Master Biologie Agrosciences
M2

Rapport de stage

Characterization and Genome-Wide Association Study of Nut Related Traits in Walnut (*Juglans regia L.*)

Julie Crabier

Stage réalisé du 06 Janvier au 26 Juin 2020
Maitre de stage: Elisabeth Dirlewanger
Co-encadrant: Anthony Bernard

Remerciements

Je tiens à remercier le **GIS Fruits** qui, en partenariat avec l'INRAE, le CTIFL et la FNFP, a financé ce stage.

Un grand merci à **Elisabeth Dirlewanger**, Directrice de Recherche à l'INRAE, unité BFP, équipe A3C, et **Anthony Bernard**, Doctorant Cifre à l'INRAE et au CTIFL, de m'avoir permis de réaliser ce stage mais aussi de m'avoir encadrée, soutenue et encouragée durant toute cette période.

Je remercie **Karima Giresse**, Directrice du CTIFL centre opérationnel de Lanxade, **Fabrice Lheureux**, Ingénieur Chef de Programme, et tout le personnel du CTIFL de m'avoir accueillie et accompagnée dans cet établissement pendant un mois.

Je remercie également **Teresa Barreneche**, **Hélène Christmann**, **Mathieu Fouche**, **Lydie Fouilhaux**, **Jacques Joly**, **Xavier Lafon**, **Loïck Le Dantec**, **José Quero-Garcia** et **Bénédicte Wenden** de m'avoir accueillie chaleureusement au sein de l'équipe A3C.

Je remercie **Camille Branchereau** et **Tina Ternjak**, Doctorantes à l'INRAE, avec qui j'ai passé d'agréables moments à discuter.

Je remercie **Armel Donkpegan**, Post-Doctorant à l'INRAE, de m'avoir aidée dans mon travail et d'avoir répondu à toutes mes questions, parfois compliquées.

Enfin, je remercie **Solenne, Audrey** et **Elina** pour leurs encouragements et leur aide précieuse à toute heure de la journée.

INRAE:

INRA, or the “Institut National de la Recherche Agronomique”, is a French agricultural research organization founded in 1946. Since January 2020, this institute has merged with the « Institut national de la Recherche en Sciences et Technologie pour l’Environnement et l’Agriculture » (IRSTEA) to become the INRAE, the “Institut National de Recherche pour l’Agriculture, l’Alimentation et l’Environnement”. Today, it is the 1st agricultural research institute in Europe but also the 2nd worldwide in number of publications. The INRAE is divided into 14 scientific divisions, corresponding to 14 different themes, including Plant Biology and Breeding. It is present on 18 regional sites, including the Nouvelle-Aquitaine Bordeaux center where I am doing my Master 2 internship (https://www.inrae.fr/nous-connaitre#chiffres).

L’Unité Mixte de Recherche (UMR) 1332 Biologie du Fruit et Pathologie: this unit is a partnership between the "Plant Health and Environment", "Plant Biology and Breeding" divisions of INRAE, and the University of Bordeaux (https://www6.bordeaux-aquitaine.inrae.fr/bfp). Among this UMR, there is the A3C team in which I work, which focuses on the Adaptation of the Cherry Tree to Climate Change (A3C). Their research aims to understand the response to climate change in the cherry tree, at different levels: genetics, epigenetics, physiological mechanisms. The long-term objective would be to improve the marker-assisted selection scheme in the cherry tree for adaptation to climate change and the production of quality fruits. The cherry tree activities are not the only ones within this team, which has been carrying out work on the plum tree and the walnut tree for three years (https://www6.bordeaux-aquitaine.inrae.fr/bfp/Recherche/Equipe-Adaptation-du-Cerisier-au-Changement-Climatique).

CTIFL:

Part of the activities of the internship were carried out at the “Centre Technique Interprofessionnel des Fruits et Légumes” (CTIFL's) operational center of Lanxade. The CTIFL is a research and development organization for the fruit and vegetable sector, from production to distribution. It was created in 1952 and is currently based in Paris. There are also five stations in addition to the head office, all located in French fruit and vegetable production areas: Carquefou, Lanxade, Balandran, St Rémy and Rungis. The Lanxade center concentrates its work on arboriculture and vegetable crops. Thus, the species studied are apple, pear, strawberry, kiwi, carrot, chestnut, hazelnut and walnut (http://www.ctifl.fr/Pages/Ctifl.aspx).

