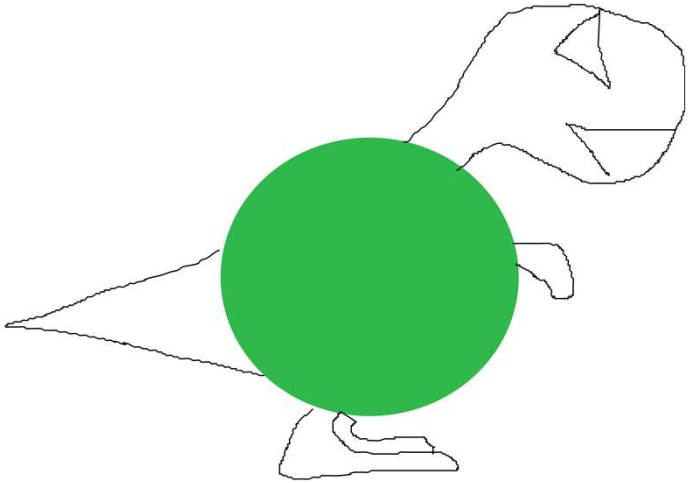
Transcriptomique par RNA-seq avec ou sans référence chez les légumineuses

Journée PEPI IBIS 06/06/2019

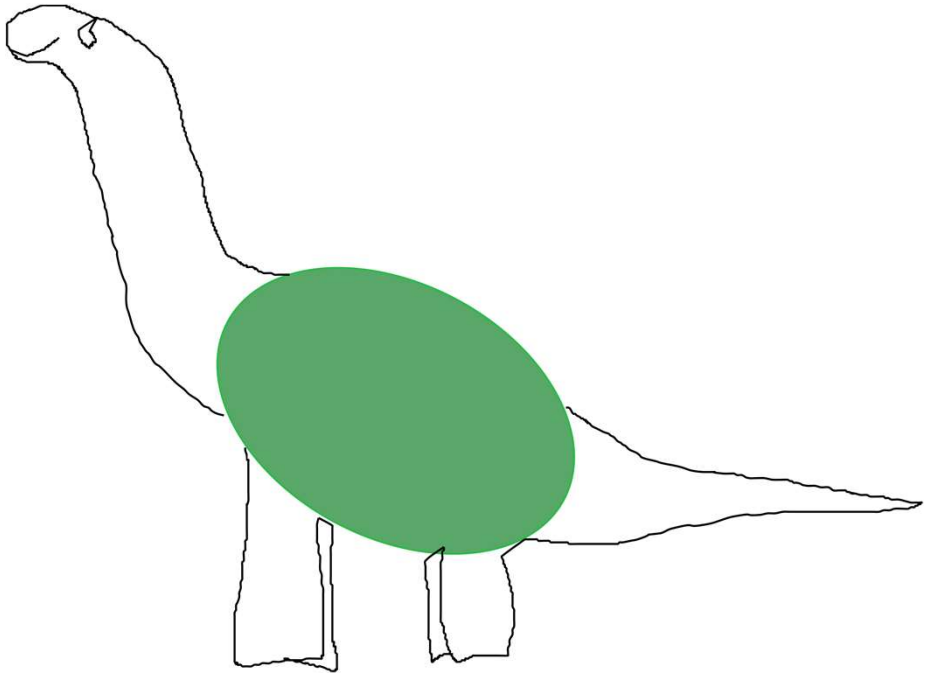

LEGUMESPARK

LA FOIRE AUX LEGUMOSAURES



Pisum "Pearanosaurus" sativum
4.5Gb diploide, génome disponible

RNA-seq séquencés dans le cadre de l'ANR REGULEG, coordonné par J.Buitnik.
Analyse en 2017 - 2019



Vicia "Fabachiosaurus" Faba
13,5Gb diploide, génome indisponible

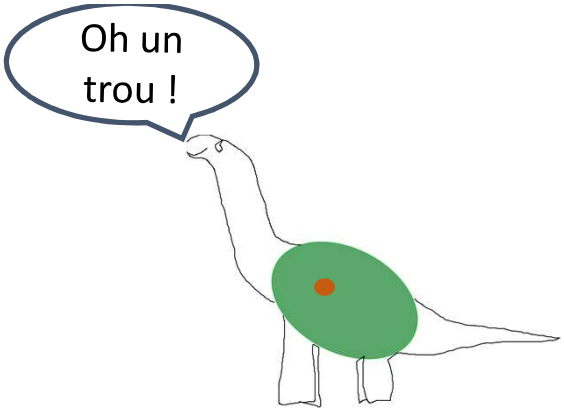
RNA-seq séquencés dans le cadre du PIA PEAMUST coordonné par J. Burstin
Analyse en 2017-2018

PROJET FINANCÉ PAR L'ANR - ANR - PROJECT FUNDED BY THE ANR

INVESTISSEMENTS D'AVENIR

PeaMUST

La bruche, terreur des fèveroles



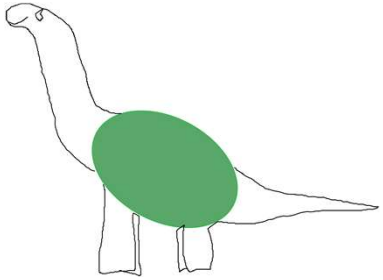
- La bruche femelle se nourrit des fleurs de fèves.
- Elle pond ensuite sur la gousse.
- Les larves rentrent dans la gousse et la graine pour se nourrir.
- L'insecte adulte sort de la graine.
- Augmente la chance de maladie pour la graine.
- Vente impossible pour la consommation humaine dans les pays acheteu (pourtour méditerranéen).

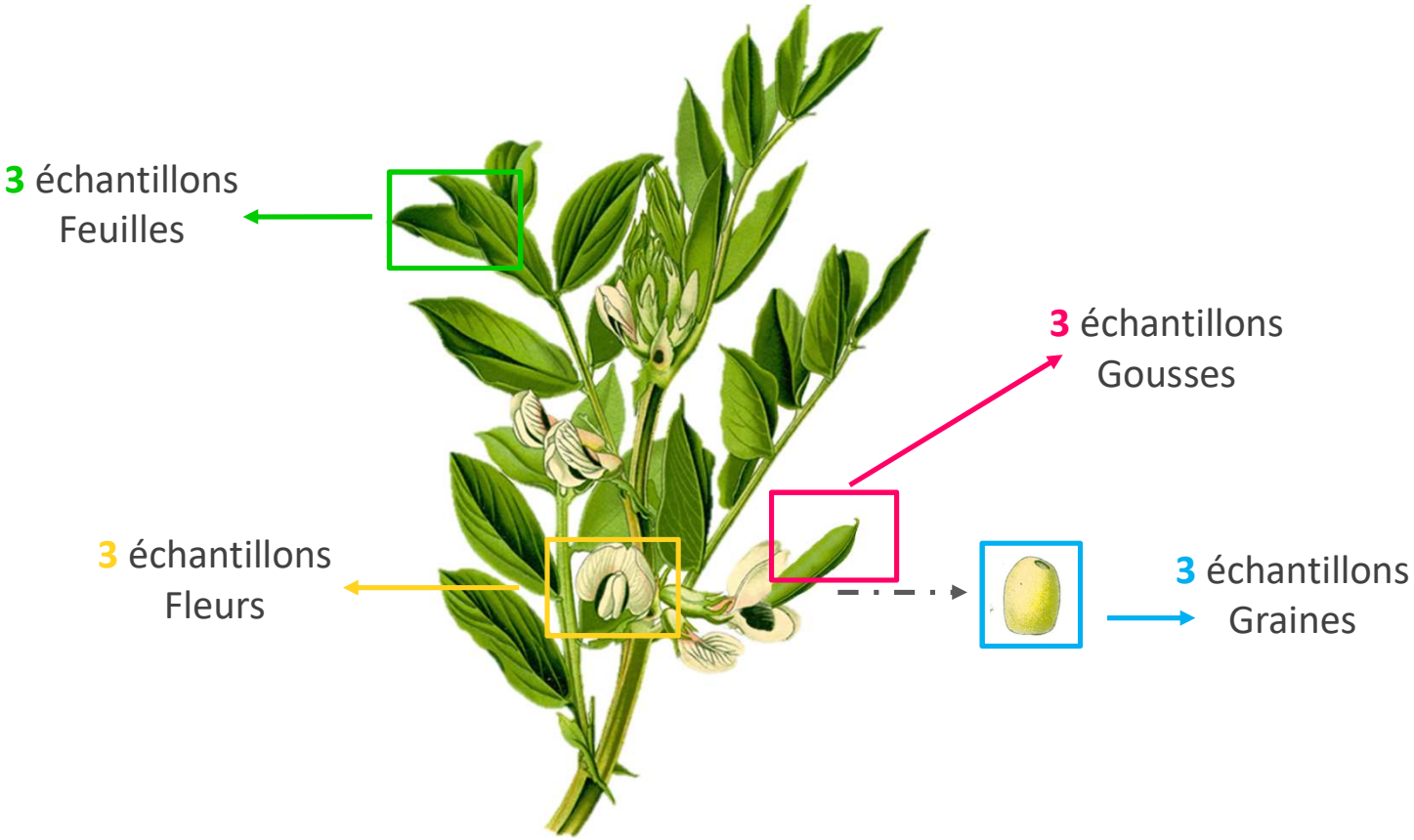
- Test en champs de 29 accessions.
- Sélection de deux accessions résistantes de façon différentes :

