~ 50M lectures x 450

Quantification : Salmon

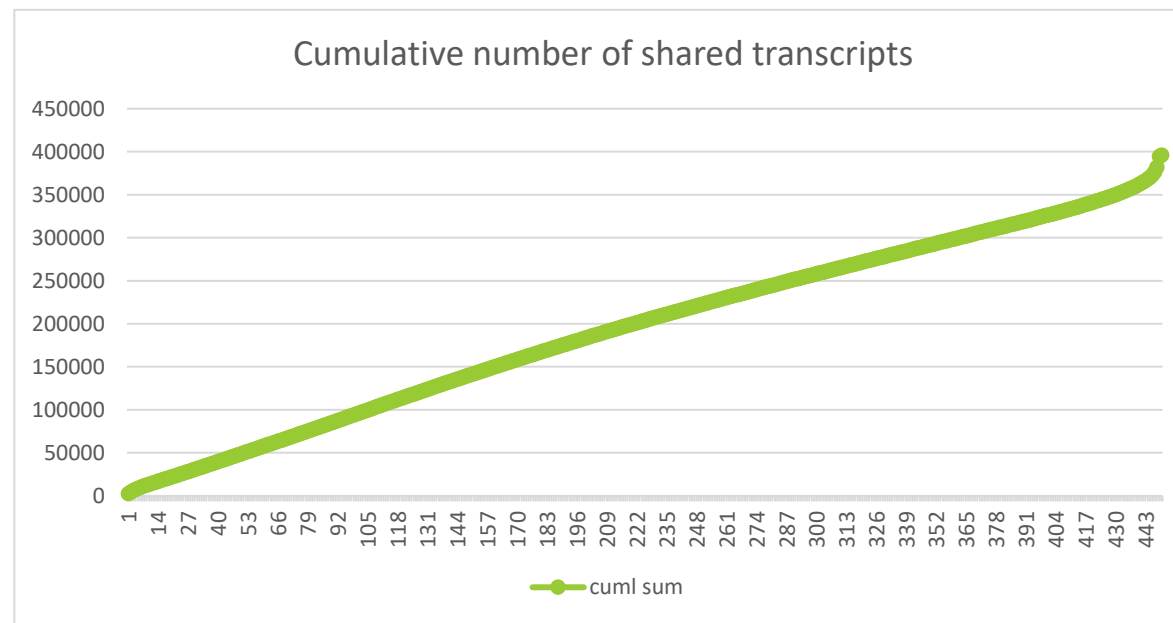
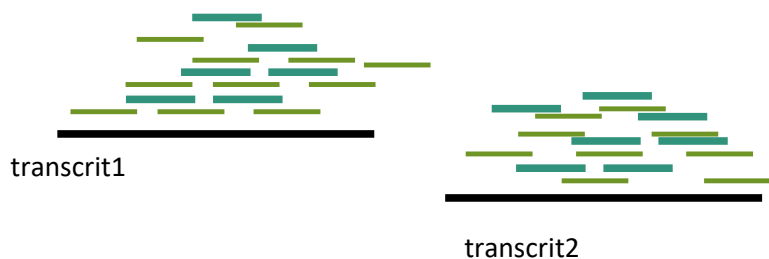


# Pan-transcriptome - Caractérisation volet 2



450 RNA-seq : ~ 50M lectures x 450

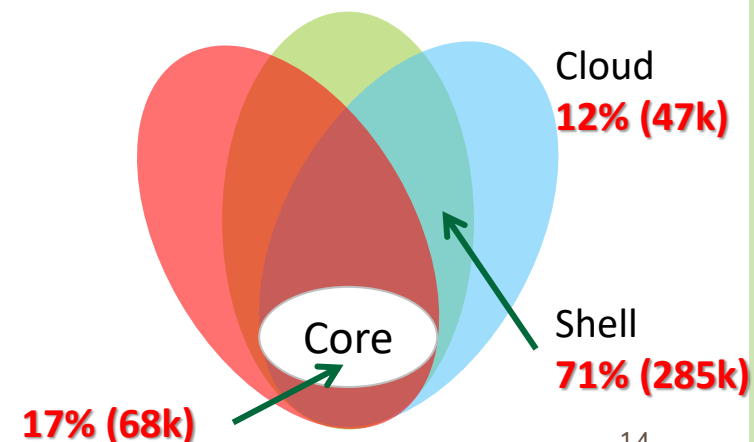
Quantification : Salmon



nombre croissant d'accessions qui partagent une isoforme

$\geq 0.01$  TPM

- 186k transcrits exprimés en moyenne par accession
- **396 221** transcrits exprimés dans au moins 1 variété : **99% du 400k PAN-Transcriptome**
- **4067 NA chez les 450**



