DéBAT : Analyse de diversité d'un panel de pré-breeding de blé tendre par une approche transcriptomique

Hélène RIMBERT [1], Frédéric CHOULET [1], Odile ARGILLIER [2], Jérôme AUZANNEAU [3], Mark DAVEY [4], Philippe DUFOUR [5], Sylvie DUTRIEZ [6], Pascal GIRAUDEAU [7], Ellen GOUDEMAND-DUGUE [8], Gemma MOLERO [9], Mickaël THROUDE [10], Hervé DUBORJAL [10], Adeline CLEMENTI [10], David GRIMBICHLER [11], Etienne PAUX [12], Sophie BOUCHET [1]*

1 - INRAE Université Clermont-Auvergne , UMR 1095, GDEC, 5 chemin de Beaulieu, 63100 Clermont-Ferrand, FRANCE
2 - Syngenta France SA, 2 avenue Gustave Eiffel, F-28000 Chartres, FRANCE
3 - AGRI-Obtentions, Chemin de la Petite Minière, 78280 Guyancourt, FRANCE
4 - BASF Innovation Center Gent, Technologiepark-Zwijnaarde 101, 9052 Gent, BELGIQUE
5 - RAGT, Rue Emile Singla, BP 3331 12033 Rodez Cedex 9, FRANCE
6 - Lidea Seeds, avenue Gaston Pheobus, 64230 Lescar, FRANCE
7 - Secobra Recherches, Centre de Bois-Henry, 78580 Maule, FRANCE
8 - FLORIMOND DESPREZ VEUVE & FILS, 59242 Cappelle-en-Pévèle, FRANCE
9 - KWS MOMONT SAS, 7 Rue de Martinval, 59246 Mons-en-Pévèle, FRANCE
10 - Limagrain Europe, Centre de recherche de Chappes, 63720 Chappes, FRANCE
11 - UCA, Plateforme AuBi & Mésocentre Clermont-Auvergne, 63000 Clermont-Ferrand, FRANCE
12 - VetAgro Sup, 89 Avenue de l'Europe, CS 82212, 63370 Lempdes, FRANCE
__Coordinatrice : Sophie BOUCHET,__ sophie.bouchet@inrae.fr

Introduction
Dans le cadre du projet DéBAT, nous avons produit un catalogue de l'ensemble des gènes exprimés dans les lignées d'un panel de 450 lignées représentatives de la diversité mondiale par la méthode RNA-seq. Douze lignées représentatives ont été séquencées en forte profondeur sur 3 tissus (tige, épi et feuille). Des niveaux d'expression différenciels ont été identifiés. Le reste du panel a été séquencé sur le mélange des 3 tissus. Ceci nous a permis d'identifier la présence absence (PAV) d'isoformes de pour 60K gènes annotés High Confidence sur la séquence de référence Chinese Spring et 83K gènes absents de cette séquence. Leur fonction a été prédite. Des analyses d'association ont été conduites sur les PAV et les SNP obtenus avec les phénotypes évalués dans le cadre du projet compagnon Ex-IGE.

Echantillonnage des lignées (Figure 1)-
Volet 1 - 12 lignées représentatives du panel BWP3 du projet Breedwheat ont été sélectionnées (Figure 1B) pour un séquençage RNA-seq forte profondeur (2 x 115 millions de reads, NovaSeq6000).
Volet 2 - Le transcriptome des 450 lignées du BWP3 (Figure 1A) a été séquencé en plus faible couverture (2x25 millions de reads) .

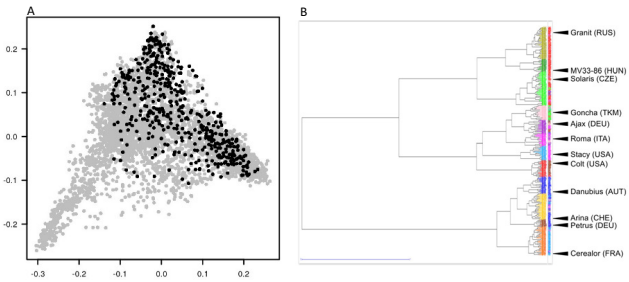


Figure 1 - (A) PcOA de 4500 lignées du CRB, représentatives de la diversité mondiale. Les points noirs correspondent aux 450 lignées du projet, adpatées à une évaluation en France; (B) Dendrogramme de Ward des 450 lignées. D'après Balfourier et al. (Science Adv 2019)

