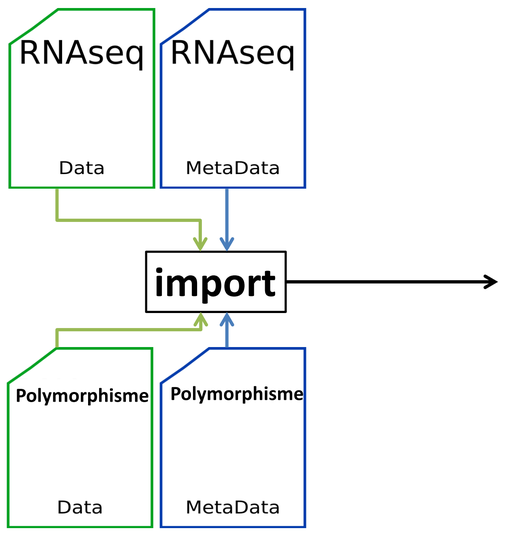
Structuration et consolidation de résultats d'analyses de RNAseq et Polymorphisme

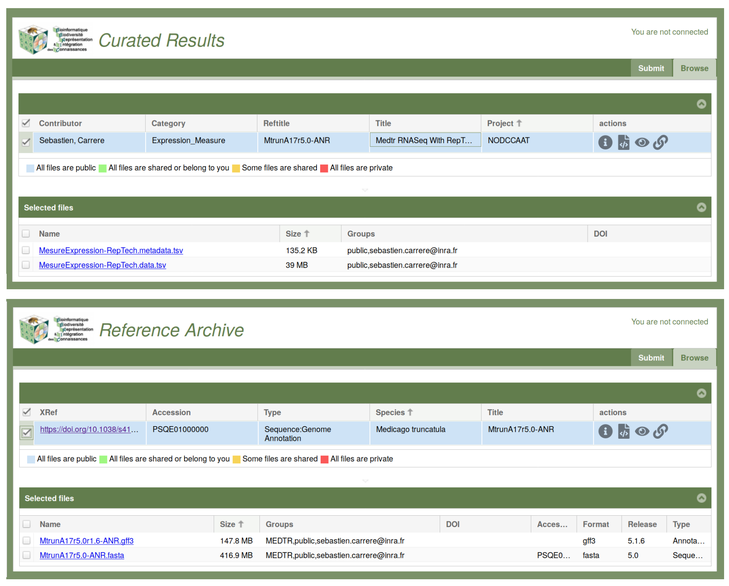
Thomas Garcia[1], Ludovic Legrand[1], Jérôme Gouzy[1], Sébastien Carrere[1]
Laboratoire des Interactions Plantes Micro-organismes
LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

NODCCAAT [1]

L'accumulation de résultats d'analyses « omiques » au cours du temps et par différents acteurs rend leur utilisation et leur intégration avec de nouveaux résultats difficile. La production de ce type de résultats sur la légumineuse modèle *Medicago truncatula*[2] sur plus de 20 ans et sur plusieurs versions d'assemblage est un cas d'usage sur le besoin de structurer et normaliser ces résultats afin que leur exploitation sur le long terme soit envisageable. Nous proposons ici une solution de structuration, transformation et normalisation automatisée de résultats d'analyse stockés sur ARCHIVE[3].

STRUCTURATION DE L'EXISTANT





Un préalable à l'intégration de résultats d'analyse est leur stockage dans des formats pérennes. Pour cela nous utilisons une instance de l'ARCHIVE. Cette brique logicielle permet de :

- fournir les **métadonnées** descriptives des résultats d'analyses (génome et annotation de référence utilisés, type d'analyse, type d'identifiants, contributeurs)
- Déclencher des **contrôles d'intégrité** se basant sur ces métadonnées
- Offrir un **entrepôt pérenne et interrogeable** depuis nos ressources de calcul

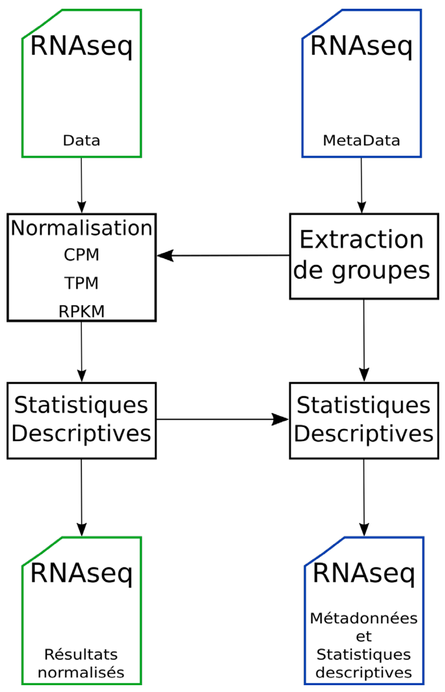
TRANSFORMATION ET NORMALISATION

Les métadonnées associées à chaque fichier de résultat permettent:

- d'identifier les groupes de répétitions techniques et biologiques
- de calculer des statistiques descriptives des résultats d'analyses

Les résultats de RNAseq sont normalisés selon trois méthodes[4].
Les résultats de polymorphisme sont transformés: chaque position polymorphe est analysée (impact fonctionnel, nature) et l'ensemble est consolidé en une matrice de polymorphismes par objet biologique et par échantillon.