# Caractérisation fonctionnelle

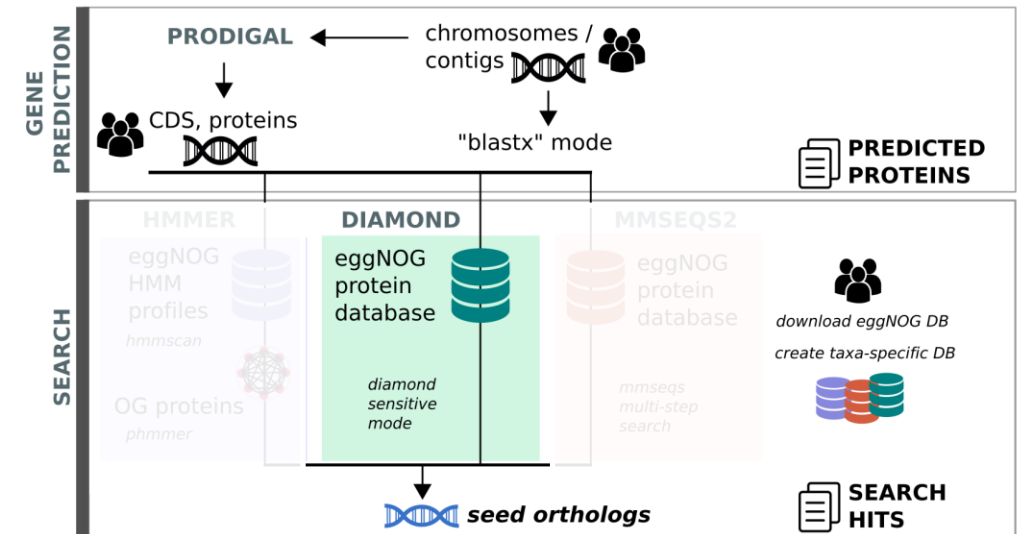
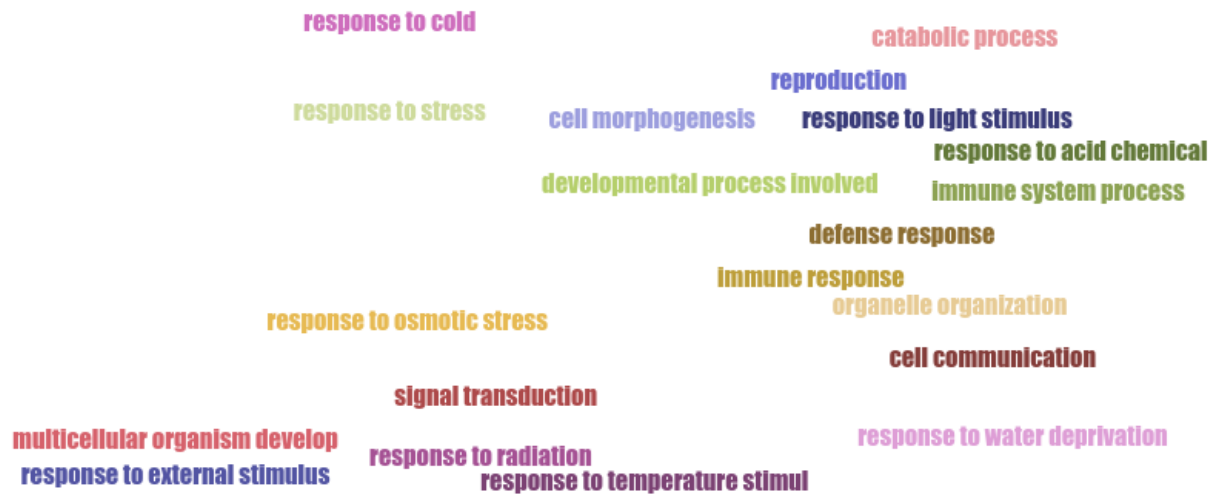
**TransDecoder** : ORFs complets -> protéines