- 159b : "Répulsion" des insectes.
- 2378 : Graines toxiques pour la bruche.

Carrillo-Perdomo, Estefanía, et al. "Identification of novel sources of resistance to seed weevils (Bruchus spp.) in a faba bean germplasm collection." Frontiers in plant science 9 (2018).

Protocole expérimental





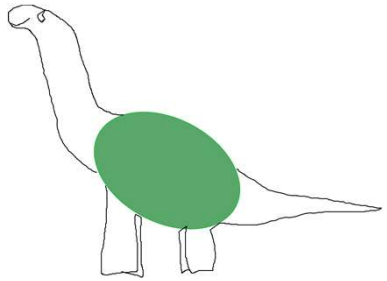
3 Cultivars :
- 159b
- Hiverna (sensible)
- 2378

Conditions de champs, bruches présentes pour tous les échantillons.

3 réplicats par tissus
Séquençage Illumina 2000/3000
2x125 paired-end stranded

Une expérience multi-tissues, multi-espèces !

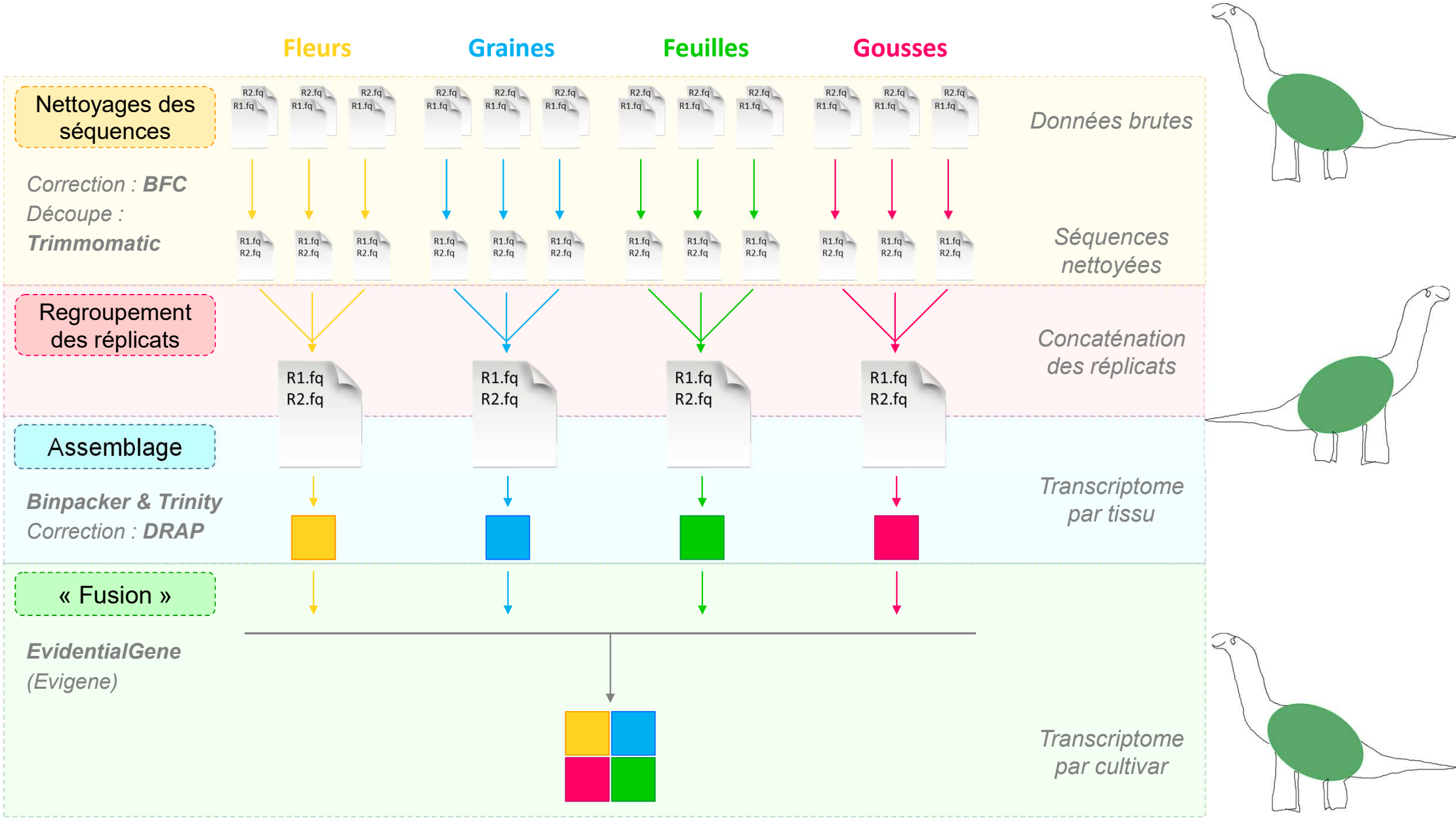
Une stratégie d'assemblage progressive



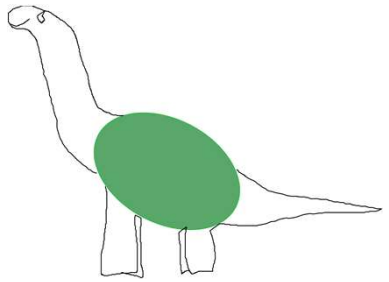
Pour comparer l'ensemble de ces échantillons, il nous faut une référence commune.

Décision d'assembler l'ensemble progressivement d'abord par tissus, ensuite par cultivars et enfin de rassembler l'ensemble.
Pilotage de la qualité des assemblages en utilisant une double validation:
- Automatique par l'outil transrate.
- A la main, par l'expertise de familles connues par G. Aubert

Smith-Unna, Richard, et al. "TransRate: reference-free quality assessment of de novo transcriptome assemblies. " Genome research 26.8 (2016): 1134-1144.


