DéBAT : Analyse de Diversité d'un panel de pré-breeding de Blé tendre par une approche Transcriptomique-projet FSOV 2018 M

Hélène Rimbert, Frédéric Choulet, Odile Argillier, Jerome Auzanneau, Mark Davey, Philippe Dufour, Sylvie Dutriez, Pascal Giraudeau, Ellen Goudemand-Dugué, Gemma Molero, et al.

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], Michael THROUDE [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 - Université Clermont Auvergne, Plateforme AuBi and Mésocentre Clermont-Auvergne, F-63000 Clermont-Ferrand, FRANCE
12 - VetAgro Sup, 89 Avenue de l'Europe, CS 82212, 63370 Lempdes, FRANCE

*__Coordinateur : Sophie BOUCHET,__ *sophie.bouchet@inrae.fr*

1 Introduction

Développer de nouvelles variétés de blé plus résistantes aux stress biotiques et abiotiques est un défi majeur de la sélection d'aujourd'hui pour l'agriculture de demain. Pour cela, les ressources génétiques constituent un réservoir de gènes et d'allèles encore largement inexploités.

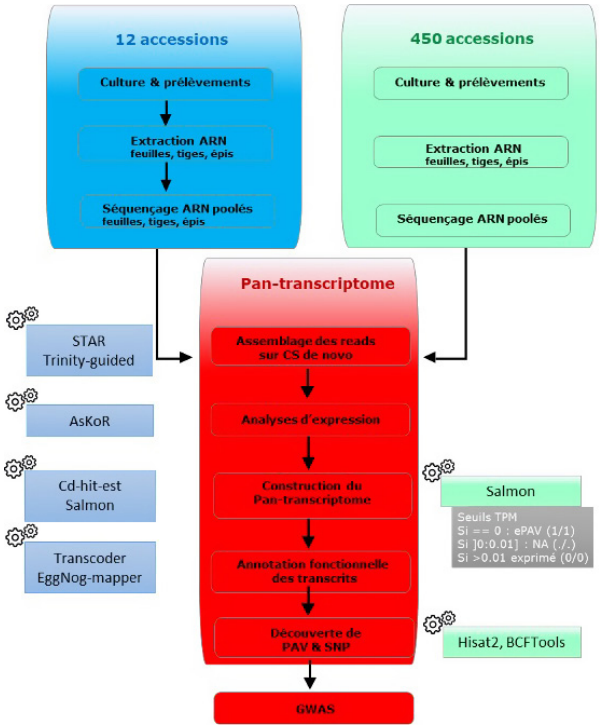


Figure 1 : *Organisation du projet DéBAT. En bleu et vert, les volets expérimentaux sur les panels de 12 et 450 lignées respectivement. En rouge, le volet d'analyses bioinformatiques et génétiques.*

Dans le cadre du projet Investissements d'Avenir BreedWheat, un panel de 450 lignées issues de la diversité mondiale et adapté à une culture en France a été sélectionné. Ce panel, appelé BWP3, a fait l'objet de caractérisations au niveau moléculaire (génotypage 350K SNP, capture d'éxome de 2000 gènes candidats) et phénotypique (rendement en condition optimale ou stressée, résistance aux maladies majeures, qualité protéique et boulangère). Le séquençage du génome entier reste une option trop coûteuse en raison de la taille du génome du blé tendre. Une alternative consiste à étudier le génome exprimé par l'intermédiaire du transcriptome.

Dans le cadre du projet DéBAT, nous avons produit un catalogue de l'ensemble des gènes exprimés dans les lignées du panel BWP3 par la méthode RNA-seq (Figure 1). Nous avons séquencé à forte profondeur les transcrits de 12 lignées représentatives de la diversité. Nous avons pu comparer les niveaux d'expression des gènes dans 3 tissus différents (tige, épi et feuille). 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 protéines sur l'ensemble du panel et de typer les SNP présents dans ces séquences. Des analyses d'association ont été conduites sur les PAV et les SNP avec les phénotypes évalués dans le cadre du projet compagnon Ex-IGE porté par Limagrain et INRAE (Figure 1).

2 Matériel et méthode

▸ Sélection de douze lignées représentatives du panel BWP3

Sur la base des résultats obtenus dans le cadre du projet BreedWheat, un arbre phylogénétique des 450 lignées du panel BWP3 a été construit et nous a permis de définir douze groupes génétiques (Figure 2). Les groupes 1 à 6

correspondent principalement à des variétés d'Europe de l'Est et du Sud ; les groupes 7 et 8 à des variétés d'Amérique du Nord ; les groupes 9 à 12 à des variétés d'Europe de l'Ouest. Nous avons ensuite sélectionné un représentant pour chacun des douze groupes (Figure 2).

Ces douze lignées sont utilisées dans le volet 1 du projet pour des analyses transcriptomiques fines.

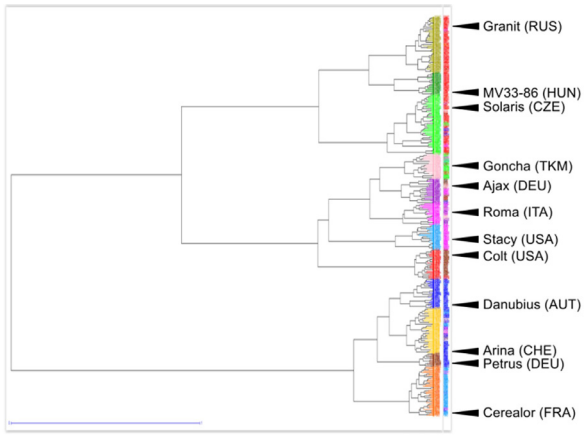


Figure 2 : *Dendrogramme de Ward montrant les relations phylogénétiques entre les 450 lignées du panel BWP3. D'après Balfourier et al 2019.*

▶ Culture et prélèvements

Les neuf partenaires privés (AgriObtentions, BASF, Caussade Semences, Florimond Desprez, KWS, Limagrain, R2N, Secobra, Syngenta) ont mis en culture 50 lignées du panel BWP3 (Annexe 1). Ces lots ont été constitués de manière à homogénéiser autant que possible les précocités des lignées pour un même partenaire (Figure 3A). L'écart moyen entre les lignées les plus précoces et les plus tardives d'un lot est d'environ 25 degrés jour, soit environ deux jours (Figure 3B). En parallèle, Biogemma a reçu les semences des 12 lignées représentatives des groupes.