Annotation Fonctionnelle :



**84 741 ORFs/nouvelles protéines avec fonction**

**TopGO (R)** : top 20 processus biologiques



# Caractérisation fonctionnelle

- Core isoforms

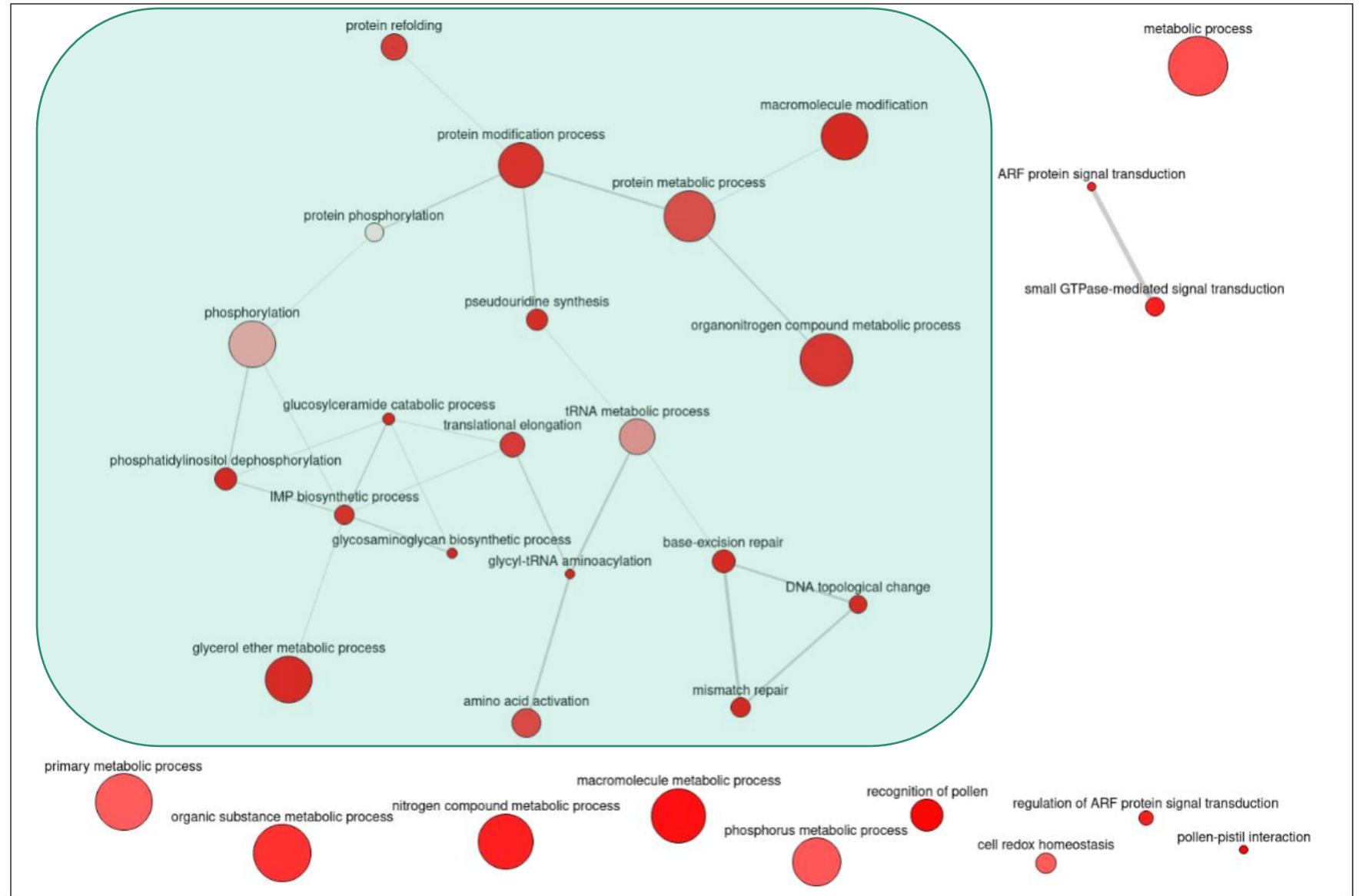
44 BP < 1e-5

Métabolisme basal

Azote

Phosphore

Reproduction (pollen)





