Analyse de données transcriptomiques d'espèces sauvages apparentées au blé

Florent MARCHAL

INRAE - UMR AGAP - Ge2POP

26 août 2024

Encadrement pédagogique
Anthony BOUREUX
.
.

Encadrement scientifique
Concetta BURGARELLA,
Nathalie CHANTRET,
Vincent RANWEZ

Problématique

Objectif

Évaluer la qualité des données de séquençage issues d'une étude antérieure ([2]) pour déterminer si celles-ci peuvent être utilisées pour rechercher des traces de sélection.

Plan

Modèles biologiques

Espèces utilisées

- 13 espèces sauvages apparentés au blé
- Famille des *Poaceae* (Graminées)
- Phylogénie connue ([2])

Modèles biologiques

Espèces utilisées

- 13 espèces sauvages apparentés au blé
- Famille des *Poaceae* (Graminées)
- Phylogénie connue ([2])

Triticum urartu

- Mode de reproduction fortement autogame
- Génome diploide de $4,8 Gpb$ ($35,5$ fois plus grand qu'*Arabidopsis thaliana* ($0,135 Gpb$))
- Partie « A » du génome du blé tendre (Froment)
- Génome de référence disponible

Données

Données brutes

- Fichier FASTQ
- 44 fichier de 24 000 000 reads

Données

Données brutes

- Fichier FASTQ
- 44 fichier de 24 000 000 reads

Referenes

- Genone de référence
- Transcriptome de référence
- Ancien transcriptome de référence (ex-nihilo)

Données

Données brutes

- Fichier FASTQ
- 44 fichier de 24 000 000 reads

Referénes

- Genone de référence
- Transcriptome de référence
- Ancien transcriptome de référence (ex-nihilo)

Données raffinés

- Fichier BAMS
- Tableaux fait avec "dNdSpiNpiS" [3]

Synonyme et non synonyme

Espèce	Individu	Séquence					
espèce_1	individu 1	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TAT Tyr
espèce_1	individu 2	ATG Met	CGT Arg	TGC Cys	CGC Arg	TGT Cys	TTT Phe
espèce_1	individu 3	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TTT Phe

Table 1 – Exemple de séquences

Synonyme et non synonyme

Espèce	Individu	Séquence					
espèce_1	individu 1	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TAT Tyr
espèce_1	individu 2	ATG Met	CGT Arg	TGC Cys	CGC Arg	TGT Cys	TTT Phe
espèce_1	individu 3	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TTT Phe

Table 1 – Exemple de séquences

- **Sites synonyme**
- **Sites non synonyme**

Synonyme et non synonyme

Espèce	Individu	Séquence					
espèce_1	individu 1	ATG Met	CGT Arg	TGC Cys	CGA **Arg**	TGT Cys	TAT Tyr
espèce_1	individu 2	ATG Met	CGT Arg	TGC Cys	CGC **Arg**	TGT Cys	TTT Phe
espèce_1	individu 3	ATG Met	CGT Arg	TGC Cys	CGA **Arg**	TGT Cys	TTT Phe

Table 2 – Exemple de séquences avec un site synonyme.

- **Sites synonyme** codons codants pour un même acide aminé
- **Sites non synonyme**

Synonyme et non synonyme

Espèce	Individu	Séquence					
espèce_1	individu 1	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TAT **Tyr**
espèce_1	individu 2	ATG Met	CGT Arg	TGC Cys	CGC Arg	TGT Cys	TTT **Phe**
espèce_1	individu 3	ATG Met	CGT Arg	TGC Cys	CGA Arg	TGT Cys	TTT **Phe**

Table 3 – Exemple de séquences avec un site non synonyme.