Le Centre de Ressources Biologiques « Céréales à Pailles » d'INRAE GDEC a distribué les graines du panel. Elles sont issues d'autofécondations contrôlées et correspondent aux mêmes lots que ceux utilisés dans les projets BreedWheat et EX-IGE. Les plantes ont été cultivées en serre dans les mêmes conditions chez les différents partenaires. Des échantillons ont été prélevés à partir de trois organes à des stades de développement différents pour maximiser la diversité des transcrits : feuilles au stade trois feuilles (Zadoks Z13), tige au stade floraison (Z65) et épillet au stade floraison (Z65) (Annexe 1).

▶ Extraction d'ARN et construction des banques

Les ARN totaux ont été extraits par Biogemma à l'aide du kit Promega SV 96 total RNA Isolation System. Dans le volet 1, pour les 12 lignées, chacun des trois échantillons de tissus fait l'objet de deux extractions. Dans le volet 2, une seule extraction a été réalisée par échantillon de tissu. La qualité des ARN extraits a été évaluée par électrophorèse capillaire sur un système LabChip 90 de Caliper (Figure 4). Pour les feuilles et les tiges, la qualité est correcte, comme montré par le score RIN (RNA Integrity Number) proche voire supérieur à 7. Pour les épillets, les profils montrent un début de dégradation, avec un score RIN de l'ordre de 5.

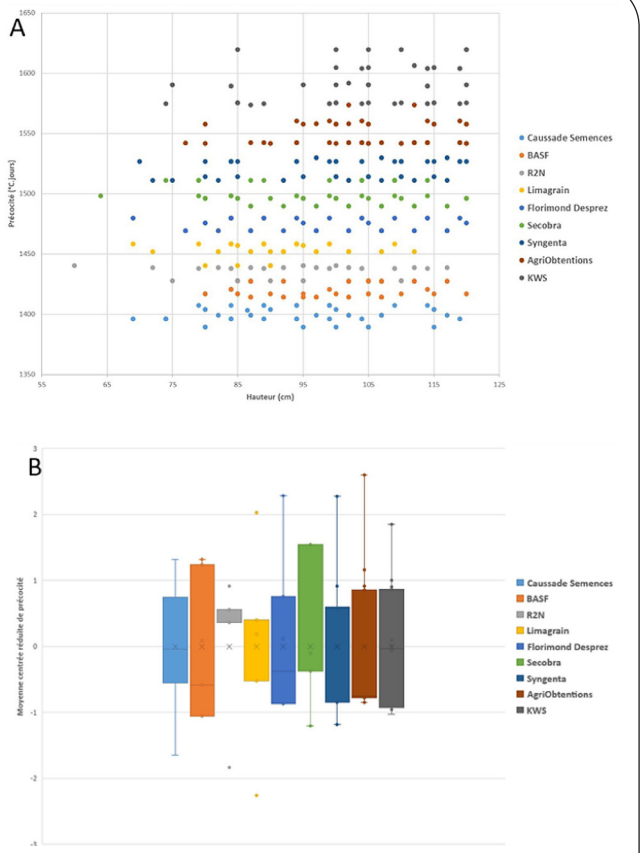


Figure 3 : *Répartition des 450 lignées du panel BWP3 aux neuf partenaires. (A) Distribution des lignées en fonction de leur hauteur et de leur précocité. (B) Moyennes centrées réduites de précocités pour chaque lot de lignées.*

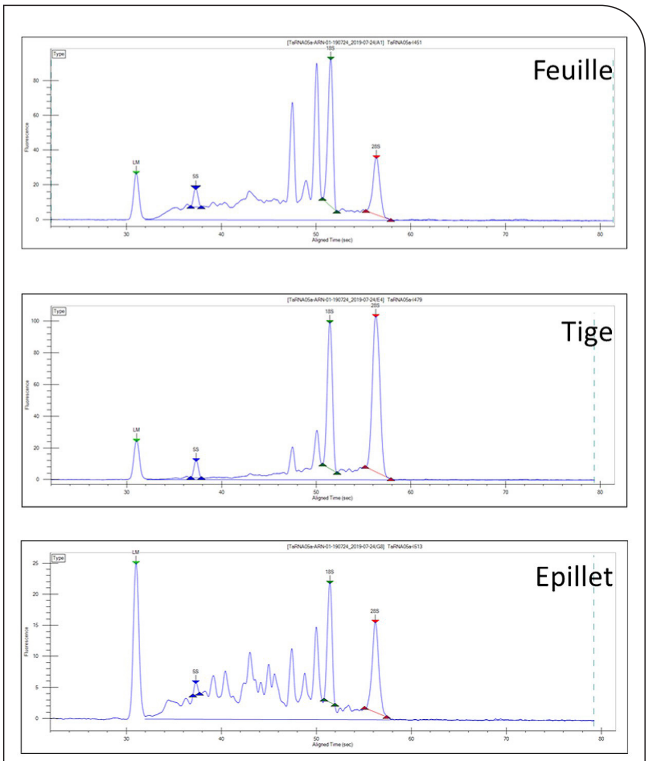


Figure 4 : *Qualité des ARN extraits par électrophorèse capillaire. Un échantillon par tissu analysé sur LabChip 90 de Caliper est représenté.*

▶ Séquençage

Les banques ont été séquencées sur Illumina NovaSeq6000 qui séquence en paires de lectures (2x150 pb). L'ensemble des extractions ont été séquencées individuellement pour les 12 lignées de référence, et poolées pour le panel. Pour le volet 1, le nombre total de banques est donc de 72 12 lignées x 3 échantillons x 2 réplicats). Dans le volet 2, pour chacune des 450 lignées, un pool équimolaire des ARN totaux provenant des différents tissus a été réalisé avant construction des banques. Chaque banque de RNAseq a été indexée par une séquence unique, permettant un pooling avant séquençage. Pour le volet 1, une moyenne de 2 x 115 millions de lectures et 35 Mb ont été produites, et pour le volet 2, une moyenne de 2 x 59 millions de lectures et 18 Mb (Figure 5A). Tous les échantillons ont une très bonne qualité (score Phred > 30) avec une moyenne supérieure à 34 (Figure 5B).

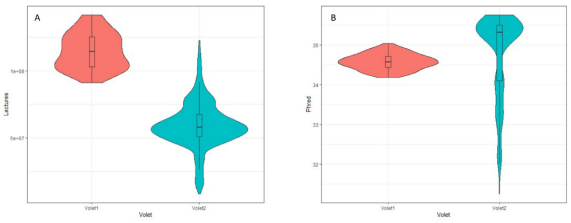


Figure 5 : *Séquençage.*
(A) Nombre de lectures pour les volets 1 (rouge) et 2 (bleu).
(B) Qualité des lectures (score PHRED) des volets 1 et 2.