# Caractérisation fonctionnelle

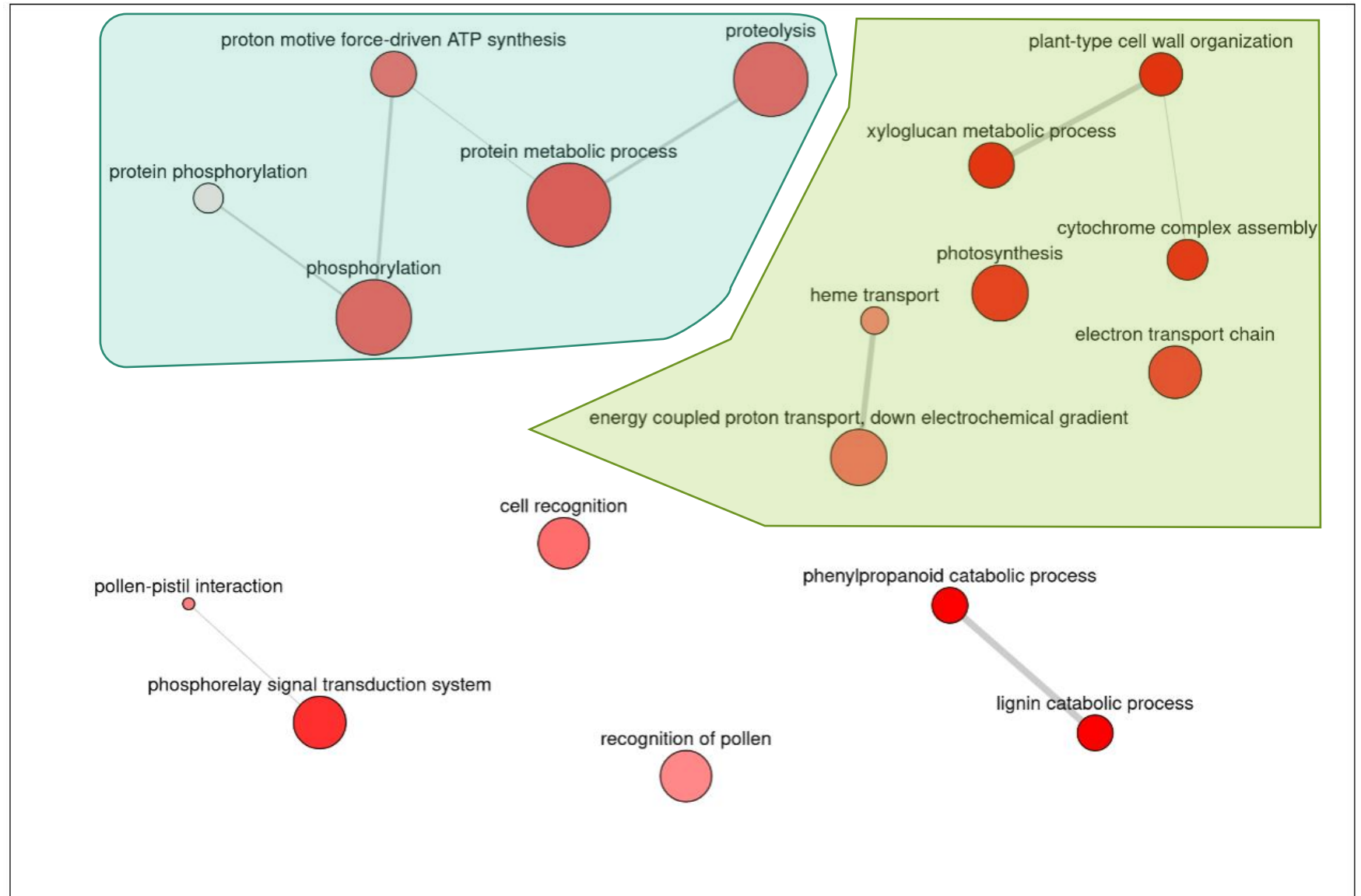
- Cloud isoforms

23 BP < 1e-5

Protéines : Modifs post-trad

Photosynthèse

Reproduction (pollen)