Des statistiques descriptives des résultats normalisés / transformés complètent les statistiques descriptives initiales.



Utilisation des métadonnées

#LIB	FILTERED	LIBSIZE	MAPPED	LIBTYPE	SAMPLE	GROUP
MEDTR-C2-T0-A	5097967	5800938	5224871	OPE2*	MEDTR-T0	C2
MEDTR-C2-T0-B	2548168	2960861	2612916	OPE2	MEDTR-T0	C2
MEDTR-C2-T1-A	5154607	5858063	5283336	OPE2	MEDTR-T1	C2

*OPE2 : Oriented Paired End

Répliquas Techniques		
MEDTR-T0	MEDTR-C2-T0-**A**	MEDTR-C2-T0-**B**
MEDTR-T1	MEDTR-C2-T1-**A**	MEDTR-C2-T1-**B**

Répliquas Biologiques		
C2	MEDTR-T0	MEDTR-T1

rnaseq-medtr-t0-c2-count
Type d'analyse | Nom de l'échantillon | Mesure

Résultats brut

id	MEDTR-C2-T0-A	MEDTR-C2-T0-B	MEDTR-C2-T0-C	MEDTR-C2-T0-D
MtrunA17_Chr1g0146181	9	22	4	5

Résultats normalisés

id	rnaseq-medtr-t0-c2-count	rnaseq-medtr-t0-c2-cpm	rnaseq-medtr-t0-c2-tpm
MtrunA17_Chr1g0146181	40	1,886	1,69

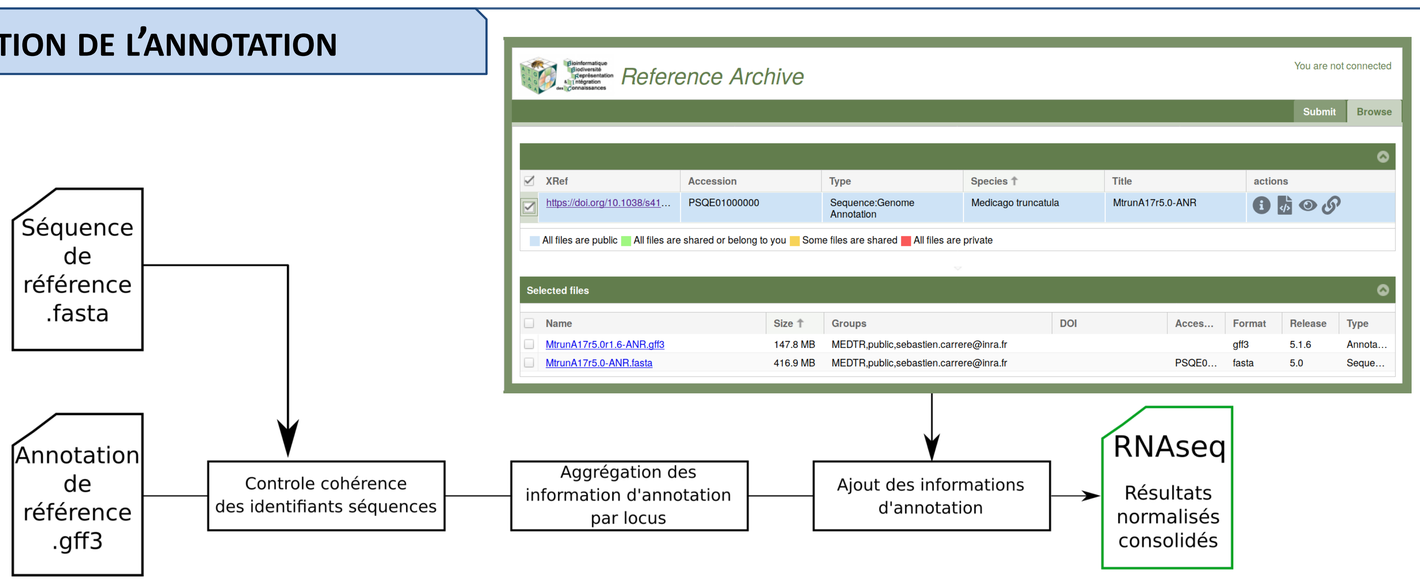
Méthodes de normalisation RNAseq:

$$TPM = \frac{r_g \times rl \times 10^6}{fl_g \times T} \qquad RPKM = \frac{r_g \times 10^9}{fl_g \times R} \qquad CPM = \frac{r_g}{\frac{R}{10^6}}$$

$$R = \sum_{g \in G} r_g \qquad T = \sum_{g \in G} \frac{r_g \times rl}{fl_g}$$

$r_g = nombre\ de\ reads\ pour\ un\ gène$
$fl = longueur\ du\ gène\ (kb)$
$T = total\ des\ reads$ (normalisés par la taille)
$R = total\ des\ reads$

INTÉGRATION DE L'ANNOTATION



Les résultats normalisés ou transformés sont complétés avec des informations d'annotations fonctionnelles.
Le fichier d'annotation au format GFF3 référencé lors du dépôt est analysé de façon à faire remonter le maximum d'informations jusqu'à l'élément le plus haut (gène).