▶ Développement d'un pipeline d'analyse

Ces analyses ont nécessité des ressources de calcul exceptionnelles. Nous avons développé un pipeline automatisé et optimisé sur le cluster de calcul HPC2 ainsi que sur l'infrastructure cloud du Mésocentre de Clermont-Ferrand OSCAR (OpenStack Cloud en Auvergne Rhône-Alpes, Figure 6). Ce pipeline inclut des étapes de nettoyage des extrémité de reads de mauvaise qualité avec Trimmomatic (Bolger *et al.* 2014), d'alignement sur transcrits par l'outil Salmon (Patro *et al.* 2017), d'alignement sur génome entier avec STAR (Dobin *et al.* 2013), d'assemblage de transcrits de novo avec Trinity (Grabherr *et al.* 2011) et de clustering de transcrits avec cdhit (Li and Godzik 2006) (Annexe 3). Les scripts sont écrits en BASH et sont utilisés dans des environnements informatiques dédiés grâce à Conda. Ils sont disponibles sur la forge GitLab d'INRAE (https://forgemia.inra.fr/fsov-debat/debat-bashpipeline).

▶ Construction du pan-transcriptome

Nous avons conduit un alignement préliminaire des données de séquençage des douze lignées du volet 1 sur la séquence de référence de Chinese Spring (CS) v1.1 à l'aide l'outil Kallisto. Les lectures de séquençage ont ensuite été alignées sur le génome de référence Chinese Spring (IWGSC refseq v2.1 (Zhu *et al.* 2021)) avec STAR. Les alignements des réplicats ont ensuite été fusionnés avec SAMTools merge (Li *et al.* 2009). Trinity a ensuite été utilisé en mode « genome guided » pour tirer profit des informations d'ancrage sur CS lors de l'assemblage des transcrits. Toutes les lectures non « mappées » ont été assemblées à part. Les transcrits obtenus ont ensuite été comparés entre échantillons pour obtenir un jeu de données non redondant. Pour cela, nous avons utilisé une méthode de « clustering » de séquences (CD-HIT-EST) en spécifiant les seuils suivants : >= 99% d'identité nucléotidique sur >= 90% de la longueur des transcrits. Par ailleurs, les transcrits < 500 bp ont été éliminés à cette étape. Enfin, les lectures Illumina ont été « re-mappées » sur ce set de transcrits avec Salmon. Le but était ici d'éliminer les contigs qui correspondent à des résidus de transcription (transcrits douteux, peu reproductibles) et aux erreurs dans le processus d'assemblage par Trinity.

▶ Caractérisation du pan-transcriptome chez les 450 lignées

Les lectures de séquençage produites dans le volet 2 ont été nettoyées avec Trimmomatic puis ont été assignées au pan-transcriptome à l'aide de Salmon. Salmon réalise un pseudo-mapping et non un alignement comme les outils de mapping classiques et nous donne directement les valeurs de Transcripts Per Million (TPM) pour chacune des lignées. Un ePAV (Présence Absence d'expression) a été considéré absent d'une lignée lorsque la valeur de TPM du transcrit est de 0. Les transcrits ayant une valeur de TPM > 0.01 sont considérés comme présents. Tous les transcrits pour lesquels la valeur de TPM était comprise entre 0 et 0.01 ont été convertis en données manquantes. Cela nous permet de générer une matrice de ePAVs pour les 400k transcrits et les 450 accessions. Une recherche de polymorphisme sur ces ePAV a produit une matrice 83K SNP supplémentaires sur le panel. Ces données sont accessibles aux partenaires du projet sur le dépôt RechercheDataGouv (RDG) (https://doi.org/10.57745/GQZF6L).

▶ Analyse fonctionnelle des gènes

Les gènes différentiellement exprimés entre tissus ont été recherchés avec AskoR(askomics/askoR) sur les 12 lignées de référence. Les contrastes d'expression entre tissus ont été évalués avec DESeq2. Les gènes correspondants aux transcrits ont été prédits grâce à l'annotation de CS et l'outil Mikado (compare). Les protéines des transcrits non annotés ont été traduites avec transdecoder. Les fonctions de ces protéines ont été prédites à partir de plusieurs bases de données de gènes orthologues (PFAM domains, KEGG pathways, Gene Ontology annotation) avec eggnog-mapper (diamond). Une étude d'enrichissement en gene ontology pour les gènes non annotés a été effectuée avecR-topGO.

▶ Infrasctructure de calcul et stockage

Les analyses ont été réalisées sur le cluster de calcul HPC2 du Mésocentre Clermont-Auvergne (UCA). Une infrastructure cloud dédiée à été déployée spécifiquement pour le projet en collaboration avec la plateforme AuBI. Les codes Ansible sont disponibles (https://github.com/HeleneRimbert/oscar-triannot).

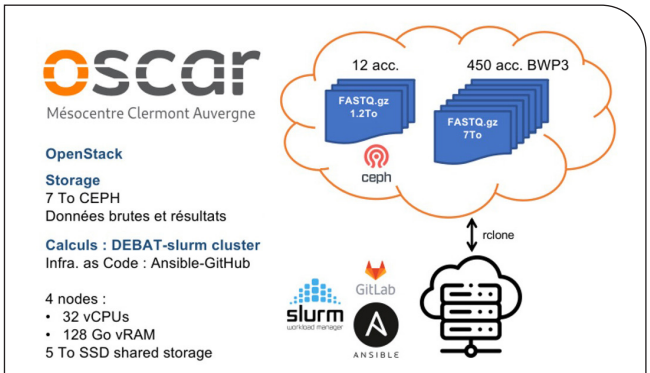


Figure 6 : *Infrastructure cloud sur OSCAR - Les données ont été stockées sur le cluster CEPH du mésocentre. Un cluster de calcul SLURM a été virtualisé sur OSCAR et déployé en utilisant Ansible.*

▶ Disponibilité des données

Les données du projet (matrices de ePAVs, SNPs, annotations fonctionnelles) sont disponibles pour les partenaires du projet sur l'entrepôt national RechercheDataGouv.
DOI : https://doi.org/10.57745/GQZF6L

3 Résultats

▶ Analyses préliminaires

Sur la base de précédentes analyses, nous estimons que les trois échantillons de tige, épis et feuille devraient nous permettre d'accéder à 82 % des gènes exprimés au cours du développement normal d'un plant de blé et 59 % des gènes codants des protéines (Pingault *et al.* 2015). D'après l'analyse préliminaire avec Kallisto, en moyenne, pour une variété donnée, 81.3 % des lectures sont alignées sur Chinese-Spring (CS) (Figure 7). Une expression supérieure à 1 TPM est détectée pour 58 430 gènes HC (« high confidence », sur ~107k prédits), soit environ 54 % des gènes HC annotés sur la séquence de référence. Ces résultats sont très proches de l'attendu (59 %). En prenant l'ensemble des 12 variétés, une expression est trouvée dans au moins un tissu d'une variété pour 72 998 gènes, soit 68 %. Il est intéressant de noter que lorsque l'on étudie la distribution des niveaux d'expression des gènes le long des chromosomes, il semble possible de détecter des introgressions. C'est notamment le cas pour la variété Solaris pour laquelle le profil de niveaux d'expression est nul sur le 1BS (Figure 8).

