**Intitulé du stage : Méthodes statistiques pour l'optimisation de la prédiction génomique par l'intégration d'informations fonctionnelles**

**Laboratoire d'accueil** : UMR AGAP Institut, Avenue Agropolis - 34398 Montpellier Cedex 5

**Encadrants** : David Pot (CIRAD, HDR, Généticien en appui aux programmes de sélection), et Vincent Garin (CIRAD, Biostatisticien)

**Contexte et problématique de l'étude**

La sélection de variétés de plantes adaptées aux évolutions du climat et des systèmes agricoles futurs est un des enjeux majeurs de la recherche agronomique actuelle. La prédiction de caractères complexes dans des organismes biologiques comme les animaux ou les plantes à l'aide d'information génétique, connue sous le nom de prédiction génomique, est une application importante des statistiques pour améliorer la sélection [1,2]. Fortement ancrée dans la théorie de modèles mixtes, la prédiction génomique s'est enrichie d'apport des statistiques Bayésiennes ainsi que plus récemment des algorithmes issus de l'apprentissage machine.
La compréhension de la construction et de la variabilité des phénotypes cibles de la sélection implique deux grands types d'équipes de recherche :

- des équipes de généticiens quantitatifs et de biostatisticiens tentant d'expliquer la variation génétique des caractères par des modèles statistiques ancrés sur la variabilité nucléotidique [1,2]
- des équipes de biologistes et physiologistes moléculaires visant quant à elles l'identification des mécanismes physiologiques et des gènes et réseaux moléculaires sous-jacents [3,4] permettant d'expliquer la construction des phénotypes finaux et leurs expressions.

Si la prédiction génomique commence à prendre en compte les informations issues des études d'identification de gènes majeurs [2,5], elle ne tire pas encore complètement parti de la compréhension fine des mécanismes moléculaires impliqués dans la construction des phénotypes. La meilleure intégration des informations génétiques et/ou biologiques dans les modèles de prédiction génomique est une stratégie prometteuse pour continuer à renforcer cette approche ayant déjà fait ces preuves [6–17]. Par exemple, des approches mobilisant des informations d'annotations fonctionnelles ont été mises en œuvre soit basées sur les Go-Terms [18] soit basées sur des propriétés positionnelles ou évolutives [19–21]. De manière plus générale, la transposition de solutions issues de la génétique animale ou humaine peut aussi être une approche intéressante dans ce domaine très dynamique à l'interface de nombreuses disciplines. Basées sur l'utilisation de larges jeux de données couvrant différents niveaux d'information biologique (variabilité nucléotidique, transcriptomique, protéomique, métabolomique), la prédiction génomique et l'intégration d'information fonctionnelle peuvent aussi nécessiter la mobilisation de compétences informatiques pour optimiser l'utilisation de ces données.

L'objectif de ce projet de Master 2 est d'explorer les bénéfices liés à l'ajout d'information fonctionnelle dans les approches de prédictions génomiques au sein d'un large dispositif de croisement multi-parental (population BCNAM comprenant plus de 3900 familles) qui a été évalué au sein de plusieurs environnements. Ces travaux s'inscrivent dans le cadre du projet ANR SorDrought « Caractérisation de nouveaux traits physiologiques pour aider l'amélioration de la tolérance au stress hydrique post-floral chez le sorgho » qui a pour objectif de développer de nouvelles méthodologies d'appui à la sélection dans un partenariat impliquant des entités de recherche publiques (INRAE, IRD, CERAAS) et privées (LIDEA Seeds et RAGT2N).

**Description des travaux à réaliser et objectifs du stage**

Les travaux à mener dans le cadre du stage s'ancreront sur un large dispositif de croisement multiparental qui a été développé en partenariat entre le CIRAD, l'Institut d'Economie Rural (IER) du Mali et l'Institut de recherche international sur les espèces de zones semi-arides tropicales (ICRISAT). Des analyses de détection de QTL prenant en compte des covariables environnementales ont déjà été effectuées [22] et l'objectif sera de développer des modèles de prédiction génomique mobilisant les informations fonctionnelles incluant les QTL déjà détectés, les positions des polymorphismes par rapport aux gènes, les annotations des gènes contenant les polymorphismes… (classes fonctionnelles, signature de sélection, appartenance à des modules d'expression) pour évaluer les bénéfices de ces approches. L'effet de la structure des populations de calibration et de validation seront aussi explorées. Pour ce faire des modèles statistiques déjà développés seront testés [18–20] et des approches alternatives pourront être mises en œuvre.
En fonction des avancées obtenues lors du stage et des questionnements qui seront soulevés des jeux de données complémentaires (populations de GWAS à large base génétique exposés à des traitements hydriques contrastants) pourront aussi être mobilisées pour tester des questions de transferts de calibration et de structure des populations d'entrainement et de validation.

