COLLEGE VEGETAL
DOCTORAL ANIMAL, ALIMENT
PAYS DE LA LOIRE MER, ENVIRONNEMENT

PhD PROPOSAL FOR THE DOCTORAL SCHOOL « Végétal, Animal, Aliment, Mer, Environnement »

GENERAL INFORMATION

Thesis title: Genomic Prediction for Exploring Subdominance in PolyPLOID Genomes

Acronym of the project: GP4PLOID

Disciplinary field 1: Statistics

Disciplinary field 2: Evolutionary genomics

Three keywords: Statistical learning, Genomic prediction, Evolution

Registration establishment: University of Angers

Research unit: Institute of Research in Horticulture and Seeds (IRHS)

Name of the thesis director HDR (Accreditation to supervise research) required: LANDES

Email address of the thesis director: claudine.landes@univ-angers.fr

Name of the thesis co-director (if applicable): HDR (Accreditation to supervise research) required: PANLOUP Fabien

Email address of the thesis co-director (if applicable): fabien.panloup@univ-angers.fr

Name of the thesis co-supervisor 1 (if applicable): RABIER Charles-Elie

Email address of the thesis co-supervisor 1 (if applicable): charles-elie.rabier@univ-angers.fr

Name of the thesis co-supervisor 2 (if applicable):

Email address of the thesis co-supervisor 2 (if applicable):

Contact(s) (mailing address and E-mail):

claudine.landes@univ-angers.fr, fabien.panloup@univ-angers.fr, charles-elie.rabier@univ-angers.fr

☒ **Doctoral school contest**

☐ **Interview**

☐ **Other (specify):**

SCIENTIFIC DESCRIPTION OF THE PhD PROJECT

Presentation of the host laboratory

The Institute of Research in Horticulture and Seeds (IRHS) gathers in Angers the main regional actors of the research in Plant Sciences and enjoys a privileged working environment in a dynamic scientific and teaching environment. The IRHS is a Joint Research Unit (UMR 1345) under the supervision of INRAE, the Rennes-Angers Agro Institute and the University of Angers. With currently more than 250 agents, including 183 permanent staff, it integrates expertise in genetics, genomics and epigenomics, physiology and ecophysiology, biochemistry, plant pathology, microbiology, modeling, bioinformatics, biostatistics and biophysics for the quality and health of horticultural species and seed production. The current thesis is proposed by the BIDEFI (BioInformatics for plant DEFense Investigation) team, which comprises 10 bioinformaticians, genomicists and statisticians working on two lines of research: the identification of phytocytokines and the evolutionary genomics of apple trees. This thesis will be carried out in collaboration with the *Laboratoire Angevin de REcherche en MAthématiques* (LAREMA) at the University of Angers.

Socio-economic and scientific context

Ancient genome duplications, which are very common in plants, seem to correspond to periods of extinction or global change. Moreover polyploids often thrive in harsh or disturbed environments. Polyploids are considered more resilient to extreme environments due to their increased genetic variation and the functional buffering of their duplicated genes, which has led to increased recognition of the short-term adaptive potential of polyploidy (Van de Peer *et al.,* 2017). Apple underwent a WGD that we dated at 27 Mya (Lallemand *et al.,* 2023). Synteny between ohnologous chromosomes is still highly conserved, making the apple tree an organism of choice for studying the evolution of genes and gene families post-WGD) (Daccord *et al.,* 2017). Understanding the role of duplicated chromosomes and their contribution to phenotype development is a major challenge in the context of climate change.

Assumptions and questions (8 lines)

Genomic subdominance has been shown in several allopolyploids. We described a similar phenomenon in apple for the first time in an autopolyploid (Lallemand et al. , 2023). We have named this phenomenon chromosomal subdominance by analogy, and have shown it for most ohnologous chromosome pairs (1-7, 2-15, 3-11, 5-10, 6-14, 8-15, 9-17 and 13-16) in cross-analyses of QTLs, RNAseq, Transposable Element content and gene retention data.

Can we confirm and capture this imbalance through genomic prediction, which takes into account allelic variations between individuals?

The main steps of the thesis and scientific procedure

Step 1 : The thesis will begin with a literature review. We will study Zingaretti *et al* (2020), where the authors focus on the performance of neural networks in allopolyploid strawberry, and autopolyploid blueberry. Tanguy Lallemand's thesis (2022) "Evolution of duplicated genes in apple", defended at IRHS, will be considered. We will also look at Jung *et al.* (2022), on the subject of genomic prediction in different environments.

Step 2 : A simulation study will be carried out using the REFPOP apple population (Jung *et al*, 2020, 2022). The phenotype will be simulated by considering various possible links between phenotype and genotype at QTLs (additivity, epistasis, dominance, non-linearity, etc.). In terms of machine learning, the preferred methods will be Genomic BLUP, random forests, Lasso, Elastic-Net, SVM, RKHS and neural networks. For each simulated trait architecture, we'll be able to extract the best statistical learning method capable of capturing the imbalance between ohnologs.

Step 3 : The mathematical aspects present in Chen et al (Statistica Sinica, 2021) on neural networks, and in Saha *et al* (2021) for Generalized Random Forests, will be studied in details. We will start with their associated packages: RandomForestsGLS (Saha *et al.,* 2021), and the DeepKriging Python code (https://github.com/aleksada/DeepKriging).

Step 4 : We will seek to improve Deep Kriging (Chen *et al.,* 2021) and Random Forests (Saha *et al.*, 2021), by developing mathematical formulas specific to genomics. In particular, we could look at prediction error, and also mathematically quantify the loss of information (in terms of prediction accuracy) when the 2 ohnologous chromosomes are not included in the prediction model (cf. Rabier and Grusea 2021, in another context).