Ceci peut s'expliquer par la translocation 1RS-1BL présente dans cette variété, impliquant le remplacement du bras court du chromosome 1B de blé par le bras court du chromosome 1R de seigle.

▶ Expression différentielle du volet 1

Les gènes sont différentiellement exprimés entre tissus sur les 12 lignées de référence (Table 1).

Les profils d'expression des échantillons clusterisent par tissus et non par phylogénie des accessions (Figure 9, 10).

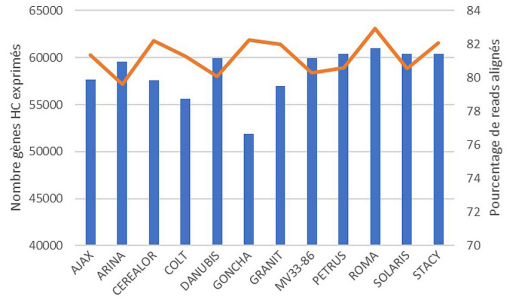


Figure 7 : *Reads et gènes alignés sur le RefSeq v1.1. % de reads en orange, nombre de gènes HC exprimés (> 1 TPM) en bleu.*

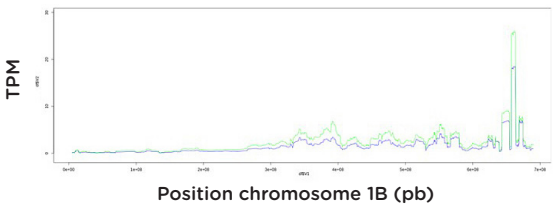


Figure 8 : *Profil d'expression des gènes du chromosome 1B chez Solaris. Les deux réplicats de RNA-seq pour le tissu feuille sont représentés (moyenne sur fenêtre glissante).*

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

Table 1 : *Nombre de gènes différentiellement exprimés chez les 12 lignées. Le tableau se lit des lignes vers les colonnes. Exemple : 5509 gènes sont surexprimés dans la feuille par rapport à l'épi.*

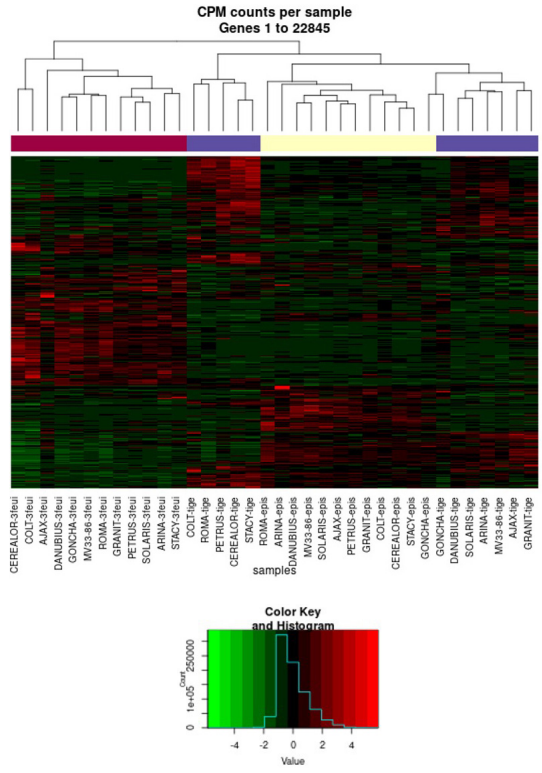


Figure 9 : *Clustering des données d'expression du des 12 lignées de référence. Réalisé avec AskoR.*
Les échantillons en magenta dans le dendrogramme correspondent aux feuilles, le violet au tiges et le jaune aux épis.

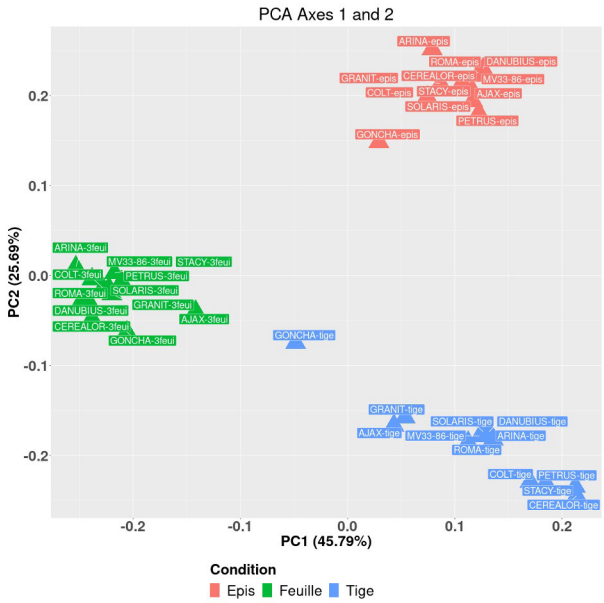


Figure 10 : *ACP des données d'expression des 12 lignées de référence dans 3 tissus.*

▶ Construction du pan-transcriptome

Au total, 331k transcrits ont été assemblés en moyenne pour chacun des 36 échantillons (Annexe 4) représentant 369 Mb.

Le clustering a permis d'obtenir un set non redondant de 7 669 122 transcrits. Ce nombre élevé s'explique par le fait que l'assemblage de novo de lectures Illumina sur transcriptome génère beaucoup de transcrits fragmentés (1 transcrit assemblé en plusieurs contigs). Par ailleurs, une part significative de transcrits correspondent à des ARN non codants, encore peu étudiés chez le blé. L'étape de re-mapping des reads sur la collection de 7M de transcrits a permis de réduire à 317 496 transcrits identifiés dans au moins 3 échantillons. Au total, 69 022 transcrits correspondent à 60 585 gènes HC prédits chez Chinese Spring, soit 56 % des gènes HC prédits.

