PROPOSITION D'UN PROJET DE THÈSE
A L'ÉCOLE DOCTORALE
« Végétal, Animal, Aliment, Mer, Environnement »

INFORMATIONS GÉNÉRALES

Titre de la thèse : Prédiction Génomique pour l'Exploration de la sous-dominance dans génomes polyPLOIDes

(F) et (GB)

Acronyme : GP4PLOID

Discipline de recherche 1 : Statistique

Discipline de recherche 2 : Génomique évolutive

Trois mots-clés : Apprentissage statistique, Prédiction génomique, Evolution

(F) et (GB)

Etablissement d'inscription : Université d'Angers

Unité d'accueil : Institut de Recherches en Horticulture et Semences (IRHS), UMR 1345 UA-INRAE-IARA

Nom, prénom du directeur·rice de thèse (HDR indispensable) : LANDES CLAUDINE

Adresse courriel : claudine.landes@univ-angers.fr

Nom, prénom du co-directeur·rice (le cas échéant) (HDR indispensable) : PANLOUP FABIEN

Adresse courriel : fabien.panloup@univ-angers.fr

Nom, prénom du co-encadrant·e de thèse 1 (le cas échéant) : RABIER CHARLES-ELIE

Adresse courriel : charles-elie.rabier@univ-angers.fr

Contact(s) (adresse postale) :

Mode de recrutement (cf. Guide du recrutement)

Le mode de recrutement du·de la doctorante dépend de la nature du financement du projet de thèse.

x Concours (CDE)

☐ Entretien (préciser dates ouverture/ fermeture) :

☐ Autre (précisez) :

DESCRIPTION SCIENTIFIQUE DU PROJET DE THÈSE

Présentation du laboratoire d'accueil

L'IRHS regroupe à Angers les principaux acteurs régionaux de la recherche en Sciences du Végétal. L'IRHS est une Unité Mixte de Recherche (UMR 1345) sous les tutelles de INRAE, l'Institut Agro Rennes-Angers et l'Université d'Angers. Comptant actuellement plus de 250 agents, l'institut intègre les expertises en génétique, génomique et épi-génomique, physiologie et écophysiologie, biochimie, phytopathologie, microbiologie, modélisation, bioinformatique, biostatistiques et biophysique au service de la qualité et de la santé des espèces horticoles et de la production de semences. La thèse proposée se situe dans le second axe de recherche de l'équipe BIDEFI (BioInformatics for plant DEFense Investigation) qui regroupe 10 personnes bio-informaticiens, génomiciens et statisticiens travaillant autour de deux axes de recherche : l'identification de phytocytokines et la génomique évolutive du pommier. Cette thèse sera effectuée en collaboration avec le Laboratoire Angevin de REcherche en MAthématiques (LAREMA) de l'université d'Angers.

Contexte socio-économique et scientifique :

Les duplications de génomes anciennes, très fréquentes chez les plantes, semblent correspondre à des périodes d'extinction ou de changement global. De plus, les polyploïdes prospèrent souvent dans des environnements difficiles ou perturbés. Les polyploïdes sont considérés comme plus résilients aux environnements extrêmes en raison de leur variation génétique accrue et de l'effet tampon de leurs gènes dupliqués, ce qui a conduit à une reconnaissance accrue du potentiel adaptatif à court terme de la polyploïdie (Van de Peer *et al.,* 2017).

Le pommier a subi une WGD que nous avons daté à 27 Mya (Lallemand *et al.,* 2023). La synténie entre les chromosomes ohnologues est encore très bien conservée ce qui fait du pommier un organisme de choix pour étudier l'évolution des gènes et des familles de gènes post-WGD) (Daccord *et al*., 2017). La compréhension du rôle des chromosomes dupliqués et de leur contribution à l'élaboration du phénotype est un enjeu majeur dans le contexte de changements climatiques.

Hypothèses et questions scientifiques (8 lignes) :

La sous-dominance génomique a été démontrée chez plusieurs allopolyploïdes. Nous avons démontré pour la première fois chez un autopolyploïde un phénomène similaire chez le pommier (Lallemand *et al.* , 2023). Nous avons baptisé ce phénomène sous-dominance chromosomique par analogie et nous l'avons mis en évidence pour la plupart des paires de chromosomes ohnologues (1-7, 2-15, 3-11, 5-10, 6-14, 8-15, 9-17 et 13-16) lors d'analyses croisées de données QTLs, RNAseq, contenu en Eléments Transposables, rétention de gènes.

> *Peut-on confirmer et capter ce déséquilibre à travers la prédiction génomique, qui prend en compte les variations alléliques entre individus ?*

On s'intéressera en particulier aux paires 1-7, 3-11, 8-15, et 13-16 qui ont été montrées comme significativement déséquilibrées en nombre de QTLs.