Fleurs
Graines
Feuilles
Gousses
Nettoyages des séquences
Correction : BFC
Découpe : Trimmomatic
Données brutes
Séquences nettoyées
Regroupement des réplicats
Concaténation des réplicats
Assemblage
Binpacker & Trinity
Correction : DRAP
Transcriptome par tissu
« Fusion »
EvidentialGene
(Evigene)
Transcriptome par cultivar
R1.fq
R2.fq

Deux assembleurs bien différents



BinPacker

- Méthode rapide
- Méthode plus stringente
- Utilise 1 threads

```
Paramètres :
        -k : taille du k-mer
(31)
        Paramètres par défaut
```

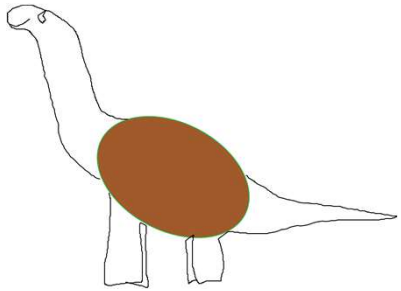
Trinity

Assembleur transcriptome *de novo* RNASeq

- Méthode de référence
- Outils compagnons
- Communauté importante

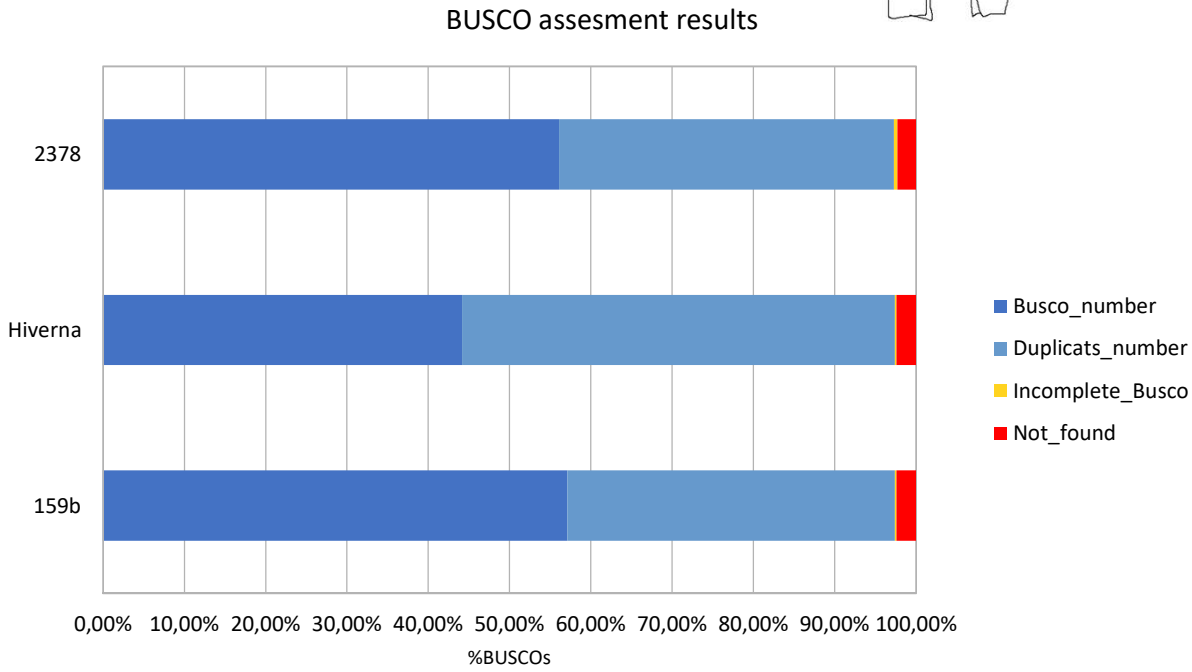
```
Paramètres :
        Paramètres par défaut
```

Résultats



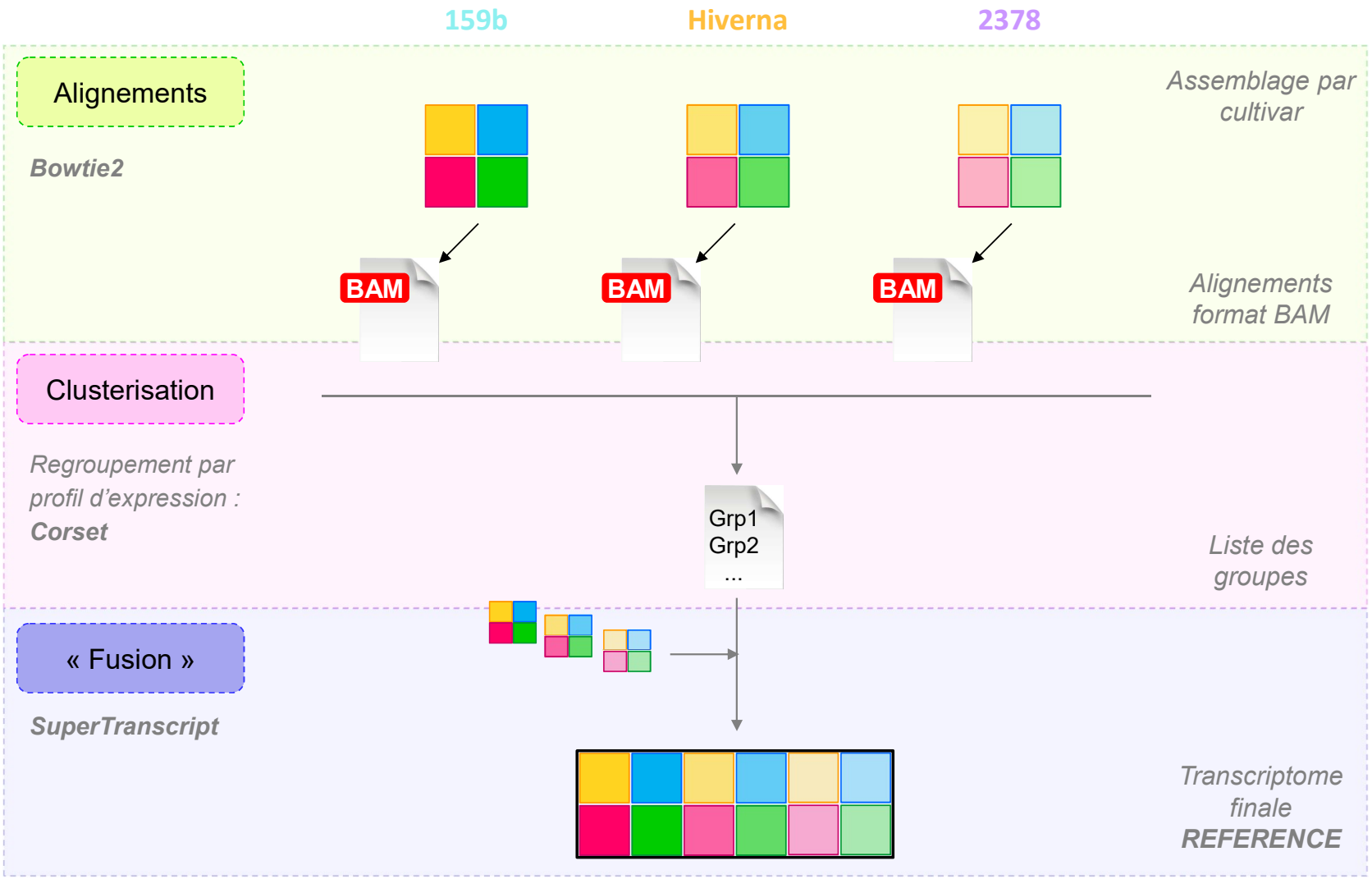
	159b	Hiverna	2378
Predicted_transcript_nb	33003	40745	35315
Smallest (base)	201	201	203
Largest (base)	19408	32071	24892
Mean_contig_length	1714.67	1668.16	1641.57
Transcript_with_ORF	27980	33651	29286
Mean_ORF_percent	75.69	75.52	76.11
n70	1588	1570	1541
n50	2089	2086	2045
n30	2817	2813	2785
%_good_mapping (reads)	0.73	0.78	0.73
Transrate_assembly_scor	0.3096	0.3546	0.3243
Transrate_optimal_score	0.3923	0.4376	0.4055
Transrate_optimal_cutoff	0.1983	0.0803	0.2281
Good_contigs_nb	29347	38280	31279