# Architecture cloud dédiée au projet

## Stockage et calcul : plateforme

### Stockage

7 To CEPH

Données brutes et résultats

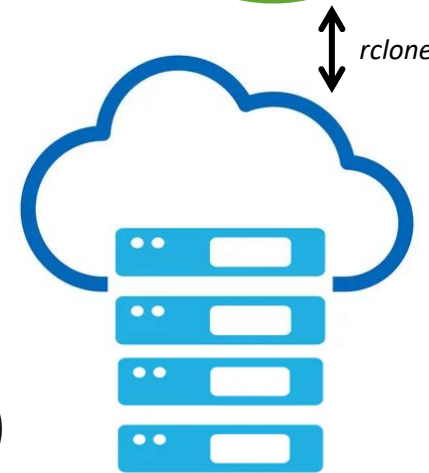
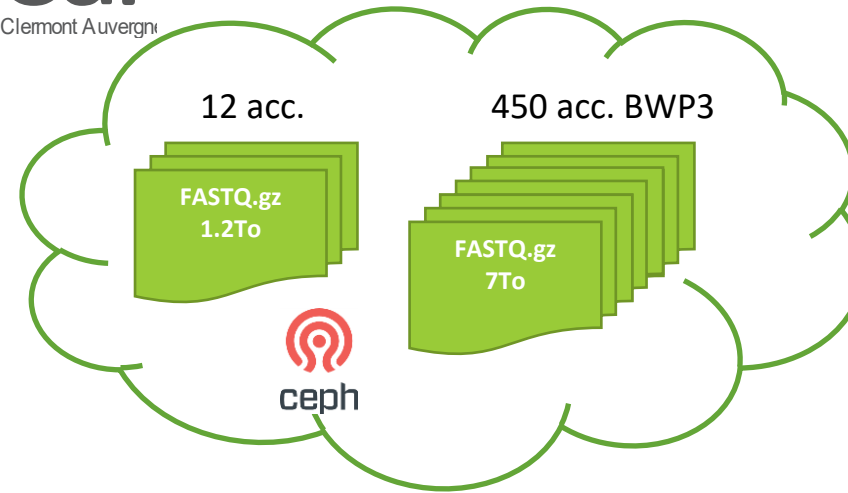
### Calculs : DEBAT-slurm cluster

Infra. as Code : Ansible-GitHub

4 nodes :

- 32 vCPUs
- 128 Go vRAM

5 To SSD shared storage



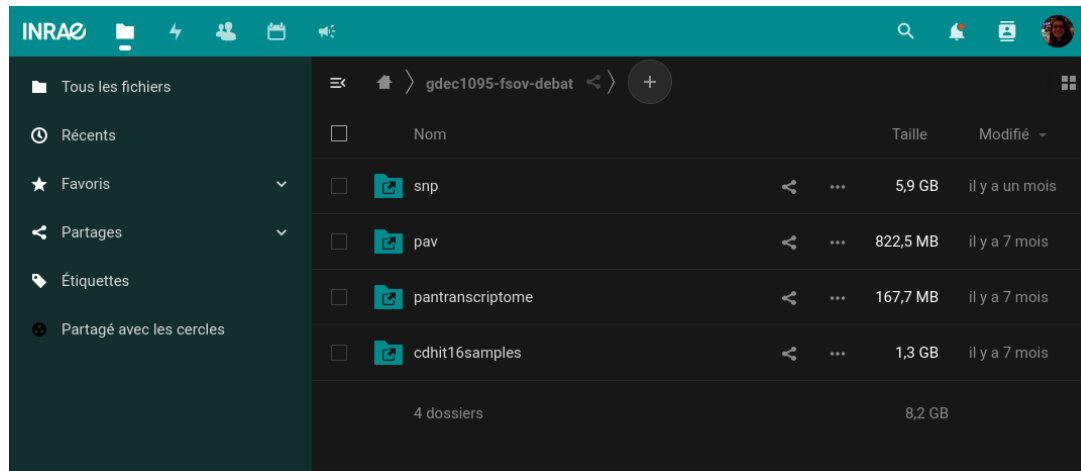
# Partage des résultats

Nextcloud@INRAE

ePAVs

SNPs

Annotations fonctionnelles



RechercheDataGouv.fr

**Accès restreint** : demander un accès avec votre e-mail pro/institutionnel via le lien:

<https://doi.org/10.57745/GQZF6L>

1 contact par partenaire

Omics DataVerse  
(www.inrae.fr)