Nous avons construit un pan-transcriptome de référence en utilisant les transcrits des 12 lignées de référence :

- prédits comme gènes chez CS (IWGSC RefSseq v2.1 : **110 909** transcrits) ;
- présents chez CS mais non prédits comme gène (**224 843**) ;
- absents chez CS (**46 935**).

Ce pan-transcriptome représente au total **382 687 transcrits potentiels**. Les transcrits du panel BWP3 ont été caractérisé par rapport à ces transcrits.

▶ Annotation fonctionnelle

Parmi les 271 778 transcrits non prédits, nous avons identifié 84741 ORFs complets d'une taille supérieure à 80 acides-aminés. Les protéines correspondantes ont été traduites avec transdecoder. Les peptides obtenus ont été comparés à une base de données de gènes orthologues. Les fonctions des proteines ont été prédites à partir des fonctions des membres du cluster d'orthologues correspondant (Table 2).

Bases de données	Nombre de protéines
PFAM	76 657
KEGG	7 321
Gene Ontology	41 955
Description littérale	78 952

Table 2 : *Annotation des protéines non annotées chez Chinese Spring.*

Parmi les gènes non annotés chez CS, d'après un test d'enrichissement en Gene Ontology (GO), on trouve un excès de gènes impliqués dans les processus liés à la réponse à l'environnement (stress biotiques, abiotiques, réponse à la lumière, température, stress osmotique, Annexe 6).

▶ Diversité des 400k pan-transcrtis

Après pseudo-alignement des reads RNA-seq du panel BWP3 avec Salmon contre les 400k transcrits de référence, avec un seuil de 0.01 TPM pour définir un transcrit comme présent chez une lignée (Figure 12), on obtient 186k transcrits exprimés en moyenne par accession et 396K transcrits exprimés dans au moins une variété (99 % du 400k PAN-Transcriptome).

En fonction du caractère rare ou conservé, les transcrits ont été classés en « core » (transcrit présent chez plus de 90 % des lignées), « shell » (de 10 % à 90 % des lignées) et « cloud » (moins de 10 % des lignées). Seuls 46k transcrits sont partagés par plus de 90% des lignées (11%), 285k appartiennent au « shell » (71%) et 66k sont rares (17 %). Voir Figure 11.

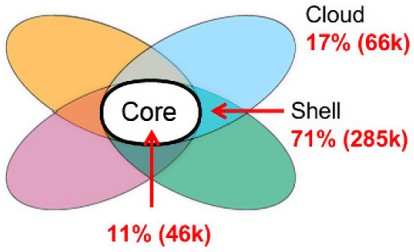


Figure 11 : *Pan-transcriptome.*

▶ Diversité génétique

Un alignement des reads RNA-seq a été réalisé sur le génome de Chinese spring v2.1 avec Salmon. Au total, 576k SNPs sont localisés dans des transcrits ayant une position physique couvrant 60k gènes annotés. Parmi eux, 83K avec une MAF > 0.01 et un nombre de données manquantes inférieur à 10 % ont été conservés pour les études d'association.

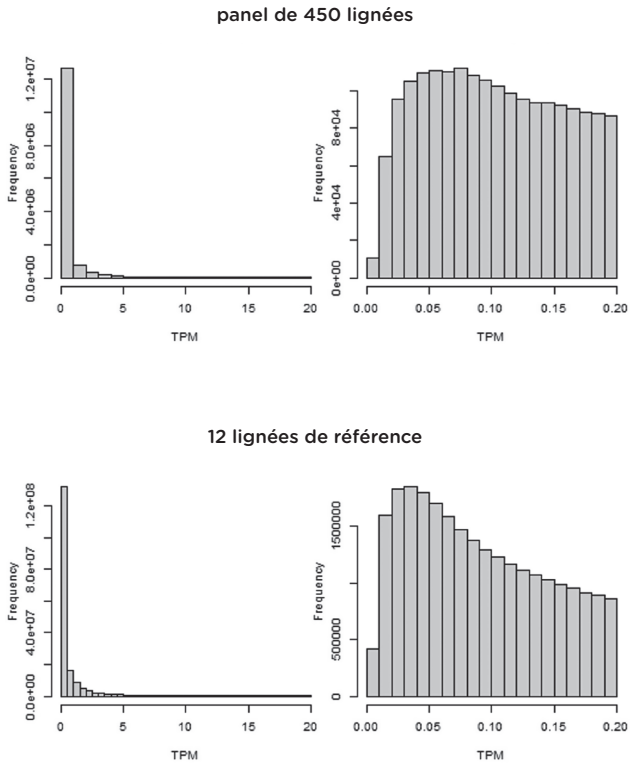


Figure 12 : *Distribution des TPM.*

Les Minimum Allele Frequency (MAF) sont équilibrées pour les ePAV. On observe un excès d'alleles rares pour les SNP (Figure 13). 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 14). Par contre les ePAV discriminent des groupes différents.

