Analyse de données transcriptomiques d'espèces sauvages apparentées au blé

Florent MARCHAL

INRAE - UMR AGAP - Ge2POP

27 août 2024

Encadrement pédagogique
Anthony BOUREUX
.
.

Encadrement scientifique
Concetta BURGARELLA,
Nathalie CHANTRET,
Vincent RANWEZ

Problématique

Objectif

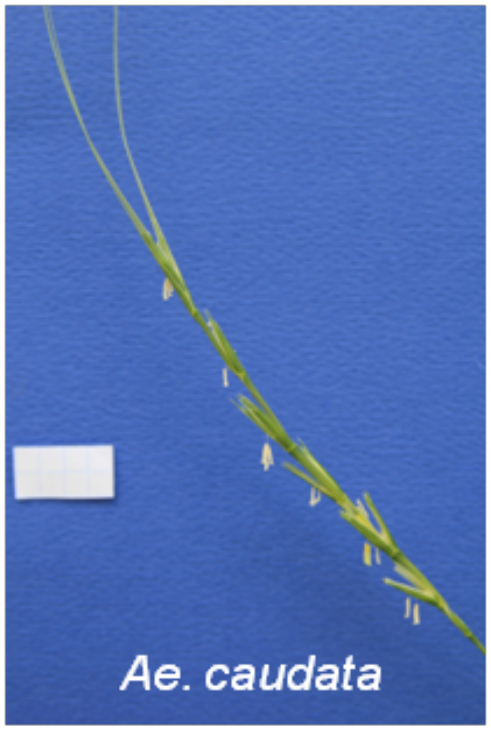
Évaluer la qualité des données transcriptomiques issues d'une étude antérieure ([Bur+24]) pour déterminer si celles-ci peuvent être utilisées pour rechercher des traces de sélection.

Plan

Modèles biologiques

Espèces utilisées

- 13 espèces sauvages apparentées au blé
- Famille des *Poaceae* (Graminées)



Modèles biologiques

Espèces utilisées

- 13 espèces sauvages apparentées au blé
- Famille des *Poaceae* (Graminées)

Triticum urartu

- Génome diploide de $4,8 Gpb$ ($35,5$ fois plus grand qu'*Arabidopsis thaliana*)
- Génome de référence disponible

SNP et Contigs

- **SNP** *Single nucléotidique polymorphism*
- **contig**
- **Sites synonymes**
- **Sites non synonymes**

SNP et Contigs

- **SNP** *Single nucléotidique polymorphisme*
- **contig** Ici, "contig" est synonyme de "gène"
- **Sites synonymes**
- **Sites non synonymes**

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 synonymes**
- **Sites non synonymes**

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

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

Polymorphisme et substitutions

Les sites peuvent s'étudier :

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 synonyme. Les substitutions synonymes sont en rouge. Les substitutions non synonymes sont en orange

Sélections

L'absence de sélection

- Les sites synonymes et non synonymes se fixent à la même vitesse

Sélections

L'absence de sélection

- Les sites synonymes et non synonymes se fixent à la même vitesse

La sélection purificatrice

- S'oppose à la fixation des sites non synonymes

Sélections

L'absence de sélection

- Les sites synonymes et non synonymes se fixent à la même vitesse

La sélection purificatrice

- S'oppose à la fixation des sites non synonymes

La sélection positive

- Favorise la fixation de sites synonymes

Sélections

L'absence de sélection

- Les sites synonymes et non synonymes se fixent à la même vitesse

La sélection purificatrice

- S'oppose à la fixation des sites non synonymes

La sélection positive

- Favorise la fixation de sites synonymes

→ Création d'un déséquilibre

Indicateurs

	Sites polymorphiques	Site Fixés
Non synonyme	P_n	D_n
Synonyme	P_s	D_s

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

Utilisation

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

Conclusion

Besoin de sites variables

→ Grand nombre de sites

→ Grand nombre de contigs

Création d'un outil

Données initiales

- Des tableaux contenant le nombre de SNP par contig

→ Tableaux générés avec "dNdSpiNpiS" ([dNd])