Résultats de transrate

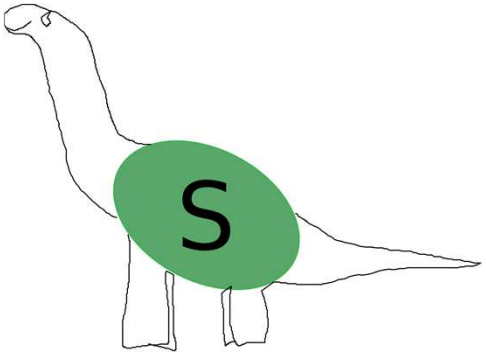


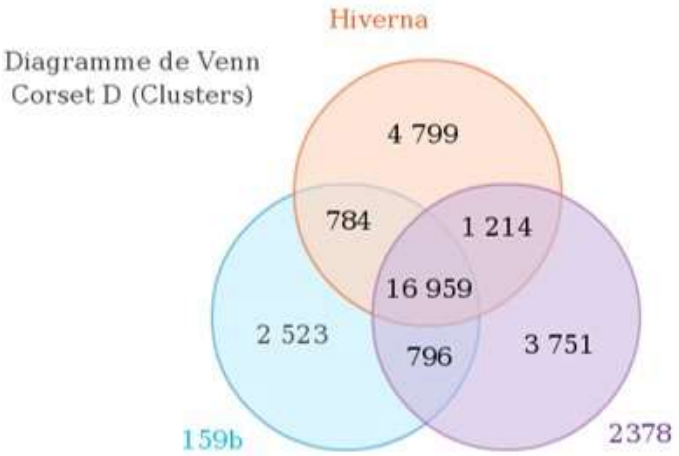
Résultats de BUSCO v1

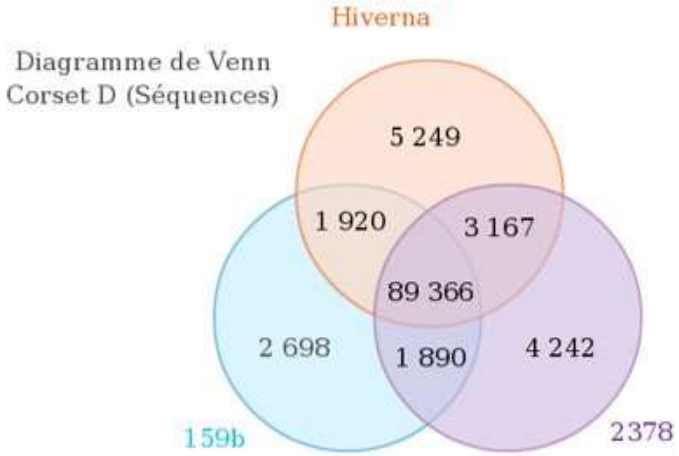
Simão, Felipe A., et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics 31.19 (2015): 3210-3212.

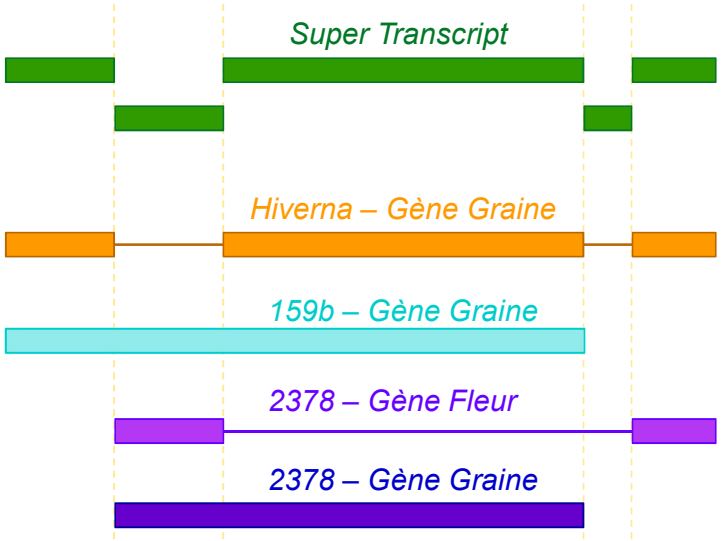

159b
Hiverna
2378
Alignements
Bowtie2
Assemblage par cultivar
BAM
BAM
BAM
Alignements format BAM
Clusterisation
Regroupement par profil d'expression : Corset
Grp1
Grp2
...
Liste des groupes
« Fusion »
SuperTranscript
Transcriptome finale REFERENCE

Supertranscripts !





