Intitulé du stage : Implémentation et comparaison de méthodes pour l'analyses de données d'essais multi-environnement et la prédiction des valeurs génétiques

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 prédiction de phénotype est un des grands défis scientifiques de la biologie moderne[1]. Depuis le début des années 2000, la prédiction génomique propose de prédire le phénotype d'individus non-observés à l'aide de leurs données génétiques et de modèles statistiques calibrés sur une population d'entraînement[2,3]. L'implémentation de cette approche a permis d'importants gains de productivité dans la sélection animale[4] et plus tard la sélection des plantes[5]. L'adaptation des modèles de prédiction génomique au contexte multi-environnementale (PGME) a ouvert la porte à une meilleure prise en comptes des interactions génotype-environnement dans le développement de variétés mieux adaptées à la diversité environnementale présente et future[6-8]. La possibilité d'extrapoler le comportement de plantes dans des situations nouvelles offerte par la PGME devient un outil précieux dans le cadre du changement climatique et de l'adaptation des systèmes agricoles à de nouvelles conditions[9].

D'un point de vue méthodologique, des développements de modèles de prédiction génomiques ont eu lieu dans plusieurs disciplines allant des modèles linéaires mixtes[10-12], au machine learning[13-15], en passant par les approches de deep learning[16,17]. Si les approches plus classiques issues des modèles mixtes conservent des propriétés intéressantes en matière d'interprétabilité et de capacités prédictives, elles requièrent souvent une plus grande puissance de calcul[18], ce qui limite leur application aux très grands jeux de données qui sont de plus en plus disponibles[19].

L'objectif principal de ce stage est d'implémenter et de comparer les grandes classes de modèle de prédiction génomique en dispositif multi-environnement. Après une phase d'initiation à des méthodes appartenant à chacune des principales catégories susmentionnées, un approfondissement d'une méthode ou d'une catégorie d'approche afin de l'optimiser et/ou proposer des améliorations pourra être mis en œuvre. Un intérêt particulier sera donné à la facilitation de l'estimation des modèles mixtes à l'aide de software ou package optimisés pour le temps de calcul (exemple : transposition de modèles estimés par ASreml dans INLA). L'intégration d'une plus grande compréhension biologique dans des modèles de machine ou de deep learning, notamment à travers des fonctions de coûts alternatives pourrait aussi être exploré.

Description des travaux à réaliser et objectifs du stage

Ce travail démarrera avec une prise en main du pipeline d'analyse développé par les équipes du CIRAD (https://gitlab.cirad.fr/agap/giv/g2amours) et des jeux de données déjà compilés prêts pour l'analyse. Après quoi, il sera attendu que le stagiaire compile 1 ou 2 jeux de données disponibles publiquement

(ex données maïs DROPS) pour étendre les possibilités de comparaison. Une intégration de ces données dans la base de données breeding management system (BMS) sera la bienvenue.

La principale activité du stage consistera à étendre une des approches abordées et/ou de proposer des schémas de comparaisons innovants, par exemple à travers des scénarios de prédiction différenciant prédiction pour de nouvelles années, de nouveaux lieux ou la combinaison des deux (lieux, année)

La synthèse de ces implémentations dans des fonctions ou librairies R/pyhton ainsi que la documentation des comparaisons de méthodes dans un rapport constitueront les deux principaux rendus attendus dans ce stage.

Profil recherché

Formation en génétique quantitative appliquée à l'amélioration des plantes, en statistiques, mathématiques appliquées, (bio-)informatique, ou autre discipline avec un fort accent sur les méthodes quantitatives

Bonnes compétences dans au moins un logiciel d'analyse de données comme R ou Python. Compétences en programmation appréciées

Expérience du travail sur un serveur de calcul ou volonté d'apprentissage

La capacité à travailler en autonomie à partir d'objectifs établis et à faire des propositions sera valorisée

Attrait pour le travail en équipe incluant notamment des statisticiens, généticiens et des sélectionneurs

Dates du stage : entre mars et septembre 2026 en fonction des calendriers des formations

Indemnités de stage : environ 600 € par mois

Contacts:

David Pot (HDR) : Généticien en appui aux programmes de sélection
Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Département BIOS - Unité mixte de recherche « amélioration génétique et l'adaptation des plantes méditerranéennes et tropicales » (AGAP)
Equipe génétique et innovation variétale (GIV)
Bâtiment 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
Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Département BIOS - Unité mixte de recherche « amélioration génétique et l'adaptation des plantes méditerranéennes et tropicales » (AGAP)
Equipe génétique et innovation variétale (GIV)
Bâtiment 3bis – bureau 159
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) :

Bienvenu, C., **Garin, V.**, Salas, N., Théra, K., Tekete, M. L., Sarathjith, M. C., Diallo, C., Berger, A., Calatayud, C., **De Bellis, F.**, Rami, J-F., **Vaksmann M.**, Segura, V., **Pot, D.**, & **De Verdal, H.** (2025). Factors influencing phenomic prediction: A case study on a large sorghum back cross nested association mapping population. *The Plant Phenome Journal*, *8*(1), e70051. https://doi.org/10.1002/ppj2.70051

De Verdal, H., Segura, V., **Pot, D.**, Salas, N., **Garin, V.**, Rakotoson, T., Raboin, L-M., **VomBrocke, K.**, Dusserre, J., **Castro Pacheco S. A.** & **Grenier, C.** (2024). Performance of phenomic selection in rice: Effects of population size and genotype-environment interactions on predictive ability. *Plos one*, *19*(12), e0309502. https://doi.org/10.1371/journal.pone.0309502

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., 2024. Characterization of adaptation mechanisms in sorghum using a multireference back-cross nested association mapping design and envirotyping. *Genetics*, *226*(4), iyae003. https://doi.org/10.1101/2023.03.11.532173

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

Références Bibliographiques

1. Washburn, J. D., Varela, J. I., Xavier, A., Chen, Q., Ertl, D., Gage, J. L., ... & de Leon, N. (2025). Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. *Genetics*, *229*(2), iyae195.
2. Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4):1819-1829
3. Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Science, 34(1):20-25.
4. Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science, 92(2):433-443
5. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González Camacho, J. M., Pérez-Elizalde, S., and Beyene, Y. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science, 22(11):961-975
6. Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., and Krishnamachari, A. (2006). Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. Crop science, 46(4):1722-1733
7. Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modelling genotype x environment interaction using pedigree and dense molecular markers. Crop Science, 52(2):707-719
8. Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., and Crossa, J. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. G3: Genes, Genomes, Genetics, 8(4):1347-1365
9. Ornella, L. A., Broccanello, C., & Balzarini, M. (2024). Plant adaptation to climate change using genomic selection and high throughput technologies. *Frontiers in Genetics*, *15*, 1471995.
10. Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W., ... & Prasanna, B. M. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3: Genes, Genomes, Genetics*, *10*(8), 2725-2739.
11. Tolhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M., & Gorjanc, G. (2022). Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, *135*(10), 3393-3415.
12. Xavier, A., Runcie, D., & Habier, D. (2025). Megavariate methods capture complex genotype-by-environment interactions. *Genetics*, *229*(4), iyae179.
13. Fernandes, I. K., Vieira, C. C., Dias, K. O., & Fernandes, S. B. (2024). Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials. *Theoretical and Applied Genetics*, *137*(8), 189
14. Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings* (Vol. 5, No. Suppl 3, p. S11). London: BioMed Central
15. Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Random forest for genomic prediction. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 633-681). Springer International Publishing
16. Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., & Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3: Genes, Genomes, Genetics*, *8*(12), 3813-3828
17. Cuevas, J., Montesinos-López, O., Juliana, P., Guzmán, C., Pérez-Rodríguez, P., González-Bucio, J., ... & Crossa, J. (2019). Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. *G3: Genes, Genomes, Genetics*, *9*(9), 2913-2924.
18. Kick, D. R., Wallace, J. G., Schnable, J. C., Kolkman, J. M., Alaca, B., Beissinger, T. M., ... & Washburn, J. D. (2023). Yield prediction through integration of genetic, environment, and management data through deep learning. *G3: Genes, Genomes, Genetics*, *13*(4), jkad006.
19. Genomes to Fields (2025). Genomes to Fields 2024 Maize Genotype by Environment Prediction Competition. CyVerse Data Commons

Unité Mixte de Recherche Amélioration Génétique des Plantes méditerranéennes et tropicales (AGAP)
Avenue Agropolis – TA A-108/03 – 34398 Montpellier cedex 5 – France.
Courriel : diragap@cirad.fr – Site internet : http/umr-agap.cirad.fr