Recherche Data Gouv > Data INRAE > Omics DataVerse >

## FSOV - DEBAT

Version 1.3

Rimbert, Hélène; Bouchet, Sophie; Choulet, Frederic; Auzanneau, Jérôme; Davey, Mark; Goudemand-Dugue, Ellen; Molero, Gemma; Dutriez, Sylvie; Throude, Mickael; Dufour, Philippe; Giraudeau, Pascal; Argillier, Odile, 2023, "FSOV - DEBAT", <https://doi.org/10.57745/GQZF6L>, Recherche Data Gouv, V1

Citer le jeu de données

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Modalités d'accès au jeu de données

Modifier le jeu de données

Lier le jeu de données

Contact Partager

Statistiques d'utilisation sur les jeux de données

93 consultations

2 téléchargements

0 citation

Description

VCF of PAVs identified based on pan-transcriptomic analysis of 450 wheat accessions.

Sujet

Agricultural Sciences

Mot-clé

genetic diversity, transcriptomique, triticum aestivum

Licence/Conditions d'utilisation des données



CC BY 4.0

# Partage des résultats

European Nucleotide Archive

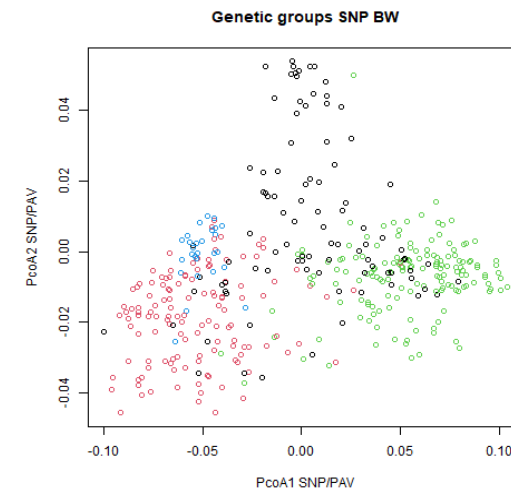
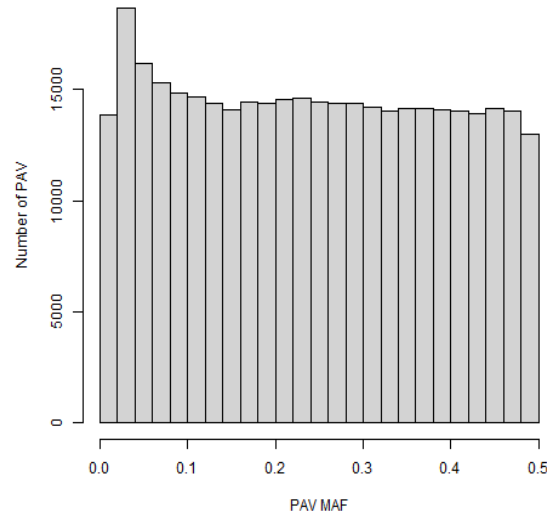
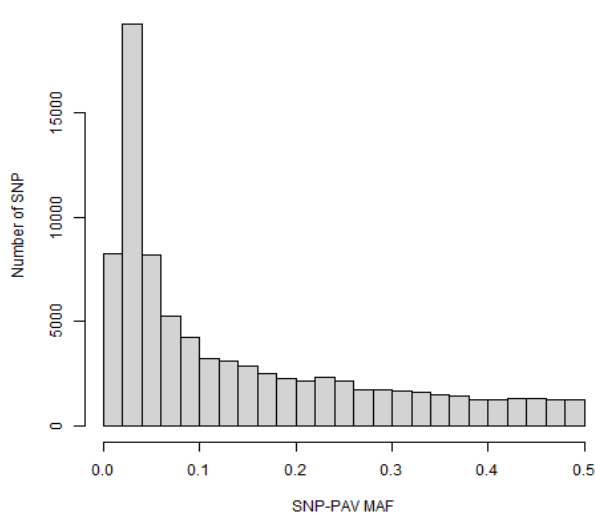