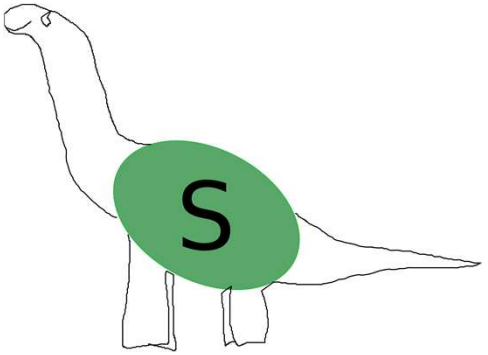


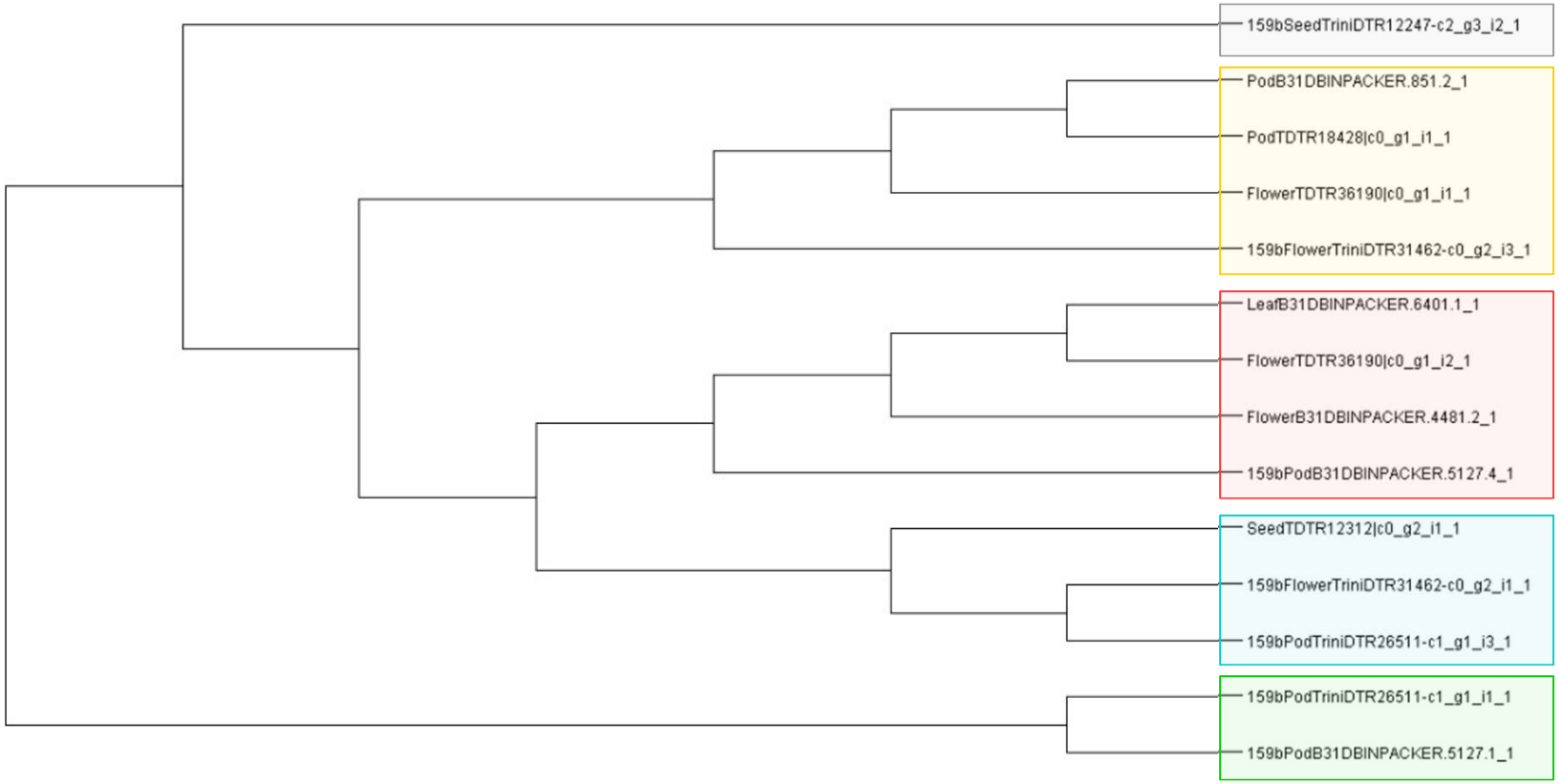


	ST	159b	Hiverna	2378
Predicted_transcript_nb	30825	33003	40745	35315
Busco_number	635	546	422	536
Duplicats_number	281	385	509	394
Incomplete_Busco	13	2	2	4
Not_found	27	23	23	22
Protein_nb	**30825**	**33081**	**40864**	**35409**

Hawkins A DK, Oshlack A, Davidson N M. SuperTranscript: a reference for analysis and visualization of the transcriptome. *BioRxiv*, 2016

Imparfait mais suffisant pour l'expression différentielle





5 gènes ?

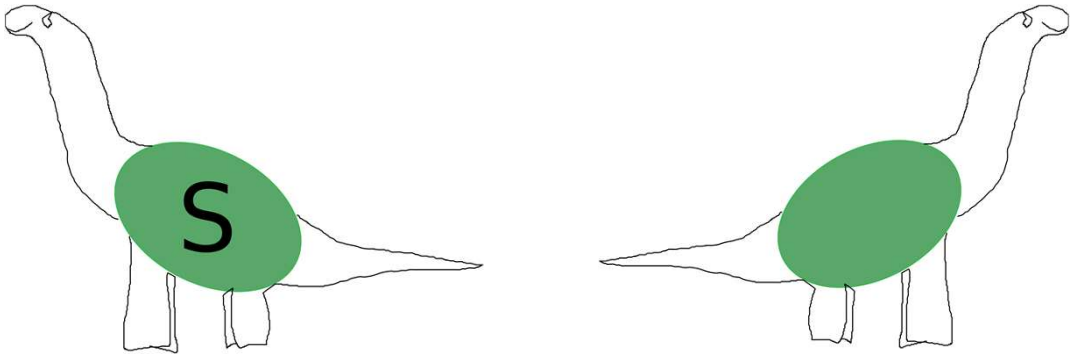
Et ensuite ?

Décision de mettre en commun nos ressources avec le projet danois NORFAB.

Création d'un transcriptome commun utilisant aussi leurs données.

Un protocole d'assemblage proche de celui présenté a été décidé.

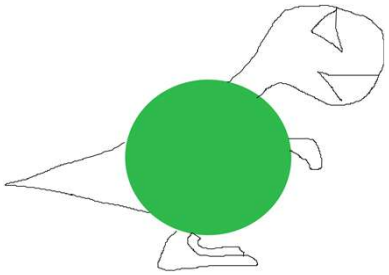
Le choix d'utiliser supertranscripts ou une autre méthode est encore en suspens.



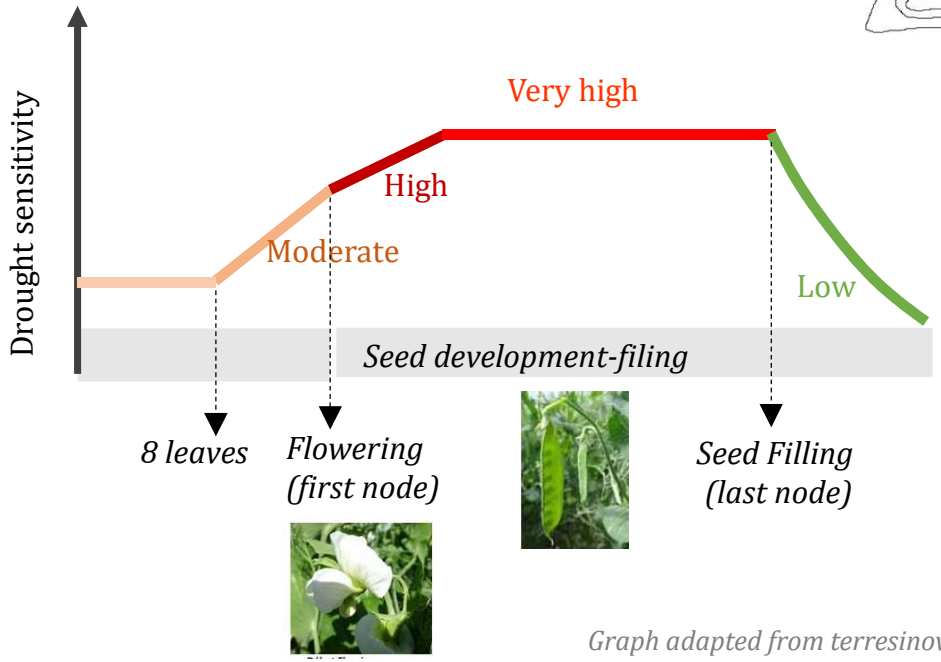
PeaMUST

AARHUS UNIVERSITET

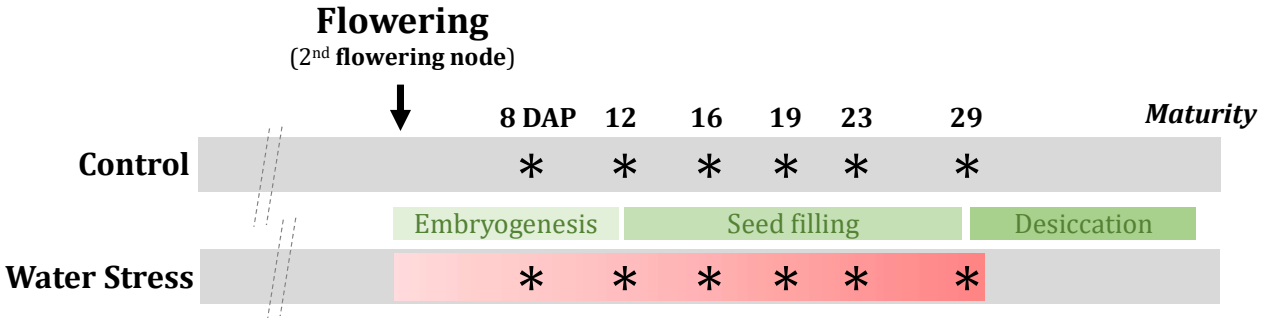
La vie rêvée des graines...