Le fichier de résultats normalisés/transformés et annoté vient enrichir une base de données spécifique à chaque version d'annotation.

```
java -jar  mergecsv-1.0.jar
 --rnaseq --mode normalize
 --data MesureExpression.RepTech.data.tsv
 --metadata MesureExpression.RepTech.metadata.tsv
 --refannot MtrunA17r5.0r1.6-ANR.gff3.gz
 --refseq MtrunA17r5.0-ANR.fasta.gz
 --idtype locus_tag
```

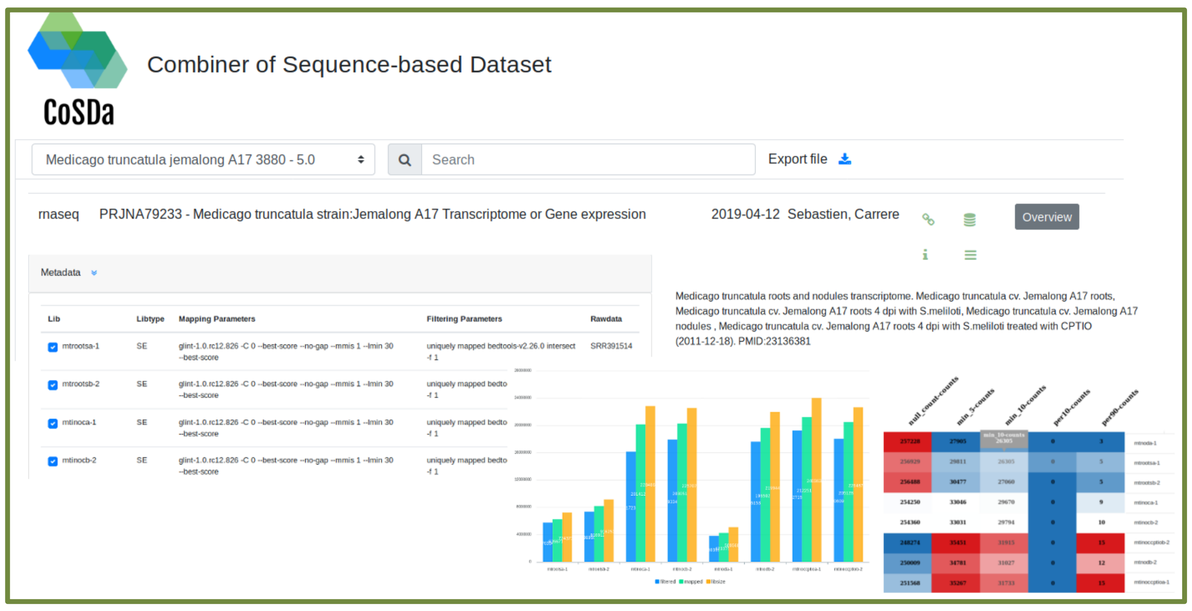
#uniqueId	Name	locus_tag	product	type	length	rnaseq-medtr-t0-c2-count	rnaseq-medtr-t0-c2-cpm	rnaseq-medtr-t0-c2-tpm	rnaseq-medtr-t0-c2-rpkm
MtrunA17Chr1:gene:87295:91102:+	MtrunA17Chr1g0146181	MtrunA17_Chr1g0146181	Putative xylogalacturonan beta-1,3-xylosyltransferase	gene	3808	179	8,42	10,04	2,81

MISE À DISPOSITION DES RÉSULTATS

Une interface web permet d'interroger les résultats ainsi consolidés.

Un aperçu sous forme de statistiques descriptives vient aider l'utilisateur à contrôler la qualité des résultats et identifier des sous-ensembles de résultats pertinents.

Ces sous-ensembles sélectionnés sont exportés afin d'être intégrés dans de nouveaux travaux d'analyse.



Références
1. http://www.agence-nationale-recherche.fr/Project-ANR-15-CE20-0012
2. Y. Pecrix, S. Evan Staton, E. Sallet, C. Lelandais-Brière., Whole-genome landscape of *Medicago truncatula* symbiotic genes, Nature Plants 4, pages1017–1025 (2018)
3. https://bbric.toulouse.inra.fr/reference
4. Wagner, G.P., Kin, K. & Lynch, Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples, V.J. Theory Biosci. 2012