FASTQ en cours de soumission sous le numero d'accèsion PRJEB73381

Accession	Secondary Accession	Title	Submission date	Release date	Status	Action
PRJEB73381	ERP158180	Characterization of the diversity of a 450 bread wheat lines pre-breeding panel	28th Feb 2024	28th Feb 2026 	Private	

# Conclusion

- 400K ePAV avec MAF équilibrée
- 83K SNP avec une MAF > 0.01 et un nombre de données manquantes inférieur à 10%
- groupes génétiques similaires à Breedwheat
- associations avec phénotypes dans le FSOV Ex-IGE

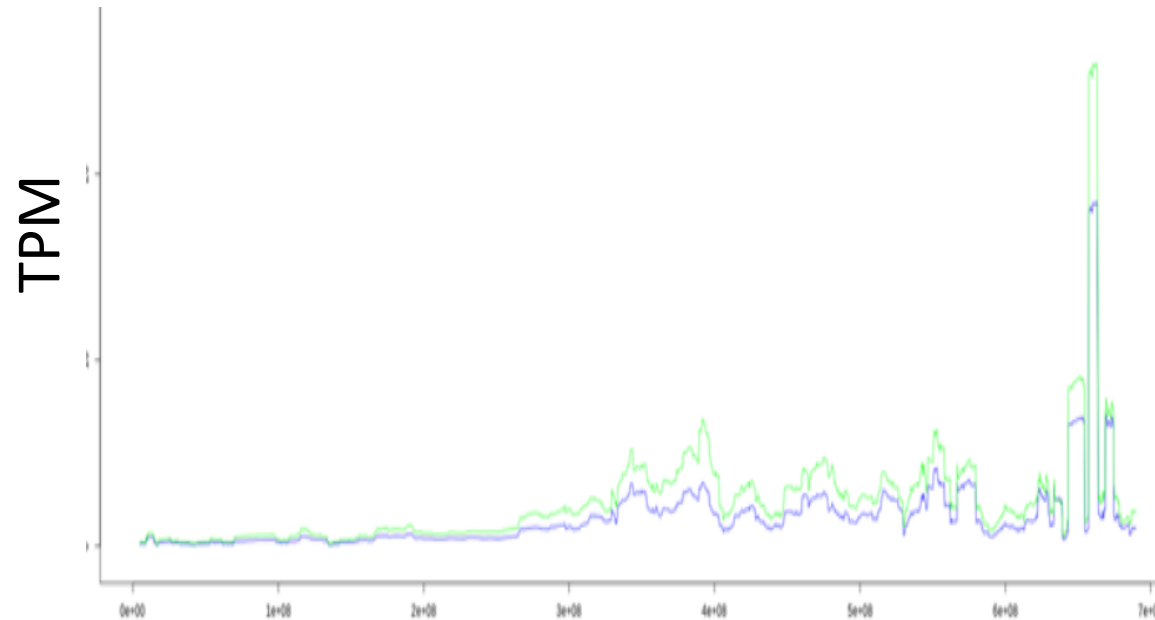


# Perspectives

- pour une partie des 500K SNP détectés, les données manquantes correspondent peut-être à des PAV: à vérifier
- moitié des ePAV absents d'une accession à l'autre: possible? beaucoup de TPM (à partir de pseudo-comptes de reads) à 0 doivent correspondre à des présences avec expression faible
- gros bruit de fond des TPM: on teste sur des introgressions connues (1RS/1BL) le niveau de bruit de fond (autour de 0)
- on teste un algorithme de segmentation pour détecter des introgressions, comparaison avec résultats puce 420K
- mapping des reads sur la séquence d'autres espèces

# Exemple TPM Solaris sur 1B (1RS/1BL)

TPM proche de 0 sur 1BS mais pas = 0, bruit de fond



Position chromosome 1B (pb)



merci

**GDEC**

Sophie Bouchet  
Clément Debiton  
Jonathan Kitt  
Pauline Lasserre-Zuber  
Fabien Leloup  
Frederic Choulet  
Etienne Paux (Vet-Agro)

**AuBI-Mésocentre UCA**

Nadia Goue  
Sammy Gilles  
David Grimbichler