- Le pois produit des graines riches en protéines
- Un stress hydrique peut influer sur la quantité et la qualité des graines.
- L'accumulation de protéines de stockages (SSP) durant le remplissage est un processus hautement régulé.
- La connaissance de ces réseaux de régulations est encore à améliorer



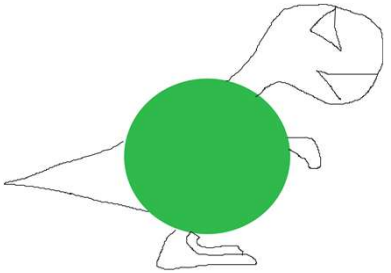
Graph adapted from terresinovia.fr



DAP = Day After Pollination

Séquençage Illumina 3000
2x125 paired stranded
~ 30M paired reads

Quand le génome va...



Genome Features	Values
Length of genome assembly	3,920,161,095
Total length of scaffolds	3,919,096,294
Number of scaffolds	24,623
N50 of scaffolds	415,940
Anchored scaffolds	10,357
Total length of contigs	3,159,358,344
Number of contigs	218,010
N50 of contigs	37,931
GC content (%)	37.6
Total length of pseudomolecules	3,234,741,624

Contig length (%)	Values
Retrotransposons (Class I)	77.8%
incl. LTR Retrotransposons	72.6%
Transposons (Class II)	5.4%
Genes	3.9%
Number of genes	44,756

- Un génome partiel (3,2Gb sur 4,3Gb).
- Le transcriptome déjà réalisé s'aligne bien sur le génome et les scores BUSCO étaient bon.
- Les ressources utilisés par l'annotation comprenaient peu de tissus de graines.
- Une annotation spécifique des SSPs a été réalisé.

RECHERCHE DE NOUVEAUX GENES

→ Concentration sur Classe U
→ Suppression des isoformes

Stringtie

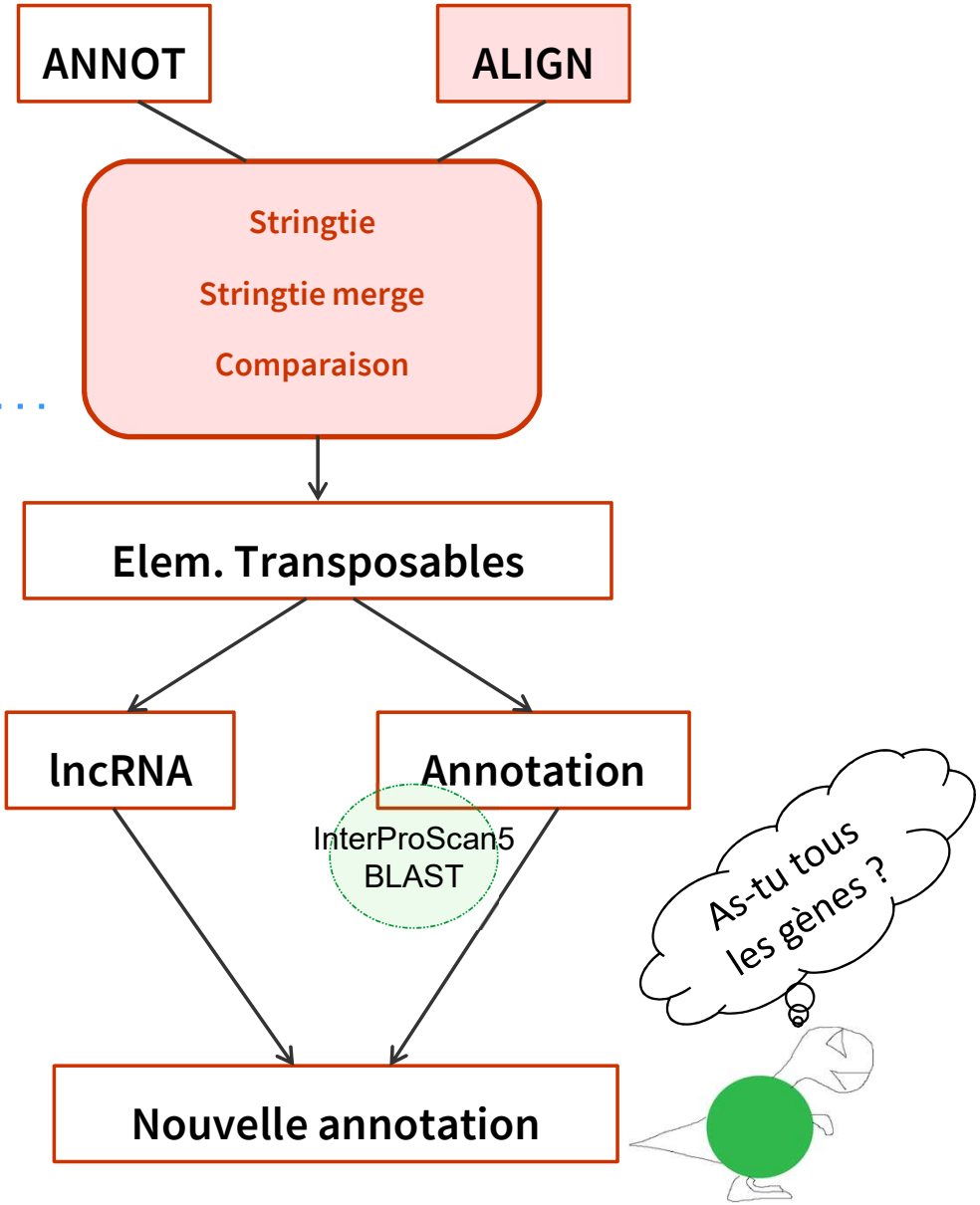
Identification des éléments transposables

→ Pas de CDS
→ Non utilisés pr les comptages
→ FeeInc

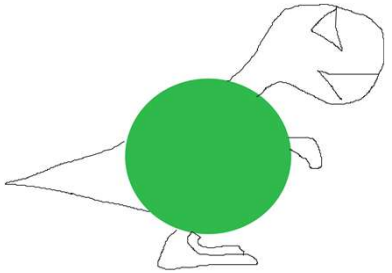
→ Transdecoder
→ Protocole : Cormier et al, 2017

Identification des lncRNA

Comptage



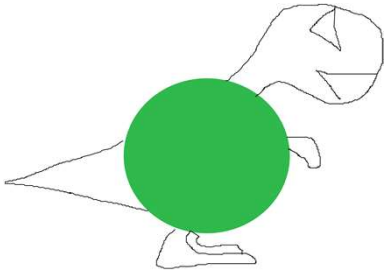
Une recherche drastique



	Classe U Sans isoformes - Non E.T.	mRNA	mRNA with CDS
Nb_genes	1 478	1 255	203
Nb_transcrits	1 551	1 270	206
Nb_exons	1 987	-	-

	Annotation	Nb_total_transcrits	Nb_transcrits_annotés	%_transcrit_annotés
Classe U Sans isoformes - Non E.T. mRNA with CDS	GO	206	10	4,8 %
	IPR / GO / Reactome / KEGG ...	206	17	8,25 %
	Blast Ath TAIR10	206	75	36,4 %
	Blast Mtr v4	206	113	54,8 %

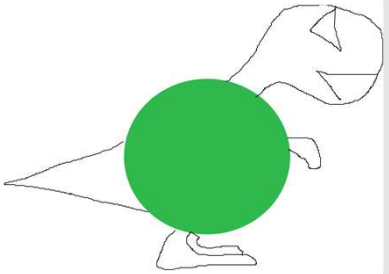
Après l'alignement...