Principales étapes de la thèse et démarche

Etape 1 :

La thèse débutera par une étude bibliographique. On étudiera Zingaretti *et al*. (2020), où les auteurs s'intéressent aux performances des réseaux de neurones chez la fraise allopolyploïde, et chez la myrtille autopolyploïde. On examinera la thèse de Tanguy Lallemand (2022) "Evolution des gènes dupliqués chez le pommier", soutenue à l'IRHS. On s'intéressera également à Jung *et al*. (2022), au sujet de la prédiction génomique dans différents environnements.

Etape 2 :

On procèdera à une étude par simulation à partir de la population de pommiers REFPOP (Jung *et al*, 2020, 2022). On simulera le phénotype en considérant différents liens possibles entre phénotype et génotype aux QTLs (additivité, epistasie, dominance, non linéarité …).
En terme de machine learning, les méthodes privilégiées seront le Genomic BLUP, les forêts aléatoires, le Lasso, l'Elastic-Net, les SVM, les RKHS, les réseaux de neurones. Pour chaque architecture de trait simulée, on pourra ainsi extraire la meilleure méthode d'apprentissage statistique capable de capter le déséquilibre entre ohnologues.
On procèdera par la suite au même type d'étude cette fois-ci sur les données réelles phénotypiques. On pourra également s'intéresser à d'autres espèces modèles allopolyploides (e.g. maïs Renny-Byfield *et al*., 2019) ou autopolyploides (myrtille Colle *et al.,* 2019).
Ceci donnera lieu à une publication appliquée.

Etape 3 :
On effectuera une étude approfondie des aspects mathématiques présents dans Chen *et al*. (Statistica Sinica, 2021) sur le réseaux de neurones, et dans Saha *et al.* (2021) pour les Forêts Aléatoires généralisées. Afin de se familiariser avec ces nouvelles méthodes, on prendra en main les packages associés : RandomForestsGLS (Saha *et al.,* 2021), et le code Python de DeepKriging (https://github.com/aleksada/DeepKriging).

Etape 4 :
On cherchera à améliorer Deep Kriging (Chen *et al.,* 2021) et les Forêts aléatoires (Saha *et al.,* 2021), en développant des formules mathématiques propres à la génomique.
On pourra notamment s'intéresser à l'erreur de prédiction, et également quantifier mathématiquement la perte d'information (en termes de précision de prédiction) lorsque les 2 chromosomes ohnologues ne sont pas inclus dans le modèle de prédiction (cf. Rabier et Grusea 2021, dans un autre contexte).

Approches méthodologiques et techniques envisagées
Données REFPOP disponibles (Jung *et al.* 2020) : la population se compose de 269 accessions et de 265 descendants issus de 27 combinaisons parentales, représentant respectivement la diversité des pommiers cultivés et le matériel de sélection européen actuel. Ces 534 génotypes sont répartis dans six pays européens permettant d'étudier les liens GxE. Les 10 traits différents ont été phénotypés (émergence florale, date de récolte, rendement, qualité du fruit …). Nous disposons également de données SNPs de haute-densité (303 329 SNPs) pour cette population.

Etant donné la proximité entre les modèles mixtes en génomique et en statistique spatiale, on s'inspirera de récents résultats mathématiques en statistique spatiale (cf. Wikle et Zammit-Mangion 2023) afin d'améliorer les méthodes existantes en prédiction génomique.
- *Réseaux de neurones* : Chen *et al.* (2021) introduisent un réseau de neurones profond où la dépendance spatiale est modélisée par l'ajout d'une couche supplémentaire permettant d'approximer le processus spatial à l'aide d'une base de fonctions.
- *Forêts Aléatoires* : Saha *et al.* (2021) proposent, afin de construire un arbre de décision, de remplacer à chaque fractionnement de noeud, le critère de moindres carrés par une optimisation prenant en compte la structure de corrélation spatiale induite par un processus Gaussien.

Compétences scientifiques et techniques requises pour le candidat
Apprentissage statistique (Forêts aléatoires, réseaux de neurones, Lasso …), Statistique en grande dimension, Modèle mixte
Maîtrise des langages de programmation en R et/ou Python
Des connaissances en évolution ou en biologie végétale seraient un plus