- **Sites synonyme** codons codants pour un même acide aminé
- **Sites non synonyme** codons ne codants pas pour un même acide aminé

Polymorphisme et substitutions

Les sites synonymes peuvent s'étudiés :

au sein d'une même population : on parle de polymorphisme

au sein d'un groupe de population : on parle de substitutions

Espèce	Individu	Séquence					
espèce_1	individu 1	ATG Met	CGT **Arg**	TGC Cys	CGA Arg	TGT **Cys**	TAT Tyr
espèce_1	individu 2	ATG Met	CGT **Arg**	TGC Cys	CGC Arg	TGT **Cys**	TTT Phe
espèce_1	individu 3	ATG Met	CGT **Arg**	TGC Cys	CGA Arg	TGT **Cys**	TTT Phe
espèce_2	individu 1	ATG Met	CGA **Arg**	TGC Cys	CGA Arg	CGT **Arg**	TTT Phe
espèce_2	individu 2	ATG Met	CGA **Arg**	TGC Cys	CGA Arg	CGT **Arg**	TTT Phe
espèce_2	individu 3	ATG Met	CGA **Arg**	TGC Cys	CGA Arg	CGT **Arg**	TTT Phe

Table 4 – Exemple de substitution synonyme et non synonymes. Les substitutions synonymes sont en rouge Substitutions non synonymes sont en orange

Sélections

Absence de sélection

- Les sites évoluent à la même vitesse.

Sélections

Absence de sélection

- Les sites évoluent à la même vitesse.

Sélection purificatrice

- S'oppose aux mutations non synonymes
- Conserve les séquences

Sélections

Absence de sélection

- Les sites évoluent à la même vitesse.

Sélection purificatrice

- S'oppose aux mutations non synonymes
- Conserve les séquences

Sélection positive

- Favorise la fixation de mutations non synonymes

Indicateurs

	Site Fixés	Sites polymorphique
Non synonyme	D_n	P_n
Synonyme	D_s	P_s

Table 5 – Indicateurs utilisés pour la recherche de traces de sélections

Utilisation

- $\frac{P_n}{P_s}$ Étude du polymorphisme
- $\frac{D_n}{D_s}$ Étude des substitutions
  - $\frac{D_n}{D_s} > 1$ Conservation de la substitution
  - $\frac{D_n}{D_s} < 1$ Élimination de la substitution

Conclusion

Besoin de sites variables

→ Suffisamment de sites

→ Suffisamment grand nombre de séquence

Création d'un outil

Données de initiales

- Nombre de SNP par contig

Definitions

- SNP : *Single nucléotidique polymorphisme*
- contig : Ici, "contig" est synonyme de "gène".

Création d'un outil

Données de initiales

- Nombre de SNP par contig

Definitions

- SNP : *Single nucléotidique polymorphisme*
- contig : Ici, "contig" est synonyme de "gène".

Objectif

- Visualiser la distribution du nombre de SNP par contig

Création d'un outil

Fonctionnement

- Chargement des données
- Création d'une matrice
- Génération de figures (MatPlotLib [4])

Reproductibilité / Tracabilité

- Génération d'un fichier README a chaque éxecution
- Disponible sur github : [5]