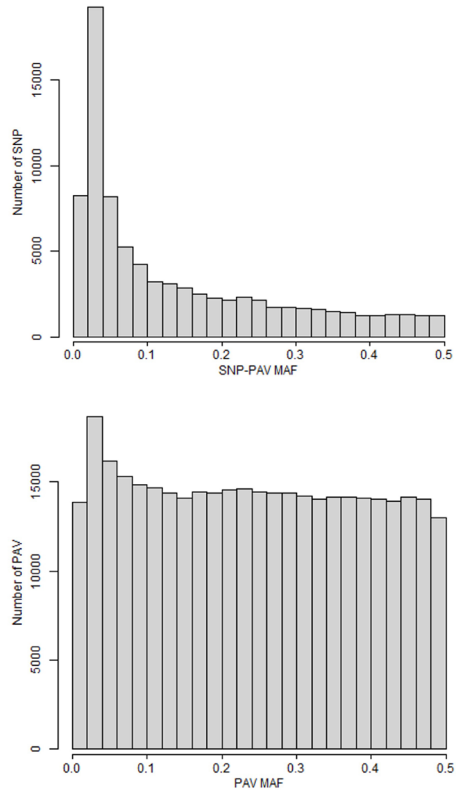


Figure 13 : *Distribution des MAF.*

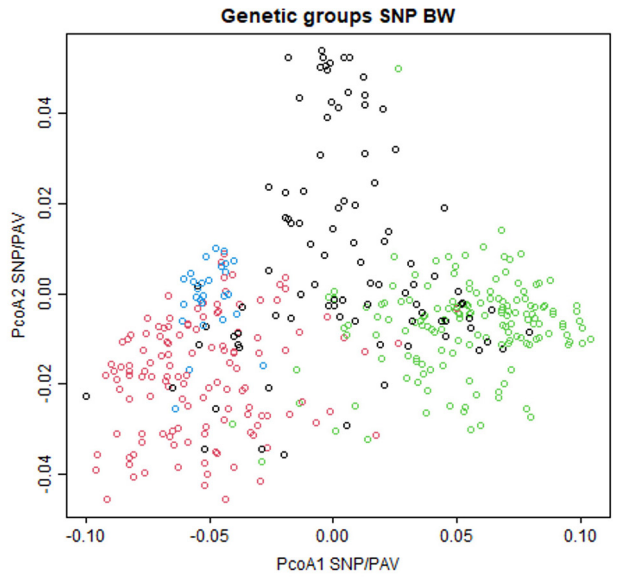


Figure 14 : *PcOA construite avec les SNPs issus d'ePAV. Les couleurs correspondent aux groupes génétiques BW (Balfourier et al, 2019)*

4 Discussion

Ce projet a permis de décrire la diversité d'expression dans 3 tissus (tige, feuille, épi) de 12 lignées représentatives de la diversité mondiale de blé tendre et d'identifier des protéines différentiellement exprimées. Pour le panel, nous avons produit une matrice de PAV de 360K transcrits. Elle a permis également de produire 83K SNP avec 8% de données manquantes.

Nous avons identifié 84K protéines non annotées chez CS dont nous avons inféré la fonction. On observe un enrichissement dans des fonctions adaptatives pour ces gènes.

▶ Construction du PAN-transcriptome

Nous avons développé une méthode hybride d'assemblage des reads RNA-seq en associant mapping avec STAR et assemblage de novo avec Trinity. Cette méthode permet de minimiser la création de fragments de gènes, seuls les reads non alignés étant assemblés de novo. Cependant, le volume de transcrits générés est très élevé (plus de 6,7 millions) et le clustering utilisé ne divise que par deux ce volume (3,8 millions). La nature hexaploïde du blé et le faible taux de divergence entre les 3 sous-génomes A, B et D étant faible (~2%), il nous est apparu indispensable de garder un fort taux d'identité (99%) afin de ne pas rassembler au sein d'un même cluster les transcrits issus de gènes homéologues.

▶ Caractérisation du pan-transcriptome

Nous avons identifié dans ces travaux une faible proportion de transcrits partagés par plus de 90% des accessions. De manière globale, seule la moitié des 400k transcrits sont exprimés pour chacune des 450 lignées. Il est possible que les paramétrages du clustering des données du volet 1 soit trop stricts et ne permettent pas de rassembler, au sein d'un même cluster, tous les isoformes d'un même gène. Cela à pour conséquence d'éclater l'information d'un même gène en autant d'isoformes portées par les différentes lignées, sous estimant alors le volume de gènes/transcrits partagés.

▶ GWAS

Les résultats des études d'association entre ces variants SNP et ePAV et des phénotypes de rendement, maladie et qualité protéique sont présentés dans le projet FSOV compagnon Ex-IGE.

▶ Détection d'introgressions

Vue la faible profondeur de séquençage, nous essayons de vérifier si les absences (très nombreuses) sont de réelles absences. Nous sommes en train de vérifier si les SNP issus des ePAV avec plus de 10% de données manquantes ne correspondent pas justement aux régions avec un réel 3ème allèle « absent ». Nous comparerons la méthode de SNP calling et une méthode de fragmentation sur profil d'expression pour identifier de grosses introgressions connues. Nous utiliserons la meilleure méthode pour typer des introgressions sur l'ensemble du génome.

Nous regarderons également si parmi les reads éliminés ou « unmapped », certains correspondent à des reads d'espèces apparentées (ventricosa, thinopyrum, dicoccoides, monococcum, speltoides, seigle...) pour avoir une validation supplémentaire des introgressions connues identifiées par notre méthode.

Références bibliographiques

askomics/askoR. [accessed 2024 Feb 27]. https://github.com/askomics/askoR

Balfourier F, Bouchet S, Robert S, De Oliveira R, Rimbert H, Kitt J, Choulet F, Paux E. 2019. Worldwide phylogeography and history of wheat genetic diversity. Sci Adv. 5(5):eaav0536. https://doi.org/10.1126/sciadv.aav0536

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652. https://doi.org/10.1038/nbt.1883

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 14(4):417–419. https://doi.org/10.1038/nmeth.4197

Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillet C, Paux E. 2015. Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. Genome Biol. 16(1):29. https://doi.org/10.1186/s13059-015-0601-9

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, *et al.* 2021. Optical maps refine the bread wheat Triticum aestivum cv. Chinese Spring genome assembly. Plant J Cell Mol Biol. 107(1):303–314. https://doi.org/10.1111/tpj.15289

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) .

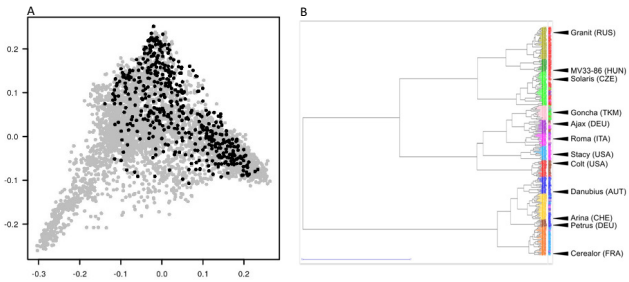


__Figure 1 - (A) PcOA de 4500 lignées du CRB, représentatives de la diversité mondiale.__ Les points noirs correspondent aux 450 lignées du projet, adpatées à une évaluation en France; __(B) Dendrogramme de Ward__ des 450 lignées. D'après Balfourier et al. (Science Adv 2019)

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.

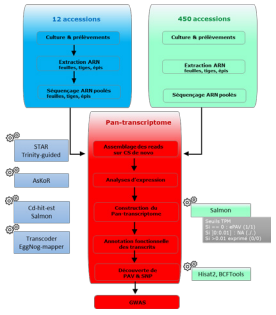


__Figure 2:__ Pipeline d'analyses

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

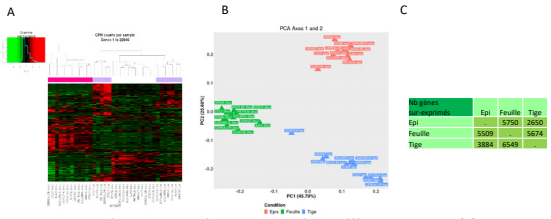


__Figure 3 – Niveaux d'expression des gènes dans 12 lignées de référence et 3 tissus ; (A) Dendogramme ; (B) ACP; (C) Nombre de gènes sur-exprimés dans les compartiments situés en ligne__
exemple: 5750 gènes sont sur-exprimés dans l'épi par rapport à la feuille

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).