Bibliographie

- Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). *Deep learning: a statistical viewpoint.* Acta numerica, 30, 87-201.
- Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., ... & Edger, P. P. (2019). *Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry.* GigaScience, 8(3), giz012.
- Chen, W., Li, Y., Reich, B. J., & Sun, Y. (2021). *Deepkriging: Spatially dependent deep neural networks for spatial prediction.* Statistica Sinica:10.5705/ss.202021.0277
- Daccord N, Celton J-M, *et al*. (2017), *High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development.* Nature Genet. 49(7): 1099-1106.
- Fan, J., Ma, C., & Zhong, Y. (2021). *A selective overview of deep learning.* Statistical science: a review journal of the Institute of Mathematical Statistics, 36(2), 264.
- Jung, M., Roth, M., Aranzana, M. J., Auwerkerken, A., Bink, M., Denancé, C., ... & Muranty, H. (2020). *The apple REFPOP—a reference population for genomics-assisted breeding in apple*. Horticulture research, *7*.
- Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... & Patocchi, A. (2022). *Genetic architecture and genomic predictive ability of apple quantitative traits across environments*. Horticulture research, 9, uhac028.
- Lallemand, T. *et al.* (2023), *Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica*, Genome Biology Evolution, 15(10): evad178
- Rabier, C. E., & Grusea, S. (2021). *Prediction in high-dimensional linear models and application to genomic selection under imperfect linkage disequilibrium.* Journal of the Royal Statistical Society Series C: Applied Statistics, 70(4), 1001-1026.
- Renny-Byfield, S., Rodgers-Melnick, E., & Ross-Ibarra, J. (2017). *Gene fractionation and function in the ancient subgenomes of maize*. Molecular biology and evolution, *34*(8), 1825-1832.
- Saha, A., Basu, S., & Datta, A. (2021). *Random forests for spatially dependent data.* Journal of the American Statistical Association, 118(541), 665-683.
- Van de Peer Y *et al.* (2017), *The evolutionary significance of polyploidy*, Nat. Rev. Genet. 10:725-732.
- Wikle, C. K., & Zammit-Mangion, A. (2023). *Statistical deep learning for spatial and spatiotemporal data*. Annual Review of Statistics and Its Application, 10, 247-270.
- Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., ... & Pérez-Enciso, M. (2020). *Exploring deep learning for complex trait genomic prediction*

ENCADREMENT DE LA THÈSE

Nom de l'unité d'accueil : Institut de Recherches en Horticulture et Semences (IRHS)
UMR 1345 UA-INRAE-IARA

Nom de l'équipe d'accueil : Equipe BIDEFI : BIoinformatics for plant DEFense Investigation

Nom du·de la directeur·rice de l'unité : JACQUES MARIE-AGNES

Nom du·de la responsable de l'équipe : LANDES CLAUDINE

Coordonnées du·de la directeur·rice de l'unité :

42 rue Georges Morel, CS 60057

49 071 Beaucouzé Cedex, FRANCE

Tel : 02 41 22 57 07

Courriel : marie-agnes.jacques@inrae.fr

Coordonnées du·de la responsable de l'équipe :

Tel : 02 41 22 57 91

Courriel : claudine.landes@univ-angers.fr
claudine.landes@inrae.fr

Directeur·rice de thèse

Nom, prénom : LANDES CLAUDINE

Fonction : Professeure d'université

Date d'obtention de l'HDR : 18/10/2011

Employeur : Université d'Angers

Taux d'encadrement doctoral dans le présent sujet : 40%

Taux d'encadrement doctoral en cours (directions et co-directions) (%) : 50%

Nombre de directions/co-directions de thèse en cours : 1

Co-directeur·rice (le cas échéant)

Nom, prénom : PANLOUP FABIEN

Fonction : Professeur d'université

Date d'obtention de l'HDR : 6/12/2014

Employeur : Université d'Angers

École doctorale de rattachement : MathStic

Taux d'encadrement doctoral dans le présent projet : 30%

Taux d'encadrement doctoral en cours (directions/co-directions/co-encadrements) (%) : 150%

Nombre de directions/co-directions/co-encadrements de thèse en cours : 1 direction et une codirection en cours (étudiants tous les deux en 3eme année).

Co-encadrant·e de thèse 1 (le cas échéant)

Nom, prénom : RABIER CHARLES-ELIE

Fonction : Maitre de Conférences

Titulaire de l'HDR : ☐ oui x non Si oui, date d'obtention de l'HDR :

Employeur : Université d'Angers

École doctorale de rattachement : ED MATH-STIC

Taux d'encadrement doctoral dans le présent projet : 30%

Taux d'encadrement doctoral en cours (directions/co-directions/co-encadrements) (%) :

Nombre de directions/co-directions/co-encadrements de thèse en cours : aucune

Co-encadrant·e de thèse 2 (le cas échéant)

Nom, prénom :

Fonction :

Titulaire de l'HDR : ☐ oui ☐ non Si oui, date d'obtention de l'HDR :

Employeur :

École doctorale de rattachement :

Taux d'encadrement doctoral dans le présent projet :

Taux d'encadrement doctoral en cours (directions/co-directions/co-encadrements) (%) :

Nombre de directions/co-directions/co-encadrements de thèse en cours :

Partenaire privé (si financement CIFRE, privé, …)

Nom, prénom :

Fonction :