→ Utilisation du mapping utilisant le transcriptome de référence de l'équipe

Création d'un outil

Données initiales

- Des tableaux contenant le nombre de SNP par contig

→ Tableaux générés avec "dNdSpiNpiS" ([dNd])

→ Utilisation du mapping utilisant le transcriptome de référence de l'équipe

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 [Hun07])

Reproductibilité / Traçabilité

- Génération d'un fichier README à chaque exécution
- Disponible sur GitHub : [Mar24]

Besoins

- Au moins 5 SNP par contig
- Sur au moins 70% des contigs

Résultats

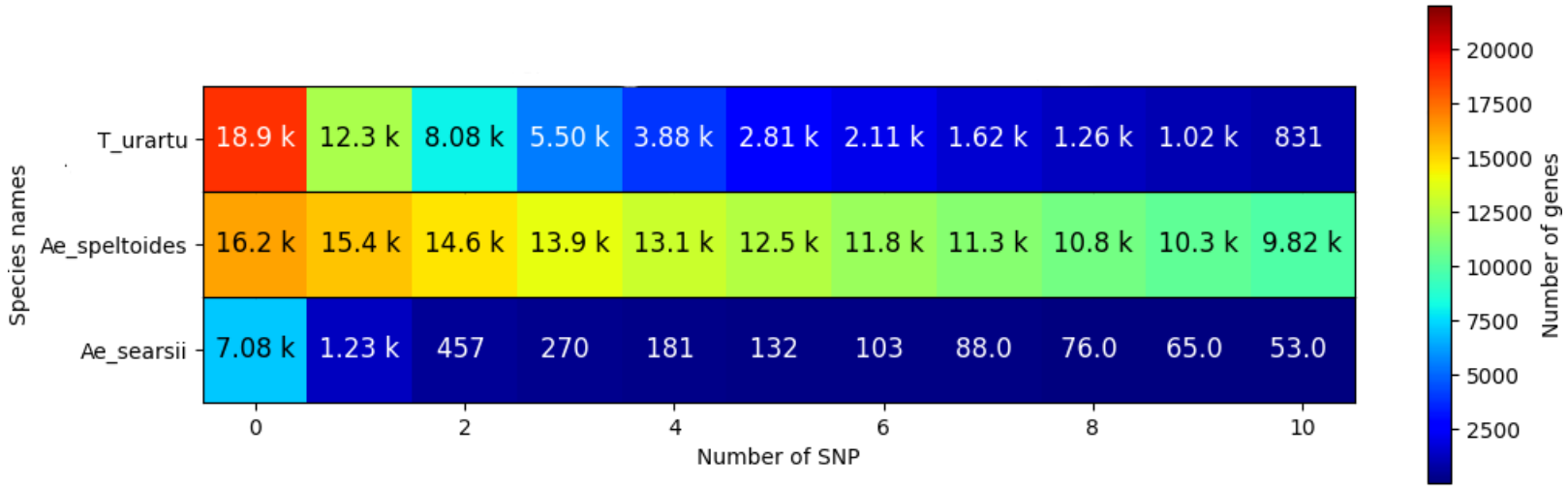


Figure 1 – Nombre de SNP par contig

Résultats

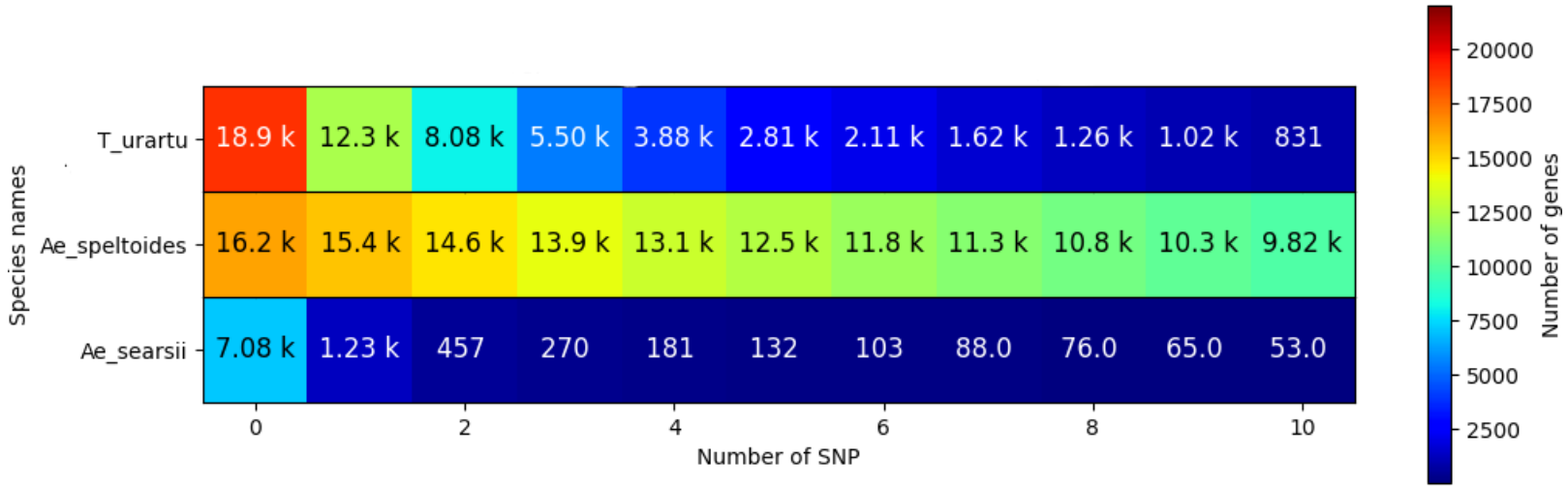


Figure 1 – Nombre de SNP par contig

Résultats

- Seule 1 des 13 espèces atteint le seuil (*Aegilops speltoides*)
- Grande variabilité dans le nombre de SNP (*Aegilops searsii*)
- *Aegilops speltoides* candidat pour une étude préparatoire

Potentielle explication des résultats précédents

- Trop peu de reads ont mappés
- Le transcriptome référence provenant de l'équipe est potentiellement incomplet

Nouveaux mappings

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

Attendus

Génome $\geq$ Transcriptome $>$ Ancien transcriptome

Outils

GeCKO [Ard+24]

- Analyses de données NGS
- « user-friendly »

Mappers

- Transcriptomes : BWA-MEM
- Génome : Minimap2

Outils

GeCKO [Ard+24]

- Analyses de données NGS
- « user-friendly »

Mappers

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

Analyses

Données brutes

- Fichiers FASTQ
- 44 fichiers de 24 000 000 reads

→ Utilisation d'un cluster de calcul

Analyses

- Nombre de reads par contig
- Nombre de contigs ayant reçu des reads
- Qualité du mapping

Analyses

Données brutes

- Fichiers FASTQ
- 44 fichiers de 24 000 000 reads

→ Utilisation d'un cluster de calcul

Analyses

- Nombre de reads par contig
- Nombre de contigs ayant reçu des reads
- Qualité du mapping