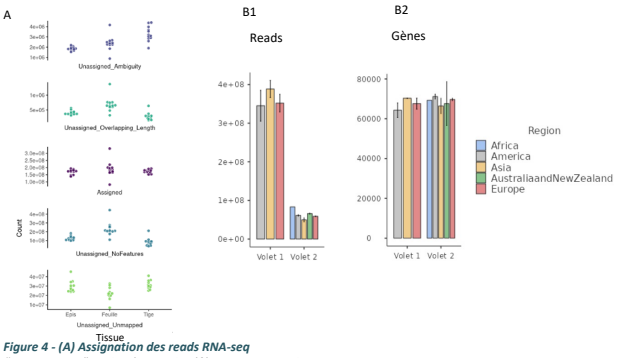


__Figure 4 - (A) Assignation des reads RNA-seq__ "Featurecount" sur le génome de référence IWGSC - Chinese spring.
Unassigned_Ambiguity: les 2 paires de reads ne mappent pas sur le même chromosome; Unassigned_Overlapping_Length: les deux paires de reads se chevauchent (minimum de 100 nucléotides de séparation 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.

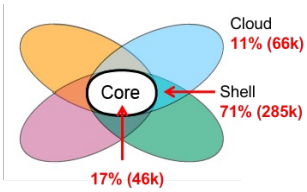


__Figure 5 – Distribution du pan-transcriptome__

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.

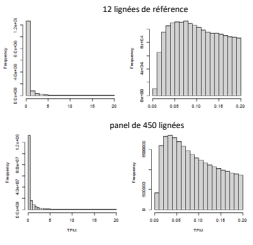


__Figure 6 – Distribution des TPM__

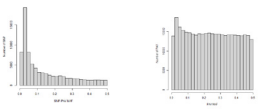


__Figure 7 – Distribution des MAF__

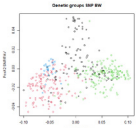


__Figure 8 – PCoA construite avec les SNP issus de ePAV__
Les couleurs correspondent aux groupes génétiques BW (Balfourier et al, 2019)

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...).

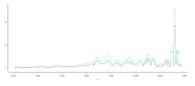
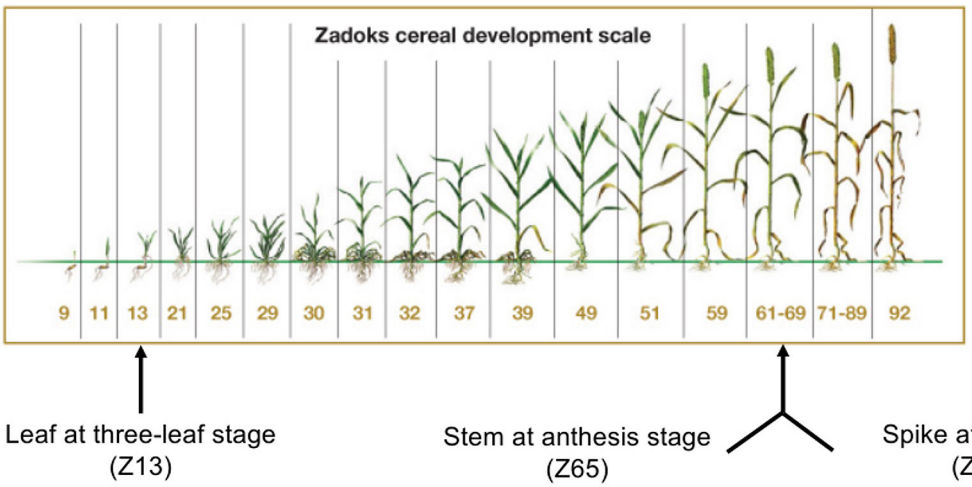


__Figure 9 - Profil d'expression des gènes du chromosome 1B chez Solaris__
Les deux réplicats de RNA-seq sur feuille sont représentés, avec une moyenne en fenêtre glissante

Annexe 1 - Développement des céréales : échelle de Zadoks



Les ARNs totaux pour les 12 lignées ont été extraits aux 3 stades de développement décrits ci-dessus pour les feuilles, tige et épis: feuille au stade 3-feuilles (Z13), tige et épis lors de l'anthèse (Z65).