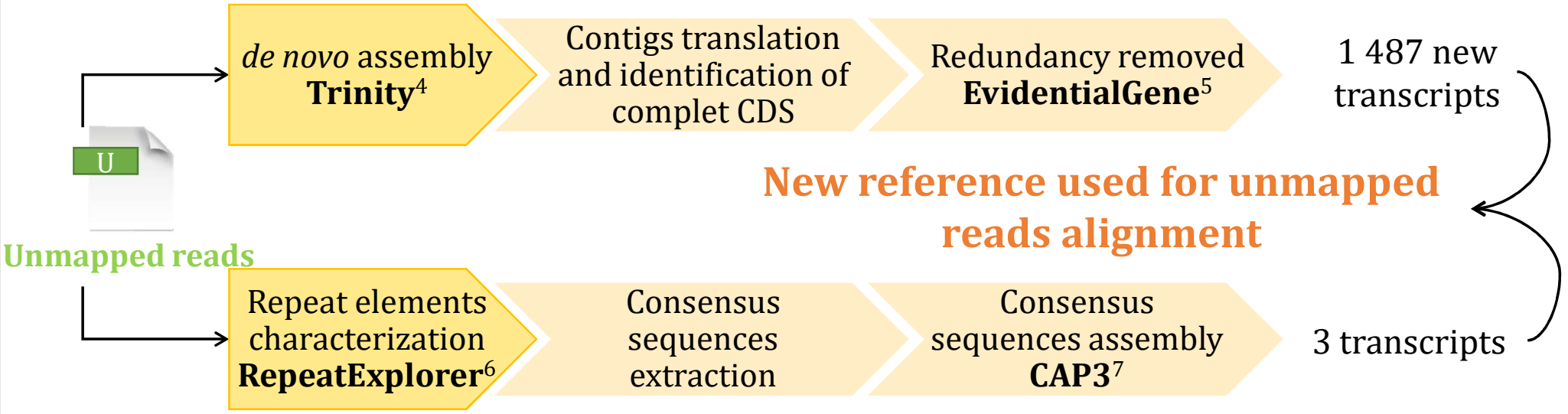


	Paired_number	%_mapped	%_reads_mapped_once	%_multimapped_reads	%_unmapped_reads
8 dap	27 404 422	96,90%	89,70%	7,20%	3,10%
12 dap	31 055 841	97,39%	89,24%	8,15%	2,61%
16 dap	27 244 707	88,87%	70,28%	18,59%	**11,13%**
19 dap	28 523 145	91,63%	68,56%	23,07%	**8,37%**
23 dap	33 062 664	96,21%	73,56%	22,65%	3,79%
29 dap	28 611 114	97,39%	87,46%	9,92%	2,61%

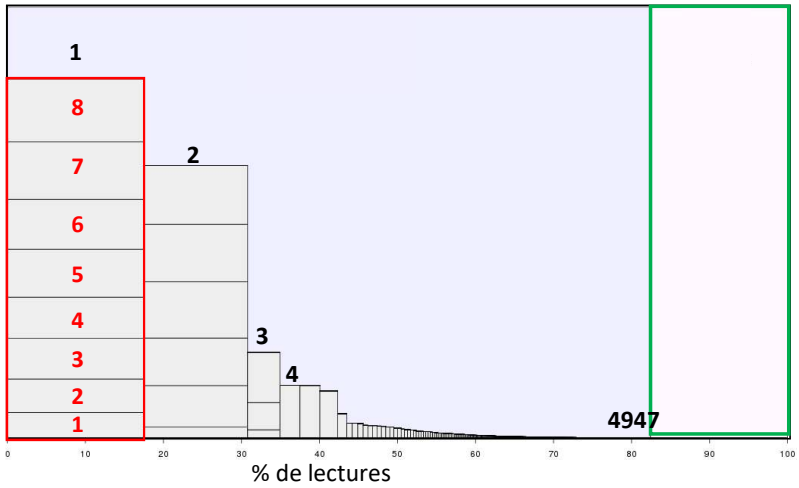
- Deux échantillons en pleine phase de remplissages ont un nombre de lectures non-alignées importants.
- On peut aussi voir une tendance sur les reads multi-mappés.
- Quel est le contenu de ces lectures ?

Récupérer les lectures non-alignées

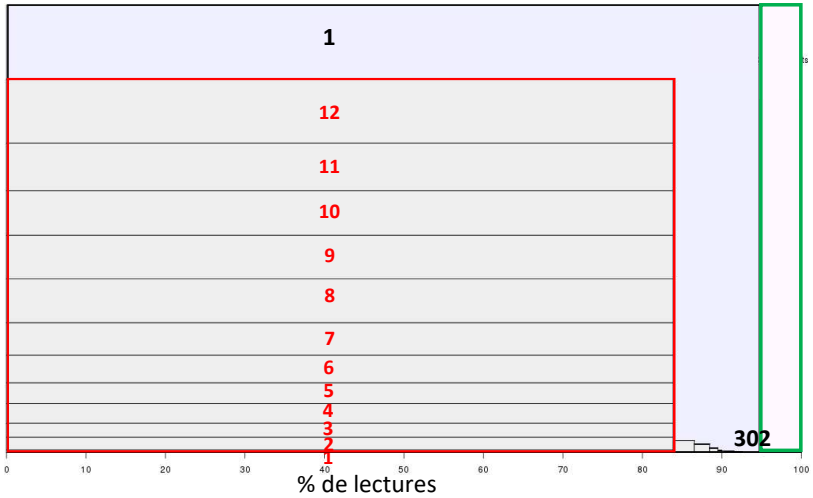




RepeatExplorer



Echantillon 8_WW

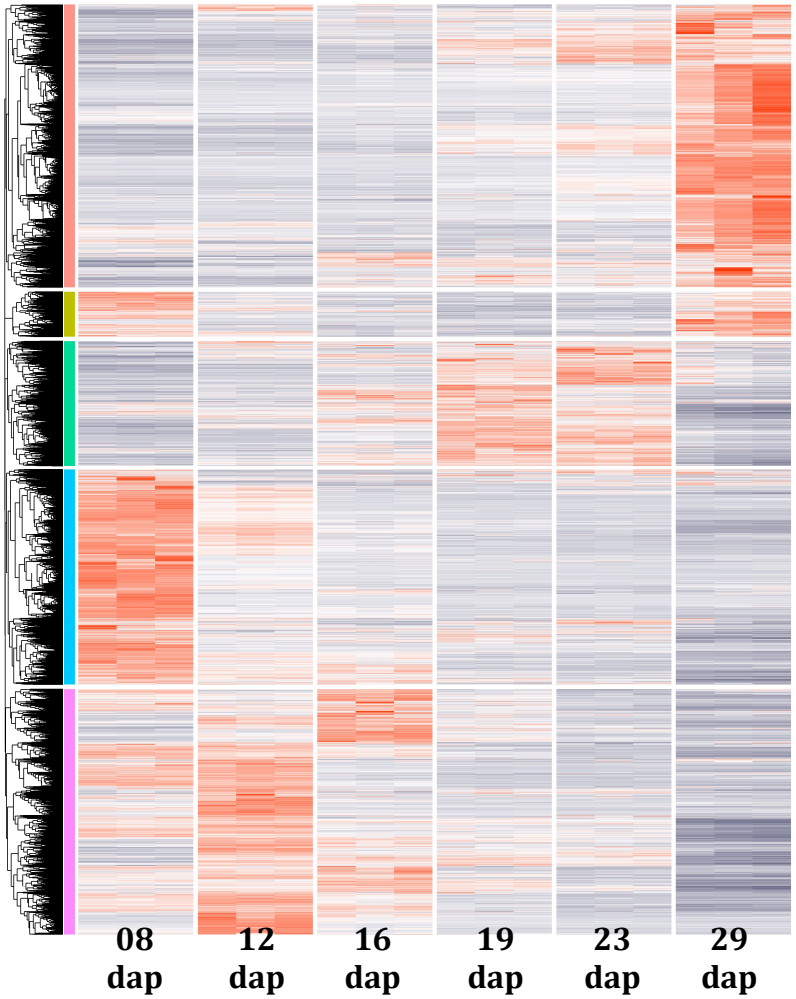


Echantillon 16_WS

	8_WW	16_WS
18S rDNA	17,5%	1,03%
25S rDNA	17,44%	0,66%
Satellite	0,02%	0%
Mitochondries	2,59%	0,32%
Non caractérisé :	32,75%	90,58%

Pour 16WS, après blast et réassemblage des clusters, on obtient 3 vicilines. Les vicilines sont très abondantes et des copies de celles-ci manquent sur le génome.