resultats

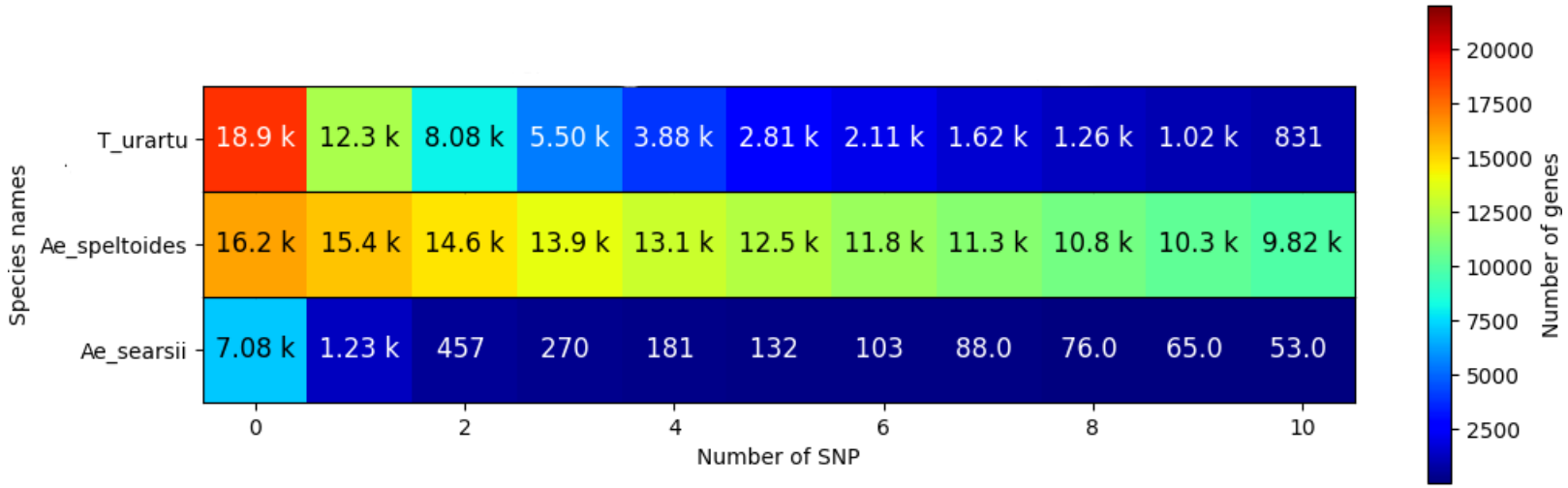

Figure 1 – Nombre de SNP par contig

Résultats

- Besoin d'au moins 5 SNP par contig sur 70% des contigs
- Seul 1 des 13 éspéces atteigne ce seuil
- Grande variabilité dans le npmbre de SNP (*Aegilops searsii*)
- *Aegilops speltoides* candidat pour une étude préparatoire

Explication des résultats précédents

- Trop peux de reads ont mappés
- Le transcriptome été référence de l'équipe est potentielment incomplet

Nouveaux mappings

- Sur le génome de référence
- Sur le transcriptome de référence
- Sur l'ancien transcriptome de référence

Attendus

$Gnome \geq transcriptome > Ancien_t ranscriptome$

Outils

GeCKO [1]

- Analyses de données NGS
- « user-friendly »

Mapper

- Transcriptomes : BWA-MEM
- Génome : Minimap2

Outils

GeCKO [1]

- Analyses de données NGS
- « user-friendly »

Mapper

- Transcriptomes : BWA-MEM
- Génome : Minimap2

Outils

GeCKO [1]

- Analyses de données NGS
- « user-friendly »

Mapper

- Transcriptomes : BWA-MEM
- Génome : Minimap2 (Non-fonctionnel)
- Génome : STAR (Arrivé trop tard)

Résultats

Analyse

- Nombre de reads par contigs
- Nombre de contigs ayants reçu des reads
- Qualité du mapping

Résultats

Analyse

- Nombre de reads par contigs
- Nombre de contigs ayants reçu des reads
- Qualité du mapping

→ **Le mapping sur l'ancien transcriptome est meilleur.**

Conclusion et Perspéctives

Conclusion

- Le jeu de donné risque de ne pas convenir.

Conclusion et Perspéctives

Conclusion

- Le jeu de donné risque de ne pas convenir.

Perspéctives

- Analyses du mapping sur le génome de référence
- La quantification des SNP n'a eu lieu que sur les "anciens BAM"

Points clef

Merci pour votre attention

Points clefs de la présentation

- Dévelopement d'un outil
- Quantification des SNP
- Remapping des BAMS
- Analyse des nouveaux BAM