Pipeline d'analyses (Figure 2)-
Volet 1 – Le transcriptome des 12 lignées de reference a été sequencé dans trois tissus (tige, feuille, épi). Un pan-transcriptome de reference avec 400K reads unique a été construit.
Volet 2 - Le transcriptome des 450 lignées du BWP3 a été séquencé en mélange. Les reads ont été mappé sur le pan-transcriptome de référence construit dans le volet 1.

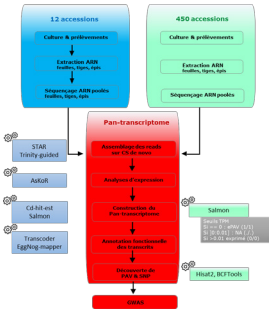
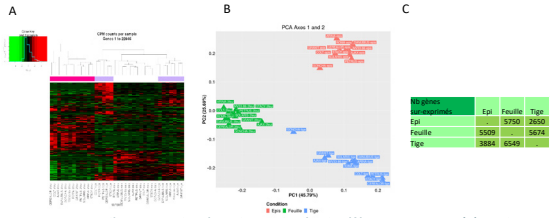


Figure 2: ***Pipeline d'analyses***

Analyse d'expression (Figure 3)-
Volet 1 – Les gènes sont différentiellement exprimés selon les tissus.



Nb gènes sur-exprimés	Epi	Feuille	Tige
Epi		5750	2650
Feuille	5509		5674
Tige	3884	6549	

Figure 3 – Niveaux d'expression des gènes dans 12 lignées de référence et 3 tissus ; (A) Dendogramme ; (B) ACP; (C) Nombre de gènes sur-exprimés dans les compartiments situés en ligne
exemple: 5750 gènes sont sur-exprimés dans l'épi par rapport à la feuille

Mapping des transcrits -
Volet 1 – Un **pan-transcriptome de 400 288 transcrits a été construit à partir des 12 lignées de référence**.
En moyenne, 300 millions de reads ont été séquencés par lignée de reference representant 400K transcrits uniques. Ils correspondent à 60K gènes annotés chez CS (HC ou LC), soit 51% des gènes annotés (Figure 4A) et 84K nouvelles protéines dont nous avons prédit la fonction. Un nombre de reads equivalent est exprimé dans tous les tissus.
Volet 2 – La plupart des 400K pan-transcrits de reference (99%) sont exprimés dans au moins une des 450 lignées (Figure 3B2). En moyenne, 186k transcrits sont exprimés dans chaque lignée (Figure 4B1). Au total, 60K gènes sont exprimés en moyenne par lignée (Figure 4B2).

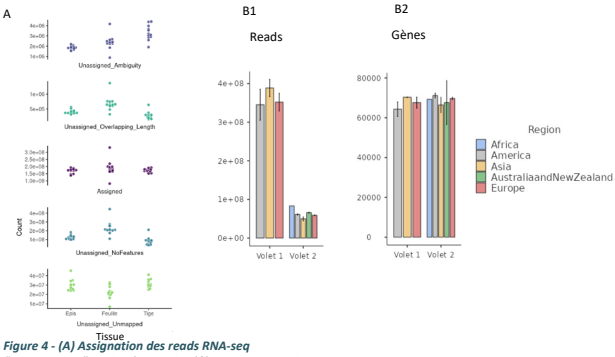


Figure 4 - (A) Assignation des reads RNA-seq "Featurecount" sur le génome de référence IWGSC - Chinese spring.
Unassigned_Ambiguity: les 2 paires de reads ne mappent pas sur le même chromosome; Unassigned_Overlapping_Length: les deux paires de reads se chevauchent (minimum de 100 nucléotides de séparation autorisé) ; Assigned: mappent sur des régions de Chinese-Spring annotées ; Unassigned_NoFeatures: mappent sur des régions de Chinese-Spring non annotées; Unassigned_Unmapped: ne mappent pas sur Chinese-Spring;
(B) Nombre de reads et de gènes presents par zone géographique