Table of content

Abbreviation List

- BIC: Bayesian Informative Criterion
- BLUPs: Best Linear Unbiased Predictions
- CTIFL: Centre Technique Interprofessionnel des Fruits et Légumes
- FAO: Food and Agriculture Organization of the United States
- FarmCPU: Fixed and random model Circulating Probability Unification
- GWAS: Genome-Wide Association Study
- INRAE: Institut National de Recherche pour l’Agriculture, l’alimentation et l’Environnement
- IRSTEA: Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture
- KASP: Kompetitive Allele Specific PCR
- KNOX: Knotted-like homeobox
- LD: Linkage Desequilibrium
- MLM: Mixed Linear Model
- MLMM: Multi-Locus Mixed-Model
- PC: Principal Component
- PCA: Principal Components Analysis
- PCR: Polymerase Chain Reaction
- QTL: Quantitative Trait Locus
- SNP: Single Nucleotide Polymorphism
- TALE: Three Amino-acid Loop Extension
- TPX2: Targeting Protein for Xklp2
- UMR: Unité Mixte de Recherche

I. Introduction:

Walnut is a tree of the family *Juglandaceae* and the genus *Juglans* which gathers more than 20 different species, divided into three botanical categories (Manning 1978). It is a monoecious, dichogamous and diploid plant with 2n = 2x = 32 chromosomes (Woodworth 1930). Its dispersion is mainly due to the wind (Germain et al. 1999). Among the species, the common walnut or *Juglans regia*, native to the Mountains of Central Asia, is of particular interest to us. Indeed, it is one of the most economically important walnut trees in the world and is widely cultivated for its fruit and wood. According to the Food and Agriculture Organization of the United Nations (FAO), in 2018 its production amounted to more than 3.6 million tonnes worldwide (FAO 2019), with China and the United States as the main producers. As the 7th producer in the world, France is facing increasing competition on the nuts market. Added to this, the effects of climate change can be observed more and more in orchards nowadays; the increase in average temperatures and the late frosts cause a loss of production (Bernard et al., 2018). The significant damage caused to crops by plant pathogens and pests must also be considered (Luedeling et al. 2011). Because of this, new studies were launched on *Juglans regia*. Thus, the genome of the "Chandler" variety was fully sequenced in 2016 by a team from the University of Davis, California, then resequenced in 2018 with 27 other accessions (Martínez-García et al. 2016; Stevens et al. 2018). It is from these data that an Axiom^{TM} DNA chip of 609,000 SNPs was developed, thus allowing the genotyping of this species (Marrano et al. 2019a). Nowadays, breeding programs are focusing on the creation of varieties adapted to these changes, while remaining efficient both in productivity and in quality.

It is in this context that the "INNOV'noyer" project, funded by the Nouvelle Aquitaine Region and leaded by the CTIFL of Lanxade in partnership with INRAE-BFP at Bordeaux and University of Davis, California. This project aims to study genetic diversity as well as phenotypic variability within the collection of genetic resources of INRAE, but also to identify the genetic determinism of traits of agronomic interest (http://www.arboriculture-fruitiere.com/articles/technique-fruit/insuffler-un-vent-de-nouveaute-dans-la-filiere-nucicole-francaise). A first Genome-Wide Association Study (GWAS) has already been carried out on phenological traits and lateral bearing on this collection, discovering new markers and improving knowledge on the genetic determinism of flowering in walnut (Bernard et al. 2020).

My internship is part of this project for the characterization of the fruits in the collection for quality criteria; weight, size, compressive force required to break the shell and yield. Then, the goal is to identify Single Nucleotide Polymorphism (SNP) markers linked to these traits and usable in selection, thanks to a GWAS analysis. The work carried out during this internship corresponds to the phenotyping of an additional year and the realization of a GWAS. Once SNPs showing an association with the studied traits identified, a search for corresponding candidate genes will be carried out, then they will be transformed in the future into Kompetitive Allele Specific PCR (KASP) markers, usable in Marker-Assisted Selection (MAS).

II. Material and Methods

1. Plant material

The plant material used comes from a collection of genetic resources of the genus *Juglans* belonging to the INRAE of Nouvelle-Aquitaine Bordeaux (Fig 1). It is an *in vivo* and *ex situ* germplasm collection maintained at the *Prunus-Juglans* Genetic Resources Center, in the Experimental Arboricultural Unit of Toulenne. It is the result of significant research work around the world carried out by Éric Germain, head of the INRAE walnut breeding program between 1988 and 2000. It includes a large part of the species that can be found within the genus *Juglans*. This collection is made up of more than 400 trees of 16 different species, aged 20 to 30 years and grafted on different rootstocks. These trees are planted in an orchard with sandy-loamy soil type, with an oceanic climate, an average of 850 mm / year of precipitation and low risk of frost.