Des résultats prometteurs



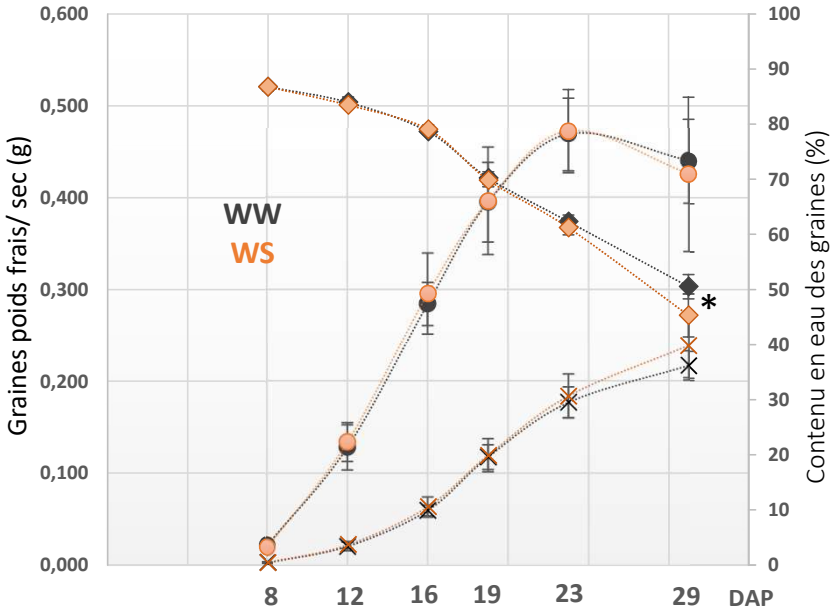
WS_vs_WW	Total_DEG	UP	DOWN
8 dap	309	165	144
12 dap	87	2	85
16 dap	92	46	46
19 dap	0	0	0
23 dap	1	1	0
29 dap	7030	3599	3431

Est-ce que le faible nombre de gènes différentiellement exprimés à 19 et 23 dap est du à :

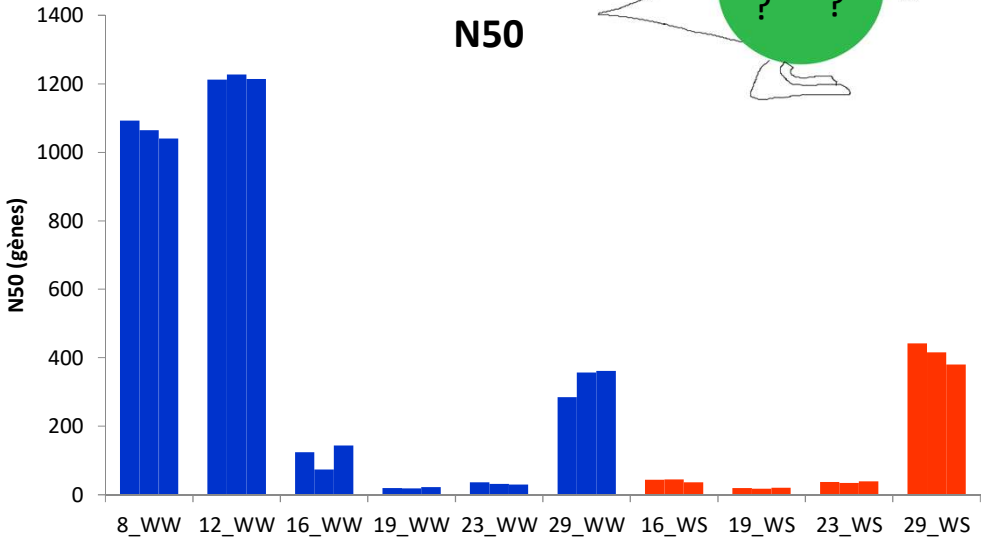
- Un effet faible du stress hydrique ?

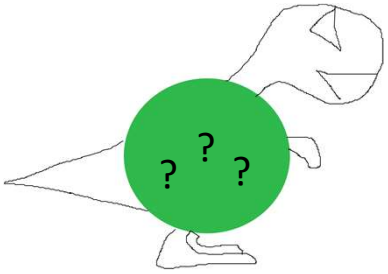
- Un problème de profondeur de lectures dû aux SSPs ?

Des questions en suspens…



Un profil WW/WS quasi-identique.



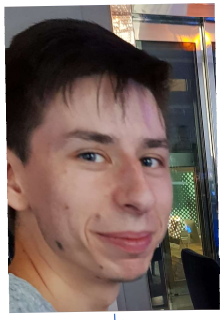


Pour les échantillons à 19 et 23 DAP, plus de la moitié des reads est alignés sur moins de 20 gènes.

On recherche des avis, des expériences identiques !

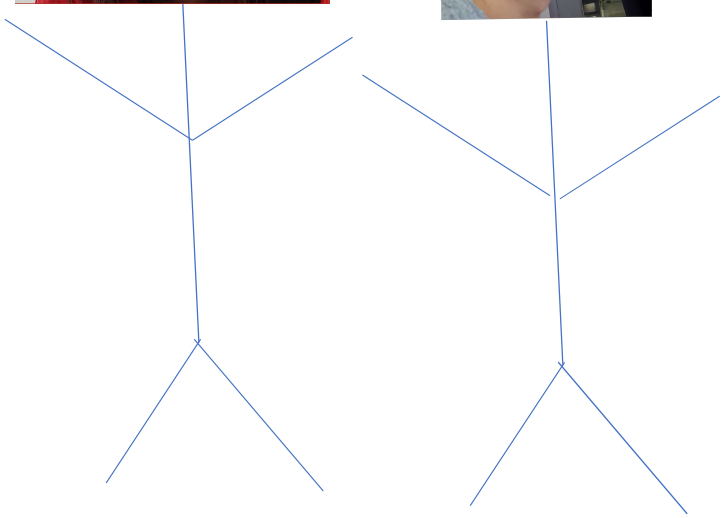
Remerciements

Morgane Terezol
CDDs IE
Assemblage féverole
Recherche de gènes Pois
Expression différentielle



Adrien Mazuel
Stage de M2
Recherche lectures non-alignées
Expression différentielle

En recherche d'emploi !



Séquençages réalisés à Genotoul.
Travaux Féverole :
Grégoire Aubert
Catherine Desmetz

Travaux Pois :
Vanessa Vernoud

Et toute l'équipe ECP et FILEAS !

ALGO EVIGENE

- Entrée : transcrits format FASTA
- Production de CDS et traduction en protéines
- Suppression des redondance : **fastanrdb**
- Clusterisation des fragments : **cd-hit-est**
- Alignements (alternatif) : **blastn**
- Classification des majeurs et alternatifs : **CDS-align**
- Sortie : okey-main, okay-alts et drop