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

Conclusion et Perspectives

Conclusion

- Le jeu de données risque de ne pas convenir

Conclusion et Perspectives

Conclusion

- Le jeu de données risque de ne pas convenir

Perspectives

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

Phylogénie

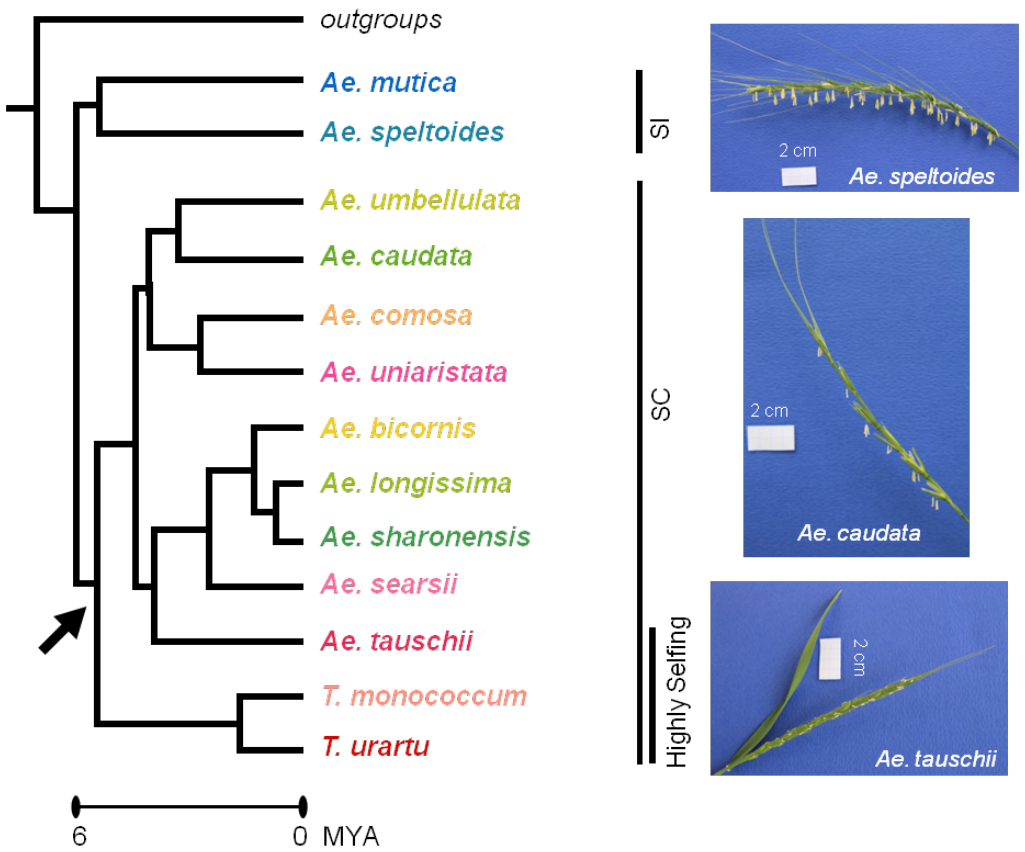
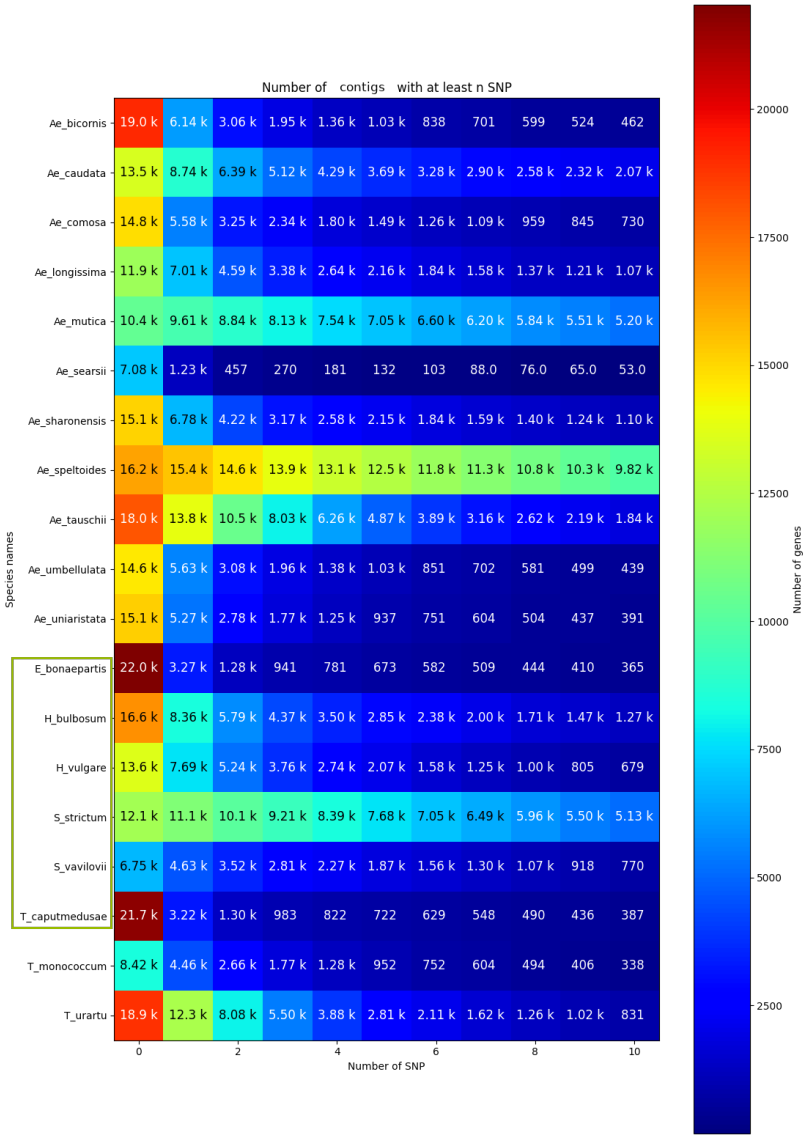
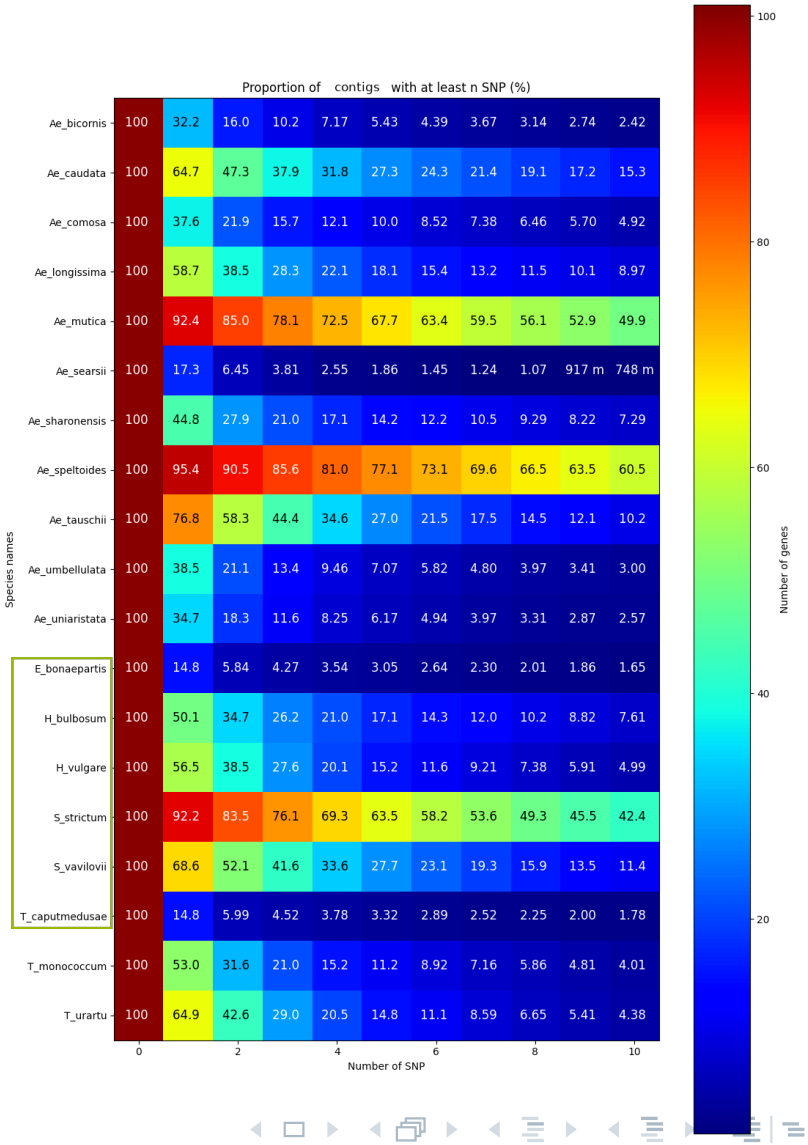