Of these 400 individuals, almost 200 accessions belong to the *Juglans regia* species and come from 23 different countries; America (United States, Chile, Canada), from all over Europe (France, Spain, Portugal, United Kingdom, Bulgaria, Romania, Hungary, Ukraine, Russia, Greece), the Middle East (Turkey, Iran, Afghanistan) as well as Central and East Asia (China, India, Japan).

For the work carried out during this internship, a subset of 170 accessions of this collection was used. It was created according to an analysis of the genetic diversity using microsatellites and represents the greatest diversity within the initial collection (Bernard et al., 2018). The nuts were harvested and gathered by accessions in batches.

Fig. 1. INRAE *Juglans* genetic resources collection.

2. Nut processing (not carried out during the internship)

Fresh nuts were dried just after the harvest according to the principle of drying by ventilation to reach a water content of less than 12%; hot, dry air will remove the water from the nuts. For this, a false-bottom dryer was used. Once dried, the nuts are kept in batches in a cool place at 2 to 4°C. This step was performed at CTIFL.

3. Phenotyping

From each batch of dried nuts, 100 nuts were selected randomly but based on their sanitary state, and then weighed. These nuts are then passed through a grader (Fig 2) in order to sort them according to their size. We will distinguish seven different sizes depending on the diameter of the nut: less than 28 mm, 28-30 mm, 30-32 mm, 32-34 mm, 34-36 mm, 36-38 mm and finally more than 38 mm. For each size, the number of nuts has been counted and the whole weighed. Then, 50 nuts are randomly selected from the initial batch of 100 and the compression force necessary to break these nuts is measured in Newton using the texturometer: 25 on the side of the suture and 25 on the face of the nut. Finally, the kernels are extracted from the 50 walnuts using a nutcracker and weighed to obtain the weight of 50 kernels as well as the breaking yield. The latter is calculated by dividing kernels weight by the weight of the 50 inshell walnuts.

During the internship, only the harvest of 2019 was phenotyped, the years 2017 and 2018 having already been made. This step was also performed at CTIFL.

A texturometer is a device used to study the physical properties of an object by applying different forces to it. In our case, a compressive force is applied to the shell of the dry nut and a force sensor will translate the deformation of the support. A specific shell nut protocol previously developed by a trainee from CTIFL was used. The device available is a TA-PLUS model, from the Lloyd Materials Testing ™ brand (Fig 3).

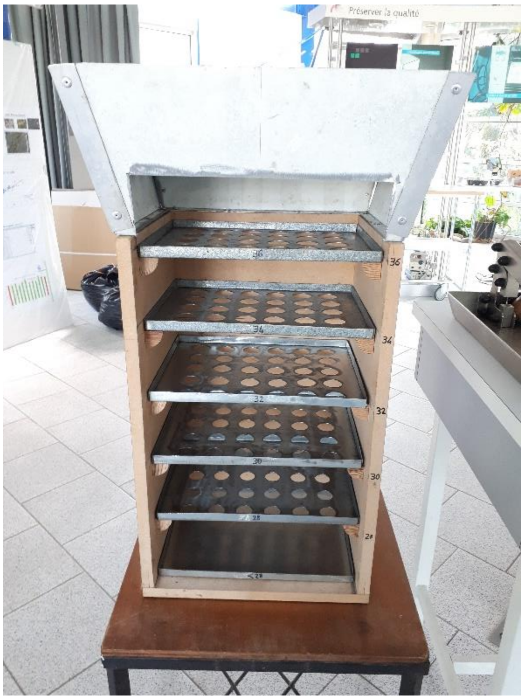


Fig. 2: Grader

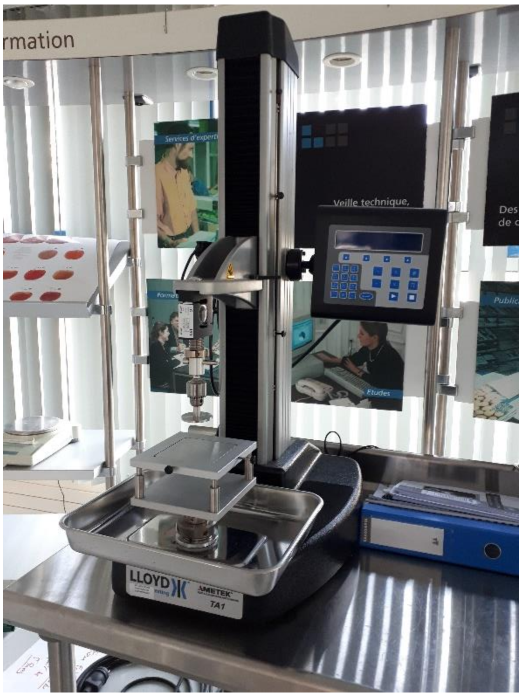


Fig. 3: Texturometer TA-PLUS Lloyd Materials Testing^{TM}

4. Analysis of phenotyping data

At INRAE, all the phenotyping data over the three years were compiled in an Excel table, from which calculations were made:

- Weight of 100 nuts, or estimated weight for few batches that contained a bit less than 100 nuts. Similarly, the weight of the nuts for each size per 100 nuts was estimated.
- The percentage of the total weight for each size, which gives us a representation of the share of each size on the batch, for each accession.
- Weight of 50 kernels, or estimated, again as some batches contained a bit less than 50 nuts during the compression force assessment step.