Methodological and technical approaches considered

REFPOP data available (Jung et al. 2020): the population consists of 269 accessions and 265 progenies from 27 parental combinations, representing respectively the diversity of cultivated apple trees and current European breeding material. These 534 genotypes are distributed across six European countries, enabling GxE relationships to be studied. The 10 different traits were phenotyped (floral emergence, harvest date, yield, fruit quality, etc.). We also have high-density SNP data (303,329 SNPs) for this population.

Given the proximity between mixed models in genomics and spatial statistics, we will built on recent mathematical results in spatial statistics (Wikle and Zammit-Mangion 2023) to improve existing methods in genomic prediction.

- Neural networks: *Chen et al.* (2021) introduce a deep neural network where spatial dependence is modeled by adding an extra layer to approximate the spatial process using a basis of functions.

- Random Forests: Saha *et al.* (2021) suggest, in order to build a decision tree, to replace the least-squares criterion at each node split by an optimization taking into account the spatial correlation structure induced by a Gaussian process.

Scientific and technical skills required by the candidate

- Statistical learning (random forests, neural networks, Lasso, etc.), high-dimensional statistics, mixed models
- Skills in R and/or Python programming languages
- Knowledge of evolution or plant biology will be appreciated

COLLEGE VEGETAL
DOCTORAL ANIMAL, ALIMENT
PAYS DE LA LOIRE MER, ENVIRONNEMENT

THESIS SUPERVISION

Unit name: IRHS (Institute of Research in Horticulture and Seeds)	**Team name:** BIDEFI
Unit director name: Marie-Agnès Jacques	**Team director name:** Claudine Landès
Mailing address of the unit director: marie-agnes.jacques@inrae.fr	**Mailing address of the team director:** claudine.landes@inrae.fr

Thesis director

Surname, first name: LANDES Claudine

Position: University Professor

Obtained date of the HDR (Accreditation to supervise research): 18/10/2011

Employer: University of Angers

Doctoral school affiliation: ED VAAME

Rate of thesis supervision in the present project (%): 40%

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%): 50%

Number of current thesis supervisions/co-supervisions: 1

Thesis co-director

Surname, first name: PANLOUP Fabien

Position: University Professor

Obtained date of the HDR (Accreditation to supervise research): 6/12/2014

Employer: University of Angers

Doctoral school affiliation: ED MathSTIC

Rate of thesis supervision in the present project (%): 30%

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%): 150

Number of current thesis supervisions/co-supervisions: 2 (currently in final year)

Thesis co-supervisor 1 (if applicable)

Surname, first name: RABIER Charles-Elie

Position: Assistant professor

Accreditation to supervise research ☐ yes ☒ no If yes, date diploma received:

Employer: University of Angers

Doctoral school affiliation: ED MATH-STIC

Rate of thesis supervision in the present project (%): 30%

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%):

Number of current thesis supervisions/co-supervisions:

Thesis co-supervisor 2 (if applicable)

Surname, first name:

Position:

Accreditation to supervise research ☐ yes ☐ no If yes, date diploma received:

Employer:

Doctoral school affiliation:

Rate of thesis supervision in the present project (%):

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%):

Number of current thesis supervisions/co-supervisions:

Private partner (if CIFRE funding, private funding…)

Surname, first name:

Position:

Employer:

Rate of thesis supervision in the present project (%):

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%):

Number of current thesis supervisions/co-supervisions:

International partner (if Cotutelle thesis)

Surname, first name:

Position:

Employer:

Rate of thesis supervision in the present project (%):

Total rate of thesis supervision in ongoing theses (supervisions and co-supervisions) (%):

Number of current thesis supervisions/co-supervisions:

Professional status of previous PhD students supervised by both director and co-supervisors (from 5 years)

Please provide the following information for each PhD students supervised

Surname, first name: LALLEMAND Tanguy

Date of PhD beginning and PhD defence: 1/11/2019 – 15/11/2022

Thesis supervision: Claudine Landès (50%) and Jean-Marc Celton (50%)

Professional status and location: engineer at seenovia in Laval
Contract profile (post-doc, fixed-term, permanent): permanent

List of publications from the thesis work:
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.

Surname, first name: LEDUC Martin

Date of PhD beginning and PhD defence: 1/10/2019 – 5/12/2022

Thesis supervision: Claudine Landès (40%) , Nathalie Leduc (30%), Jérémy Clotault (30%)

Professional status and location: engineer at NDP Systems in Angers
Contract profile (post-doc, fixed-term, permanent): permanent

List of publications from the thesis work:
Leduc M, Lallemand T, 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, 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.

Five main recent publications of the supervisors on thesis subject:
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, 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.

THESIS FUNDING

Origin(s) of the thesis funding: ED VAAME
Gross monthly salary: 2100€ (2024), 2200€ (2025), 2300€ (2026)
Thesis funding state: Non acquired
Funding beginning date/duration of the thesis funding: 1/10/2024

Date: March 20, 2024

Name, signature of unit director:

Marie-Agnès Jacques

Name, signature of team director:

Claudine Landès

Name, signature of thesis project director:

Claudine Landès

All sections must be filled in. Once completed, please save the proposal form in PDF format using the following naming: Supervisor Name_Unit_Subject Acronym_EN.pdf
Please also send a Word version to make it easier to change the layout if necessary.

Documents to be send to: ed-vaame@doctorat-paysdelaloire.fr