Phylogénie

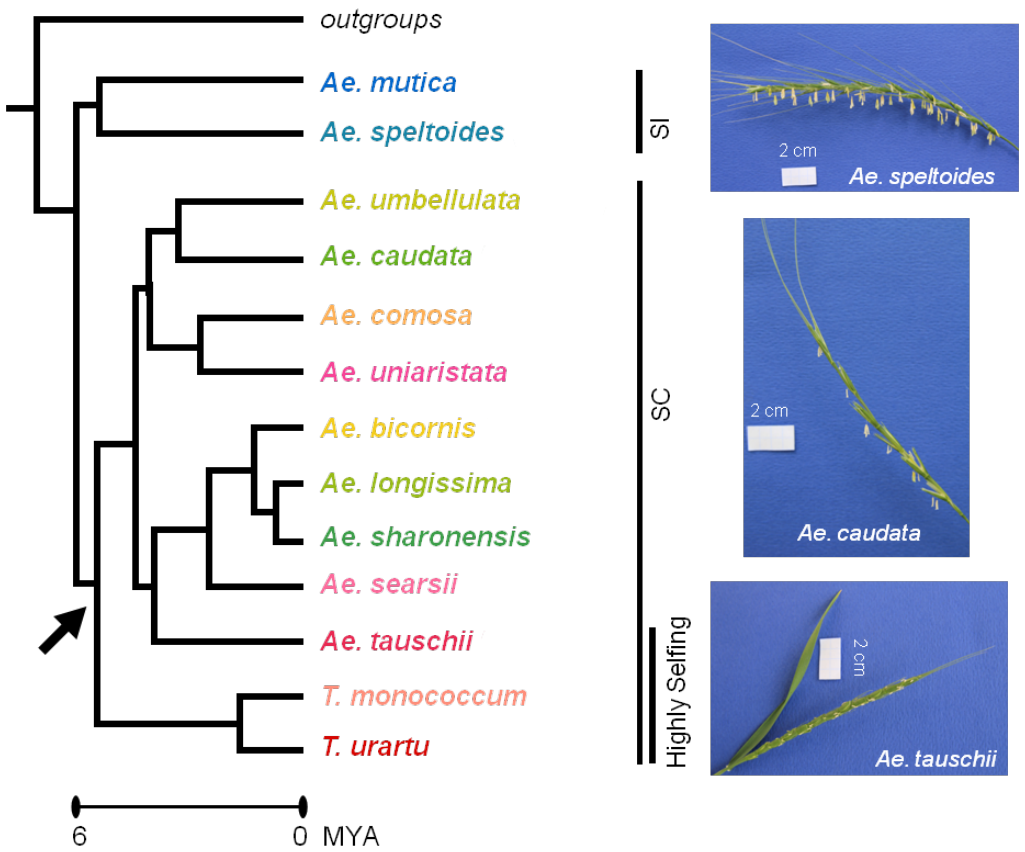

Figure 2 – Relation phylogénétique entre les 13 espèces diploïdes du genre *Aegilops* / *Triticum*. Les couleurs représentent un gradient d'auto-fécondation. Les espèces heterogame (*SI*) strictes sont bleues, les espèces avec un mode de reproduction mixte (*SC*) sont en vert / jaune et les espèces autogame (*Highly Selfing*) sont en rouge. Cette figure est issue de [2] et sa légende a été adaptée et traduite par l'auteur de ce rapport.