**Profil recherché**
Formation en statistiques, mathématiques appliquées, bio-informatique, ou autre domaine avec un fort accent sur les méthodes quantitatives, en particulier les modèles mixtes et/ou les approches bayésiennes
Fort intérêt pour le test et le développement de méthodologies statistiques
Intérêt pour les applications dans le domaine de l'amélioration des plantes avec un fort attrait pour les biostatistiques et la bio-informatique
Utilisation d'un langage d'analyse et de programmation (R, Python, C++)
Attrait pour le travail en équipe incluant notamment des généticiens, des sélectionneurs et des statisticiens
Le désir de travailler dans un contexte international lie au développement de l'agriculture est un plus.

**Perspectives de suite**
Une bourse de thèse ciblant les aspects de compréhension du déterminisme génétique des caractères d'adaptation au stress hydrique post floral et le développement de méthodes de prédiction des valeurs génétiques a été obtenue dans le cadre du projet ANR SorDrought. En fonction de l'intérêt de l'étudiant de master 2 pour cette thématique et son intégration au sein de l'équipe encadrante, la poursuite en thèse pourra être envisagée.

**Dates du stage :** entre Janvier et Septembre 2024 en fonction des calendriers des formations

**Indemnités de stage :** environ 600 € par mois

**Contacts:**

David Pot (HDR) : Généticien en appui aux programmes de sélection
CIRAD Agricultural Research for Development
Biological Systems Department
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants"
Genetics and Varietal Innovation Team
Building 3 - Office 129
TA A-108 / 03 - Avenue Agropolis - 34398 Montpellier Cedex 5
France
E-mail: david.pot@cirad.fr
Téléphone : +33 6 51 75 13 76

Vincent Garin : Biostatisticien en appui aux programmes de sélection
CIRAD Agricultural Research for Development
Biological Systems Department
AGAP Mixed Research Unit "Genetic Improvement and Adaptation of Mediterranean and Tropical Plants"
Genetics and Varietal Innovation Team
Building 3bis – Office 152
TA A-108 / 03 - Avenue Agropolis - 34398 Montpellier Cedex 5
France
E-mail: vincent.garin@cirad.fr

## Publications pertinentes de l'équipe (les personnes de l'équipe sont indiquées en gras) :

**Burgarella, C**., Berger, A., Glémin, S., David, J., Terrier, N., Deu, M., **Pot, D**., 2021. The Road to Sorghum Domestication: Evidence From Nucleotide Diversity and Gene Expression Patterns. Front. Plant Sci. 12, 1706. https://doi.org/10.3389/fpls.2021.666075

**Garin, V**., Choudhary, S., Murugesan, T., Kaliamoorthy, S., Diancumba, M., Hajjarpoor, A., Chellapilla, T.S., Gupta, S.K., Kholovà, J., 2023a. Characterization of the Pearl Millet Cultivation Environments in India: Status and Perspectives Enabled by Expanded Data Analytics and Digital Tools. Agronomy 13, 1607. https://doi.org/10.3390/agronomy13061607

**Garin, V**., Diallo, C., Tekete, M.L., Thera, K., Guitton, B., Dagno, K., Diallo, A.G., Kouressy, M., Leiser, W., Rattunde, F., Sissoko, I., Toure, A., Nebie, B., Samake, M., Kholova, J., **Frouin, J., Pot, D.**, **Vaksmann, M**., Weltzien, E., Teme, N., Rami, J.-F., 2023b. Characterization of adaptation mechanisms in sorghum using a multi-reference back-cross nested association mapping design and envirotyping. https://doi.org/10.1101/2023.03.11.532173

**Garin, V**., Malosetti, M., van Eeuwijk, F., 2020a. The usefulness of multi-parent multi-environment QTL analysis: an illustration in different NAM populations (preprint). Genetics. https://doi.org/10.1101/2020.02.03.931626

**Garin, V**., Wimmer, V., Borchardt, D., Malosetti, M., van Eeuwijk, F., 2020b. The influence of QTL allelic diversity on QTL detection in multi-parent populations: a simulation study in sugar beet (preprint). Genetics. https://doi.org/10.1101/2020.02.04.930677