Then, the data was visualized on Rstudio thanks to the **tidyverse** (Wickham 2019), **corrplot** (Wei et al. 2017) and **PerformanceAnalytics** (Peterson et al. 2020) packages, which make it possible to produce histograms to observe the distribution of the data as well as correlation matrices, with calculation of the p-values of the correlation coefficients.

For each trait, broad-sense heritability H^2 was calculated using the following formula:

$$H^2 = \frac{\sigma_G^2}{[\sigma_G^2 + (\frac{\sigma_\varepsilon^2}{N_y})]}$$

With:

σ_G^2= variance of the genotypic effect

σ_ε^2= variance of the residuals

N_y= number of years = 3 (from 2017 to 2019)

The variances were calculated using the mixed linear model **lme4** package of Rstudio (Bates et al. 2015).

5. Genotyping data (Obtained before the internship)

Genotyping data was previously obtained using the 609,000 SNP Axiom^{TM} DNA chip, developed in 2019 by an American team from University of Davis in California (Marrano et al. 2019a). The data set of genotyping results is published and available for free access in the previous GWAS analysis on flowering (Bernard et al. 2020).

After receiving genotyping data, it is important to filter it based on the quality of the SNPs. 364,275 SNPs were kept after data cleaning.

6. Structure of the population (not performed during the internship)

The power of a GWAS is strongly linked to the structure of the population used. Indeed, the more genetically close the individuals within a population are, the less effective is the detection of associated SNPs; we increase the possibility of getting false positives. Therefore, the structure must be studied and integrated into the GWAS model to limit these risks. Usually, for this the STRUCTURE software is used; however, there are other methods more suitable for large datasets as in our case. Here, it is the sNMF function of the **LEA** (Frichot et al., 2015) package on the Rstudio software that is used to determine the best number of K genetic groups in our population. Two distinct groups have been identified, depending on the origin of the accessions: one group representing Western Europe and America, and a second group representing Eastern Europe and Asia.

7. GWAS

GWAS is a genetic analysis that identifies loci or alleles whose polymorphism is involved in the variation of a phenotype. The aim is therefore to study a possible statistical association between phenotypic variation and genetic variation within a large population. As part of this analysis, after the phenotyping and genotyping stages, comes the statistical analysis itself which is carried out thanks to the use of linear models. A linear model follows the following formula:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{R}$$

With **Y** as the phenotype, which is the variable to be explained, **μ** the mean, **Xβ** as the covariates, **Gγ** the effects of the tested SNP and **R** the residuals. The genetic variation corresponds to the explanatory variable.

There are different approaches to GWAS with different linear models. Here, it was performed on Rstudio using the GAPIT (Lipka et al. 2012) package.

The method used here is called the "two-steps approach" and is meant for large datasets gathering two or three years. For this, the work is divided into two steps: first, Best Linear Unbiased Predictions (BLUPs) were calculated to integrate the effect of the year and were used as phenotyping data for the rest of the analysis. Then, the kinship matrix and Principal Component Analysis (PCA) matrix, which account for the structure of the population and the relatedness between individuals, are considered in the covariates (**Xβ**). For this, the "model selection" function of the package will first be used to recalculate the population structure using a Mixed Linear Model (MLM) and determine the number of groups present. This number of Principal Components (PC) to integrate into the GWAS is selected with the Bayesian Information Criterion (BIC): a high score indicates the number of ancestral groups most suitable for the analysis. Finally, the analysis itself is carried out by applying two multi-locus models to determine the significant associations; MLMM (Segura et al. 2012) and FarmCPU (Liu et al. 2016).

MLMM is a Multi-Locus Mixed-Model, which integrates both a kinship matrix and cofactors. Unlike a single-locus model, it will consider several loci at the same time. By integrating cofactors and the kinship matrix at the same time, MLMM limits the appearance of "false positives" and allows a greater detection power than a model integrating only one of the two (Kaler et al. 2020).

Fixed And Random Model Circulating Probability Unification (FarmCPU) is a newer model based on the MLMM model. It divides MLMM into two sub-models: a random effects model and a fixed effects model. FarmCPU was demonstrated as the most able to control the occurrence of false positives and false negatives compared to other models (Kaler et al. 2020).