Entreprise :

Taux d'encadrement doctoral dans le présent projet :

Taux d'encadrement doctoral en cours (directions/co-directions/co-encadrements) (%) :

Nombre de directions/co-directions/co-encadrements de thèse en cours :

Partenaire international (si thèse en co-tutelle)

Nom, prénom :

Fonction : Employeur :

Taux d'encadrement doctoral dans le présent projet :

Taux d'encadrement doctoral en cours (directions/co-directions/co-encadrements) (%) :

Nombre de directions/co-directions/co-encadrements de thèse en cours :

Devenir des anciens doctorants du·de la directeur·rice et co-directeur(s)/co-encadrant(s) de thèse (depuis 5 ans)

Compléter les informations suivantes pour chaque ancien doctorant

Nom, prénom : LALLEMAND Tanguy

Date de début et de fin de thèse : 1/11/2019 au 30/10/2022 (soutenue le 15/11/2022)

Direction de thèse : Claudine Landès (50%) et J-Marc Celton (50%)

Emploi actuel, lieu : Ingénieur chez SEENOVIA

Contrat (post-doc, CDD, CDI) : CDI

Liste des publications issues de ce travail de thèse :
Articles issus de la thèse : 3 (dont 3 en premier auteur)
•**Lallemand T,** Leduc M, Desmazières A, Aubourg S, Rizzon C, Celton J-M, **Landès C** (2023). Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica. *Genome Biology and Evolution*, *15*(10), evad178.
•**Lallemand T**, Aubourg S, Celton J-M, **Landès C** (2022). "Chromosome dominance in apple after whole genome duplication". ISHS Acta Horticulturae 1362, doi:10.17660/ActaHortic.2023.1362.9
•**Lallemand T**, Leduc M, **Landès C**, Rizzon C, Lerat E (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes*, *11*(9), 1046.

Nom, prénom : LEDUC Martin

Date de début et de fin de thèse : 1/10/2019 au 30/09/2022 (soutenue le 5/12/2022)

Direction de thèse : Claudine Landès (40%), Nathalie Leduc (30%) et Jérémy Clotault (30%)

Emploi actuel, lieu : Ingénieur NDP Systèmes

Contrat (post-doc, CDD, CDI) : CDI

Liste des publications issues de ce travail de thèse : 2 (dont une en premier auteur)
•Lallemand T, **Leduc M**, Desmazières A, Aubourg S, Rizzon C, Celton J-M, Landès C (2023). Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica. *Genome Biology and Evolution*, *15*(10), evad178.
•**Leduc M**, Lallemand T, Landès C, Rizzon C, Lerat E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes*, *11*(9), 1046.

Publications majeures des 5 dernières années du·de la directeur·rice de thèse et co-directeur(s)/co-encadrant(s) sur le sujet de thèse :
Lallemand T, Leduc M, Desmazières A, Aubourg S, Rizzon C, Celton J-M, **Landès C** (2023). "Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica". Genome Biol. Evol. 15(10), https://doi.org/10.1093/gbe/evad178

Lallemand T, Aubourg S, Celton J-M, **Landès C** (2022). "Chromosome dominance in apple after whole genome duplication". ISHS Acta Horticulturae 1362, doi:10.17660/ActaHortic.2023.1362.9

Lallemand T, Leduc M, **Landès C**, Rizzon C, Lerat E (2020). "An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice". Genes, 11(9):1046, https://www.mdpi.com/2073-4425/11/9/1046.

C-E Rabier, S Grusea. "Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium", Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol 70(4), 2021, doi.org/10.1111/rssc.12496.

C-E Rabier, B Mangin, S Grusea. "On the accuracy in high dimensional models and its application to genomic selection", Scandinavian Journal of Statistics, Vol 46(1), 2019, https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12352.

S Gadat, **F Panloup**. "Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity". Stochastic Processes and Applications. 156, 2023. 312–348.

FINANCEMENT DE LA THÈSE

Origine(s) du financement de la thèse : ED VAAME

Salaire brut mensuel :

- 2024 : 2 100.00 € brut mensuel soit un coût chargé de 2 957.22 €
- 2025 : 2 200.00 € brut mensuel soit un coût chargé de 3 098.04 €
- 2026 : 2 300.00 € brut mensuel soit un coût chargé de 3 238.86 €

État du financement de la thèse : Demande

Date du début/durée du financement de la thèse

(Au format JJ/MM/AA, pour renseigner le contrat) : 01/10/2024

Date : le 19 mars 2024

Nom, signature du·de la directeur·rice d'unité :

Marie-Agnès JACQUES

Nom, signature du·de la responsable de l'équipe :

Claudine LANDES

Nom, signature du·de la directeur·rice de thèse :

Claudine LANDES

Toutes les rubriques de ce document doivent être remplies.