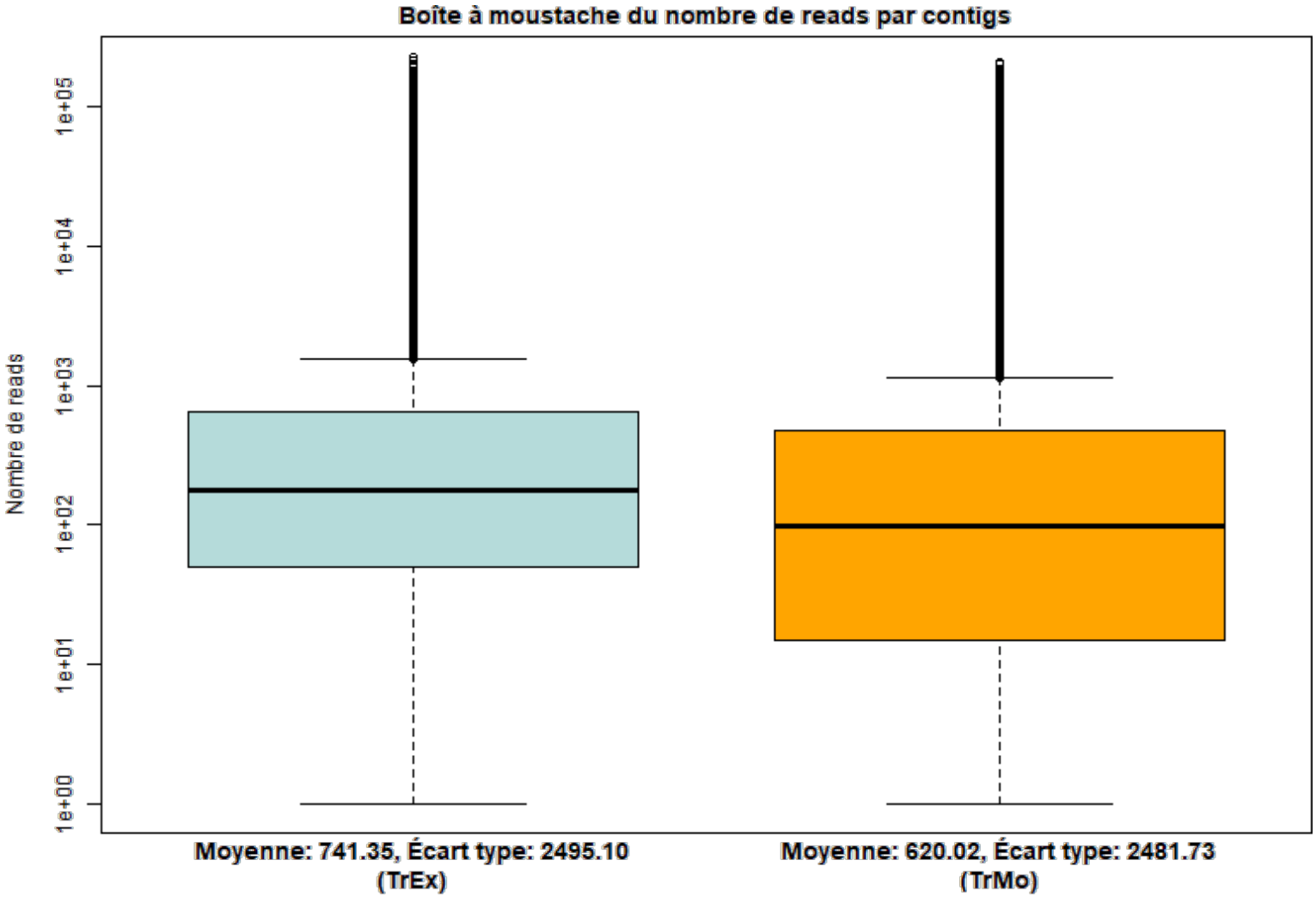

Figure 3 – Boite à moustache du nombre de reads par contigs

Références I

[1] Morgane Ardisson et al. « GeCKO : user-friendly workflows for genotyping complex genomes using target enrichment capture. A use case on the large tetraploid durum wheat genome. ». In : (mars 2024). doi : 10.21203/rs.3.rs-4123643/v1.

[2] Concetta Burgarella et al. « Mating systems and recombination landscape strongly shape genetic diversity and selection in wheat relatives ». In : *Evolution Letters* (août 2024), qrae039. issn : 2056-3744. doi : 10.1093/evlett/qrae039. url : https://doi.org/10.1093/evlett/qrae039 (visité le 17/08/2024).

Références II

[3] dNdSpiNpiS. *PopPhyl*. url : https://kimura.univ-montp2.fr/PopPhyl/index.php?section=tools (visité le 19/08/2024).

[4] John D. Hunter. « Matplotlib : A 2D Graphics Environment ». In : *Computing in Science & Engineering* 9.3 (mai 2007). Conference Name : Computing in Science & Engineering, p. 90-95. issn : 1558-366X. doi : 10.1109/MCSE.2007.55. url : https://ieeexplore.ieee.org/document/4160265 (visité le 19/08/2024).

Références III

[5] Florent Marchal. *F-Marchal/M1BioinfoInternship2024-INRAE_AGAP_GE2POP*. original-date : 2024-07-26T07 :35 :22Z. Juill. 2024. url : https://github.com/F-Marchal/M1BioinfoInternship2024-INRAE_AGAP_GE2POP (visité le 02/08/2024).