Pan-transcriptome (Figure 5):-
Chaque transcrit a été attribué aux compartiments "core" (plus de 90% des lignées partagent le transcript), "shell" (entre 10 et 90%) ou "cloud" (moins de 10%). Au total, 71% des transcrits sont présents dans le "shell", 17% dans le "core" et 11% dans le "cloud". En moyenne, un transcript est présent chez la moitié des lignées du panel.

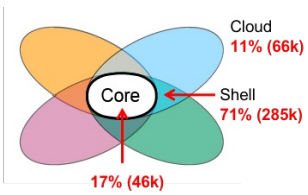


Figure 5 – Distribution du pan-transcriptome

Analyse fonctionnelle - La construction du pan-transcriptome a permis l'identification de **84714 nouvelles protéines** pour lesquelles une fonction a été prédite (Transdecoder, eggNOG-mapper). L'enrichissement en GO de ces protéines (R-topGO) a montré une prévalence pour les processus adaptatifs (stress biotiques, abiotiques, lumière, température, stress osmotique ...).

ePAVs et SNP calling–
Nous avons utilisé un seuil de 0,01 TPM pour déclarer un transcrit présent dans une lignée (Figure 6). Les transcrits avec un TPM compris entre 0 et 0,01 ont été déclarés en données manquantes. Après filtre sur données manquantes (<0,1) nous avons gardé 362 899 marqueurs ePAV.
Un SNP calling a été fait parmi les 400K transcrits. Après filtre sur données manquantes, nous avons gardé 83K SNP.
Les Minimum Allele Frequency (MAF) sont équilibrées pour les ePAV. On observe un excès d'alleles rares pour les SNP (Figure 7). En moyenne, les SNPs comptent 8% de données manquantes, et les ePAV présentent très peu de données manquantes (au maximum 2% par marqueur).
Les SNP issus des ePAV discriminent les mêmes groupes génétiques que les SNP Breedwheat (Figure 8). Par contre les ePAV discriminent des groupes différents.

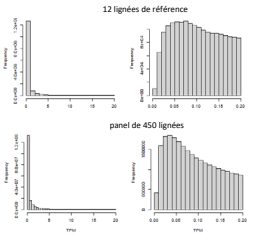


Figure 6 – Distribution des TPM

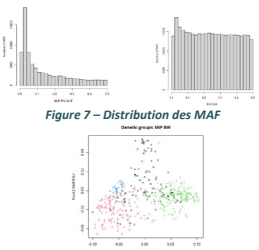


Figure 7 – Distribution des MAF

Figure 8 – PCoA construite avec les SNP issus de ePAV
Les couleurs correspondent aux groupes génétiques BW (Balfourier et al, 2019)

Détection automatique d'introgressions –
Une analyse est en cours pour détecter automatiquement les introgressions. Par exemple l'introgression de seigle 1BS/1R est déctectée chez Solaris avec une expression nulle des gènes de blé tendre sur le bras court (Figure 9).

Perspectives –
Des études d'associations ont été conduites avec les marqueurs produits dans ce projet et les phenotypes produits de le projet compagnon Ex-IGE. En éliminant les marqueurs SNP avec beaucoup de données manquantes, nous avons probablement éliminé tous les marqueurs avec trois allèles (absence et presence avec SNP). Il serait intéressant d'exploiter toutes les données Breedwheat et DéBAT pour valider notre méthode de detection sur des introgressions connues. Nous regarderons également si parmi les reads éliminés ou "unmapped", certains correspondent à des reads d'espèces apparentées (ventricosa, thinopyrum, dicoccoides, monococcum, speltoides, seigle...).

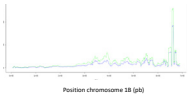


Figure 9 - Profil d'expression des gènes du chromosome 1B chez Solaris
Les deux réplicats de RNA-seq sur feuille sont représentés, avec une moyenne en fenêtre glissante