**Garin, V**., Wimmer, V., Mezmouk, S., Malosetti, M., van Eeuwijk, F., 2017. How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. TAG Theor. Appl. Genet. Theor. Angew. Genet. 130, 1753–1764. https://doi.org/10.1007/s00122-017-2923-3

**Hennet, L.,** Berger, A., Trabanco, N., Ricciuti, E., Dufayard, J.-F., Bocs, S., Bastianelli, D., Bonnal, L., Roques, S., Rossini, L., Luquet, D., Terrier, N., **Pot, D**., 2020. Transcriptional Regulation of Sorghum Stem Composition: Key Players Identified Through Co-expression Gene Network and Comparative Genomics Analyses. Front. Plant Sci. 11. https://doi.org/10.3389/fpls.2020.00224

**Nguyen, V.H**., Morantte, R.I.Z., Lopena, V., Verdeprado, H., Murori, R., Ndayiragije, A., Katiyar, S.K., Islam, M.R., Juma, R.U., Flandez-Galvez, H., Glaszmann, J.-C., Cobb, J.N., **Bartholomé, J**., 2023. Multi-environment Genomic Selection in Rice Elite Breeding Lines. Rice 16, 7. https://doi.org/10.1186/s12284-023-00623-6

## Références Bibliographiques

1. Hickey, J. M., Chiurugwi, T., Mackay, I. & Powell, W. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* **49**, 1297–1303 (2017).
2. Xu, Y. *et al.* Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Commun.* **1**, 100005 (2020).
3. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911–920 (2010).
4. Brooks, M. D. *et al.* ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. *Plant Physiol.* **185**, 49–66 (2021).
5. Liu, X. *et al.* Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. *Front. Plant Sci.* **10**, 1129 (2019).
6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
7. Westhues, M. *et al.* Omics-based hybrid prediction in maize. *TAG Theor. Appl. Genet. Theor. Angew. Genet.* **130**, 1927–1939 (2017).
8. Westhues, M., Heuer, C., Thaller, G., Fernando, R. & Melchinger, A. E. Efficient genetic value prediction using incomplete omics data. *TAG Theor. Appl. Genet. Theor. Angew. Genet.* **132**, 1211–1222 (2019).
9. Schrag, T. A. *et al.* Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics* **208**, 1373–1385 (2018).
10. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).
11. Zhengcao, L. Integrating Omics Data into Genomic Prediction. (Georg-August-University Göttingen, 2019).
12. Hu, X., Xie, W., Wu, C. & Xu, S. A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnol. J.* **17**, 2011–2020 (2019).
13. Azodi, C. B., Pardo, J., VanBuren, R., Campos, G. de los & Shiu, S.-H. Transcriptome-Based Prediction of Complex Traits in Maize. *Plant Cell* **32**, 139–151 (2020).
14. Chateigner, A. *et al.* Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC Genomics* **21**, 1–16 (2020).
15. de las Heras-Saldana, S. *et al.* Use of gene expression and whole-genome sequence information to improve the accuracy of genomic prediction for carcass traits in Hanwoo cattle. *Genet. Sel. Evol.* **52**, 54 (2020).
16. Gage, J. L. *et al. Variation in upstream open reading frames contributes to allelic diversity in protein abundance*. 2021.05.25.445499 https://www.biorxiv.org/content/10.1101/2021.05.25.445499v2 (2021) doi:10.1101/2021.05.25.445499.

17. Giri, A., Khaipho-Burch, M., Buckler, E. S. & Ramstein, G. P. Haplotype associated RNA expression (HARE) improves prediction of complex traits in maize. *PLOS Genet.* **17**, e1009568 (2021).
18. Edwards, S. M., Sørensen, I. F., Sarup, P., Mackay, T. F. C. & Sørensen, P. Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. *Genetics* **203**, 1871–1883 (2016).
19. Mollandin, F., Gilbert, H., Croiseau, P. & Rau, A. Accounting for overlapping annotations in genomic prediction models of complex traits. *BMC Bioinformatics* **23**, 365 (2022).
20. Ramstein, G. P. *et al.* Dominance Effects and Functional Enrichments Improve Prediction of Agronomic Traits in Hybrid Maize. *Genetics* **215**, 215–230 (2020).
21. Ramstein, G. P. & Buckler, E. S. Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. *Genome Biol.* **23**, 183 (2022).
22. Garin, V. *et al.* Characterization of adaptation mechanisms in sorghum using a multi-reference back-cross nested association mapping design and envirotyping. 2023.03.11.532173 Preprint at https://doi.org/10.1101/2023.03.11.532173 (2023).