Figure 2 – Relation phylogénétique entre les 13 espèces diploïdes du genre *Aegilops* / *Triticum*. Les couleurs représentent un gradient d'auto-fécondation. Les espèces heterogame (*SI*) strictes sont bleues, les espèces avec un mode de reproduction mixte (*SC*) sont en vert / jaune et les espèces autogame (*Highly Selfing*) sont en rouge. Cette figure est issue de [Bur+24] et sa légende a été adaptée et traduite par l'auteur de ce rapport.

Heatmaps complètes (Analyse des SNP)





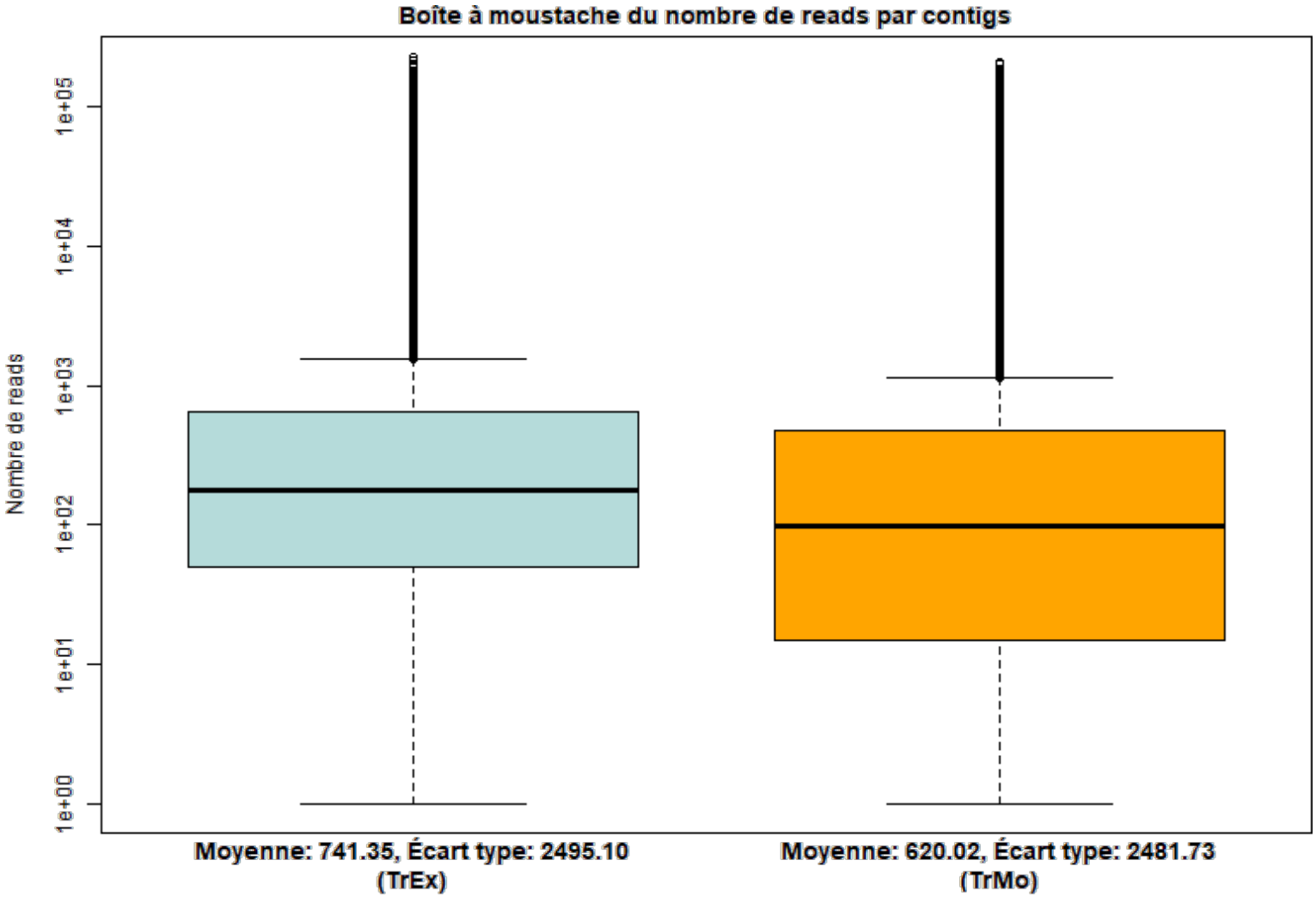


Figure 4 – Boîte à moustaches du nombre de reads par contig

Références I

[Ard+24] 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.

[Bur+24] 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

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

[Hun07] 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

[Mar24] 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).