BLUPs, or Best Linear Unbiased Predictions, are generally used in mixed linear models to predict random effects. In our case, genotypic effects are considered as random whereas environmental factors, the “year” effect here, are considered as fixed. BLUPs will allow data to be adjusted and new data centred on zero. In GWAS, they are used to consider the effect of several factors as covariates in the model, such as the year or the place of culture.

As the analysis progresses, the GAPIT package will create several files for each multi-locus model, containing the results in the form of a graph called Manhattan Plot, Quantile-Quantile Diagrams (Q-Q plot) as well as several Excel files with, among others, the names and positions of the different associations detected.

At the end of the analysis, it was necessary to adjust the Bonferroni significance threshold of the Manhattan plot, automatically established at 1% risk of error by the software, so that it corresponds to a threshold of 5% which is less strict and usually used in multi-locus GWAS models.

8. Linkage Disequilibrium (LD)

We talk about Linkage Disequilibrium when alleles at different loci are associated in a non-random pattern and transmitted together to the descendants. Their frequency therefore differs from a theoretical allelic frequency. Moreover, it is important to note that an association detected in a LD block can hide another position in the same block and actually associated with the characteristic studied. Therefore, it is important to be interested in these in association genetics. By looking for all the genes included in the LD block of the detected association, it is possible to potentially find candidate genes linked to the trait studied.

In this case, LD blocks were analyzed on the Haploview software (Barrett et al. 2005), thanks to its “Linkage Format” function. For each trait studied, the analysis start with two files created beforehand on the Plink software, containing the data from the SNPs on a window of 100,000 bp on either side of the association detected during the GWAS. This window is arbitrarily determined within the limits of the operation of the software (200,000 bp maximum in total). Two methods of analysis to detect these LD blocks were used; Solid Spine of LD and Confidence Interval (by Gabriel et al.). The first is less strict and usually gives larger LD blocks. Haploview will display all the SNPs in the window indicated at the start, as well as their LD blocks, in the form of an "LD plot". Each SNP is assigned a number, so we simply search for the number of the SNP detected during the GWAS analysis on the graph and identify the limits of the LD block to which it belongs. For SNPs in complete linkage equilibrium, a window of 50 kb on both sides was investigated.

9. Candidates Genes

The search for candidate genes was done on the Rstudio software, using a script requiring two types of files to operate: a first file containing the annotations of the genome of the walnut *Juglans regia* (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Juglans_regia/100/) and a second file with the information found during GWAS analysis on associated SNPs (chromosome, name, position, bounds of the LD block and associated character). The latter is in two copies in our situation; one for each analysis method used when searching for LD blocks.

At first, the script only search for the full-fledged genes included in the limits, which it will compile in a first document. Then, it searches among all the annotations and not just the coding sequences, producing a second document with many more matches. In the end, 24 documents are obtained: four documents per trait, including two for each LD blocks analysis method. For the following analyses, only the results obtained with Solid Spine of LD including all the annotations were kept.

For each trait, each associated SNP can correspond to several annotations, identified by a “LOC” number. It is then a question of finding the molecule corresponding to the LOC number on an annotation file of the reference sequence. In this way, we have several names of molecules, known or not, for each association detected in an LD block, the function of which must be sought.

10. Calculation of R^2 and allelic effect

In association genetics, when working on quantitative traits, each association detected for a trait can explain a certain percentage of the phenotypic variation observed. This value is called R^2 and is calculated using the GAPIT package during the GWAS analysis, for each association. However, for each significant association found, it is necessary to adjust this value taking into account the R^2 value of all associations; this is the “genome-wide correction”. For this, the average of R^2 of all the other markers is subtracted from the value R^2 of the significant marker. However, given the significant time that this process requires, it is possible to use the average of only 5% of the markers randomly selected, the value remaining unchanged. This calculation was done using a script provided by the American team from University of Davis in California also working on nuts; thereby we obtain the average R^2 of 5% of the markers randomly selected, which we can then subtract from the R^2 of our significant associations.

The allelic effect is the difference in mean of the trait measured with one allele or the other. It is given by the GAPIT package during the GWAS analysis. The sign of the allelic effect is linked to the "major" allele of the trait, the allele present in higher proportion in the population. For instance, a negative value for the allelic effect means that the major allele is linked to a low value for the trait studied (a low weight or compression force for example).