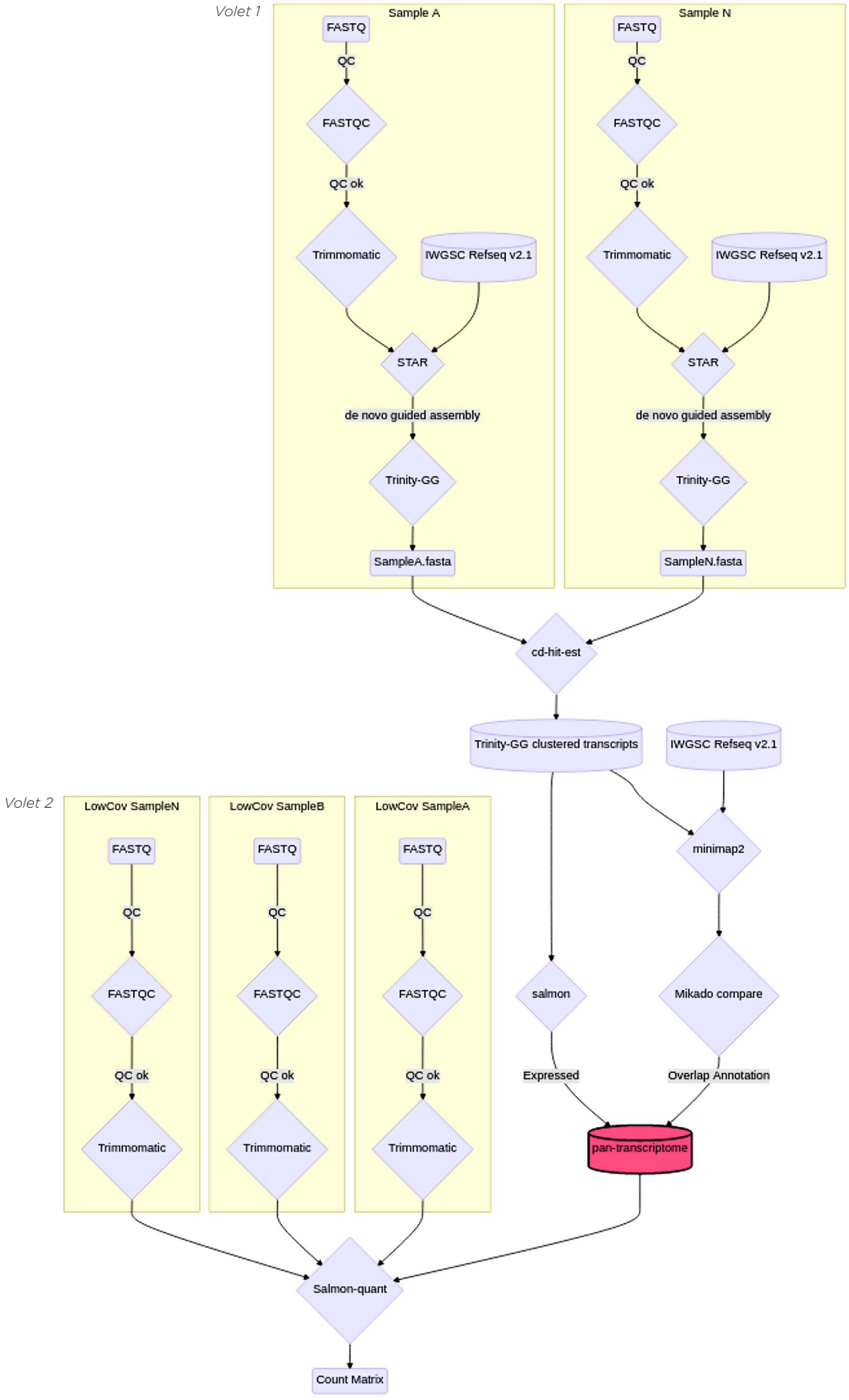
Annexe 2 - Statistiques générales sur les volets 1 et 2

Statistiques descriptives

	Volet	Region	N	Moyenne	Médiane	Ecart-type	Minimum	Maximum
reads	Volet 1	Africa	0	NaN	NaN	NaN	NaN	NaN
		America	6	345196942.33	321238205.00	97584054.12	244349741	491116797
		Asia	3	388500046.33	406920223	38361222.21	344402909	414177007
		AustraliaandNewZealand	0	NaN	NaN	NaN	NaN	NaN
		Europe	27	351918479.00	324778734	118367698.88	196754204	818399260
	Volet 2	Africa	1	83282343.00	83282343	NaN	83282343	83282343
		America	85	61225169.64	59610767	18136537.05	16578660	105053688
		Asia	21	49742318.48	53149172	24188768.87	54867	82171919
		AustraliaandNewZealand	2	65936004.00	65936004.00	1994332.45	64525798	67346210
		Europe	334	58942530.89	57438753.00	18612428.72	8404892	209096620
expressed	Volet 1	Africa	0	NaN	NaN	NaN	NaN	NaN
		America	6	64302.50	64302.50	8953.62	56129	72476
		Asia	3	70311.00	70311	0.00	70311	70311
		AustraliaandNewZealand	0	NaN	NaN	NaN	NaN	NaN
		Europe	27	67588.56	70560	14280.51	36068	85381
	Volet 2	Africa	1	69255.00	69255	NaN	69255	69255
		America	85	71162.11	72531	11396.79	39575	97949
		Asia	21	66432.05	68817	17695.02	10101	93357
		AustraliaandNewZealand	2	67604.00	67604.00	15619.99	56559	78649
		Europe	334	69673.70	69631.00	11335.62	36068	98617
ePAVs	Volet 1	Africa	0	NaN	NaN	NaN	NaN	NaN
		America	0	NaN	NaN	NaN	NaN	NaN
		Asia	0	NaN	NaN	NaN	NaN	NaN
		AustraliaandNewZealand	0	NaN	NaN	NaN	NaN	NaN
		Europe	0	NaN	NaN	NaN	NaN	NaN
	Volet 2	Africa	1	175947.00	175947	NaN	175947	175947
		America	85	195055.66	194866	15449.73	162674	244703
		Asia	21	215117.14	202266	46767.86	174775	390187
		AustraliaandNewZealand	2	200009.00	200009.00	19820.20	185994	214024
		Europe	334	197342.51	194887.00	15390.23	158230	277157

Statistiques descriptives sur le séquençage des deux volets (reads), sur le nombre de gènes exprimés (expressed) et le nombre de ePAVs.

Annexe 3 - Processus bioinformatique de traitement des données RNA-seq des volets 1 et 2



Annexe 4 - Assemblage guidé Trinity-STAR

sample	tissue	raw assembly	total bp assembled
AJAX	3feui	341 287	382 Mb
AJAX	epis	329 885	354 Mb
AJAX	tige	288 368	310 Mb
ARINA	3feui	205 692	203 Mb
ARINA	epis	362 054	373 Mb
ARINA	tige	373 537	438 Mb
CEREALOR	3feui	209 924	209 Mb
CEREALOR	epis	350 777	381 Mb
CEREALOR	tige	330 445	380 Mb
COLT	3feui	275 229	264 Mb
COLT	epis	312 064	327 Mb
COLT	tige	322 559	373 Mb
DANUBIUS	3feui	343 339	405 Mb
DANUBIUS	epis	361 218	385 Mb
DANUBIUS	tige	349 607	398 Mb
GONCHA	3feui	266 205	287 Mb
GONCHA	epis	335 651	364 Mb
GONCHA	tige	250 922	260 Mb
GRANIT	3feui	327 367	390 Mb
GRANIT	epis	358 813	394 Mb
GRANIT	tige	295 902	315 Mb
MV3386	3feui	279 661	302 Mb
MV3386	epis	373 449	417 Mb
MV3386	tige	312 978	341 Mb
PETRUS	3feui	367 998	446 Mb
PETRUS	epis	403 147	448 Mb
PETRUS	tige	331 887	366 Mb
ROMA	3feui	217 222	214 Mb
ROMA	epis	422 680	494 Mb
ROMA	tige	288 848	310 Mb
SOLARIS	3feui	344 702	407 Mb
SOLARIS	epis	348 433	375 Mb
SOLARIS	tige	365 944	427 Mb
STACY	3feui	350 399	424 Mb
STACY	epis	383 251	438 Mb
STACY	tige	307 192	349 Mb