III. Results

1. Phenotyping data analysis

The data collected during phenotyping over the three years were gathered to be compared. The means and standard deviations are relatively constant over the years, for all traits studied (Table 1). For instance, for the weight of 100 nuts, the mean was of 1,109.64 g in 2017, 1,192.80 g in 2018, and 1,152.86 g in 2019. However, when observing the ranges, we note that variation is important within a year, particularly for the weight of 100 nuts (minimum of 521.74 g and maximum of 2,278.20 g for 2017 for example), due to the different accessions. After calculating broad-sense heritability, we obtain high values ranging from 0.88 to 0.95, both suture and face strength having lower values with 0.89 and 0.88, respectively.

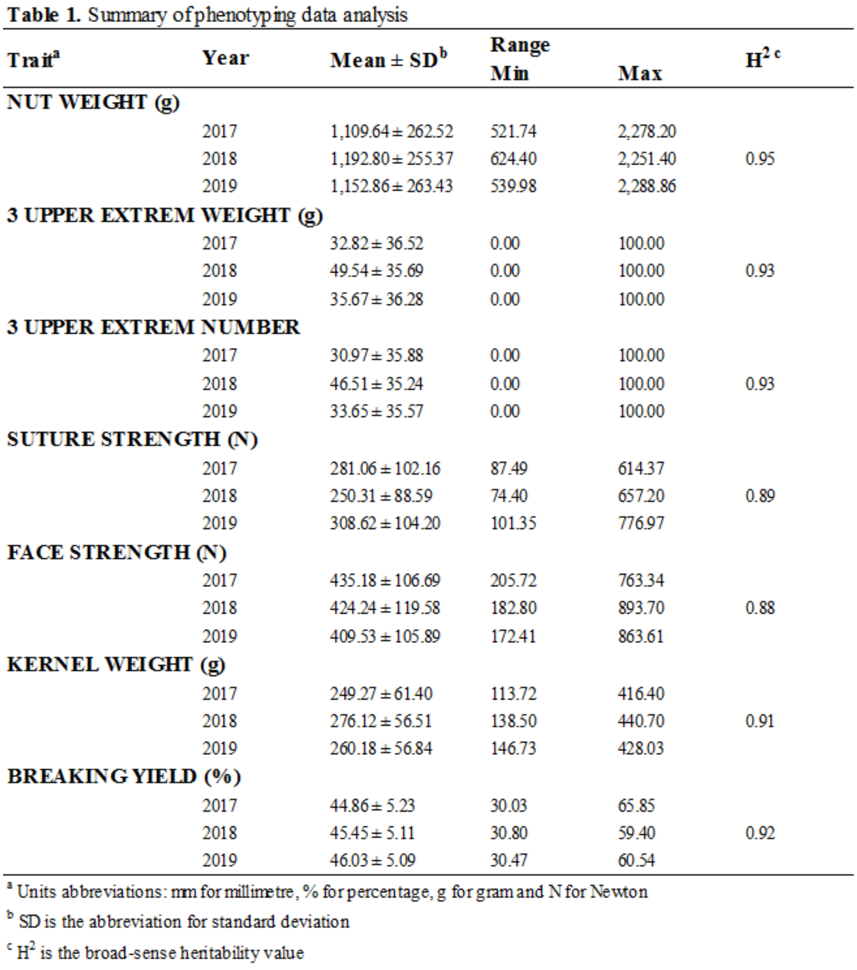


Table 1. Summary of phenotyping data analysis

Trait[a]	Year	Mean ± SD[b]	Range Min	Max	H^2 [c]
NUT WEIGHT (g)					
	2017	1,109.64 ± 262.52	521.74	2,278.20	
	2018	1,192.80 ± 255.37	624.40	2,251.40	0.95
	2019	1,152.86 ± 263.43	539.98	2,288.86	
3 UPPER EXTREM WEIGHT (g)					
	2017	32.82 ± 36.52	0.00	100.00	
	2018	49.54 ± 35.69	0.00	100.00	0.93
	2019	35.67 ± 36.28	0.00	100.00	
3 UPPER EXTREM NUMBER					
	2017	30.97 ± 35.88	0.00	100.00	
	2018	46.51 ± 35.24	0.00	100.00	0.93
	2019	33.65 ± 35.57	0.00	100.00	
SUTURE STRENGTH (N)					
	2017	281.06 ± 102.16	87.49	614.37	
	2018	250.31 ± 88.59	74.40	657.20	0.89
	2019	308.62 ± 104.20	101.35	776.97	
FACE STRENGTH (N)					
	2017	435.18 ± 106.69	205.72	763.34	
	2018	424.24 ± 119.58	182.80	893.70	0.88
	2019	409.53 ± 105.89	172.41	863.61	
KERNEL WEIGHT (g)					
	2017	249.27 ± 61.40	113.72	416.40	
	2018	276.12 ± 56.51	138.50	440.70	0.91
	2019	260.18 ± 56.84	146.73	428.03	
BREAKING YIELD (%)					
	2017	44.86 ± 5.23	30.03	65.85	
	2018	45.45 ± 5.11	30.80	59.40	0.92
	2019	46.03 ± 5.09	30.47	60.54	

[a] Units abbreviations: mm for millimetre, % for percentage, g for gram and N for Newton

[b] SD is the abbreviation for standard deviation

[c] H^2 is the broad-sense heritability value

Then, the relationships between the different traits were evaluated using a Pearson correlation matrix between all the traits over the three years of phenotyping (Fig. 4). We can see that there is a significant positive correlation between the nut weight and the kernel weight (0.84), between the compression force on the suture and on the face (0.68) as well as between the extreme groups and the kernel and nut weight (0.72 and 0.78 respectively). These traits evolve together: when the nut weight increases, the kernel weight also increases. Likewise, the compressive force required to break the shell at the suture will increase if the compressive force necessary on the face increases.

Conversely, there is a significant negative correlation between the compression force exerted on the suture or on the face and the breaking yield (respectively -0.48 and -0.58): a low compression force would be linked to an important breaking yield.

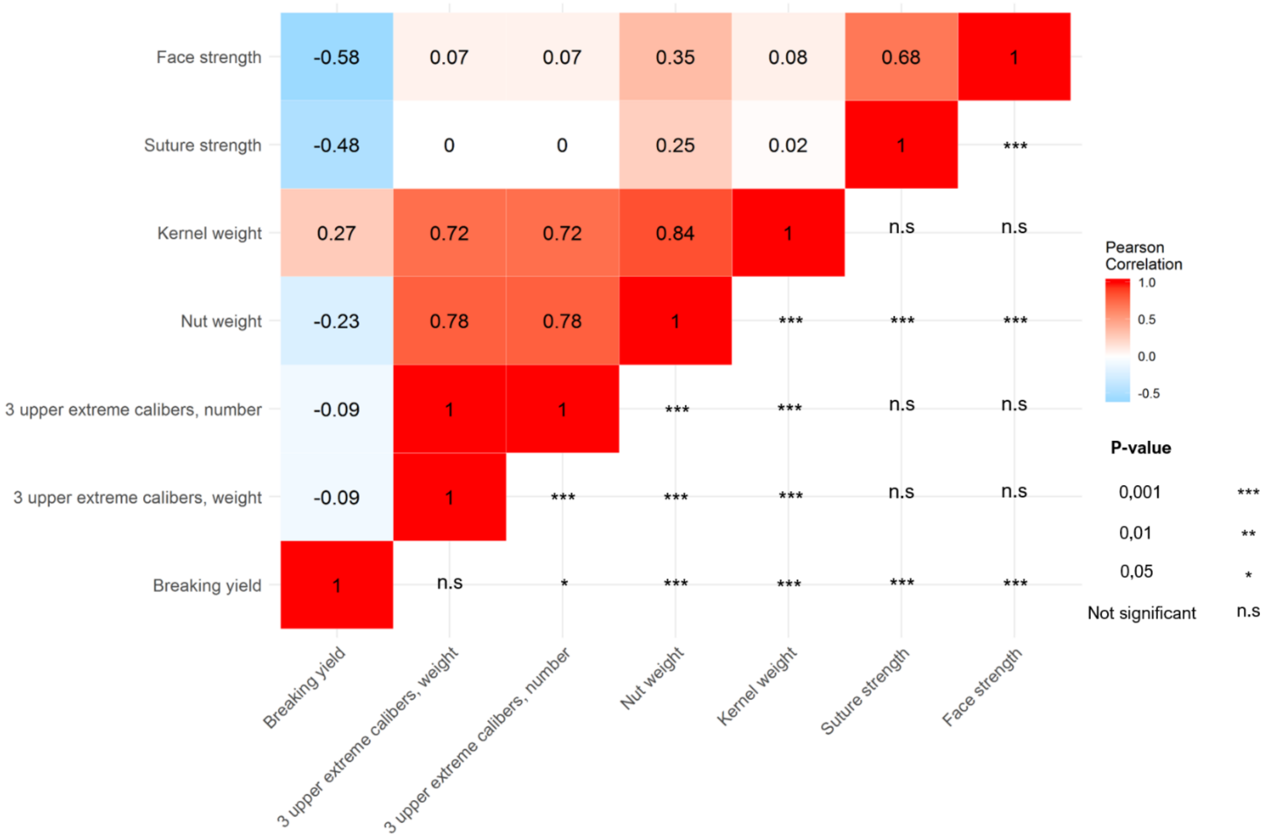


Fig. 4. Correlation matrix using three years data for all the traits

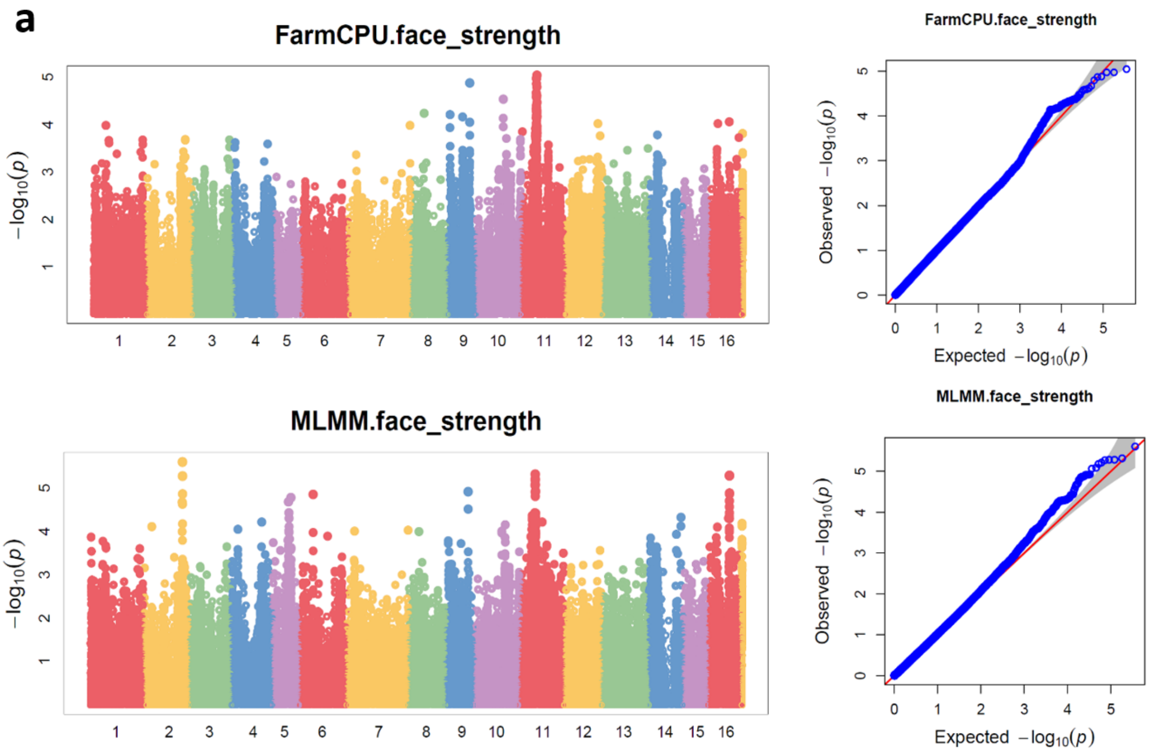
2. GWAS

The results analyzed from the GWAS are presented in the form of a Manhattan plot and a Q-Q plot (Fig. 5). For each character, we have two Manhattan plots, one for each model, that will be compared to determine the associations detected. Only six out of seven traits were used for GWAS; “3 upper extrem calibres” number and weight being similar and already having a trait related to nut weight, the first one was kept for analysis. Here, only two traits are shown for illustration, the other figures are available in the Supplemental data.

For the compression force on the face (Fig 5a), with both models, no signal exceeds the 1% Bonferroni threshold, however it is still possible to consider associations that are found in both models. Two signals stand out: one on chromosome 2 using MLMM with the SNP "AX-170865366" (position 32,096,896, p-value = 2.49E-6) and one on chromosome 11 using FarmCPU with the SNP "AX-170722428" (position 13,511,848, p -value = 9.10E-6). Interestingly, even with no significance, a signal also on chromosome 11 is found close using MLMM, with the SNP "AX-170721865" (position 13,163,274, p-value = 4.83E-6).

For the breaking yield (Fig 5b), with the FarmCPU model we obtained two significant associations: one at the end of the chromosome 14 and the second on chromosome 16. On the other hand, with the MLMM model there does not seem to be any association exceeding both significance thresholds of Bonferroni at 1% and 5%. However, one can find signals on chromosomes 14 and 16 like on the first model.

Q-Q plots are used to assess the relevance of fitting a data distribution to a theoretical model. They represent the observed p-values as a function of the expected p-values. When the blue curve follows the y = x line (red), the data is correctly adjusted with the model. This is the case for the compression force on the face, for both models and the breaking yield with the MLMM model. On the other hand, for the breaking yield with the FarmCPU model, the blue curve goes below the red line.



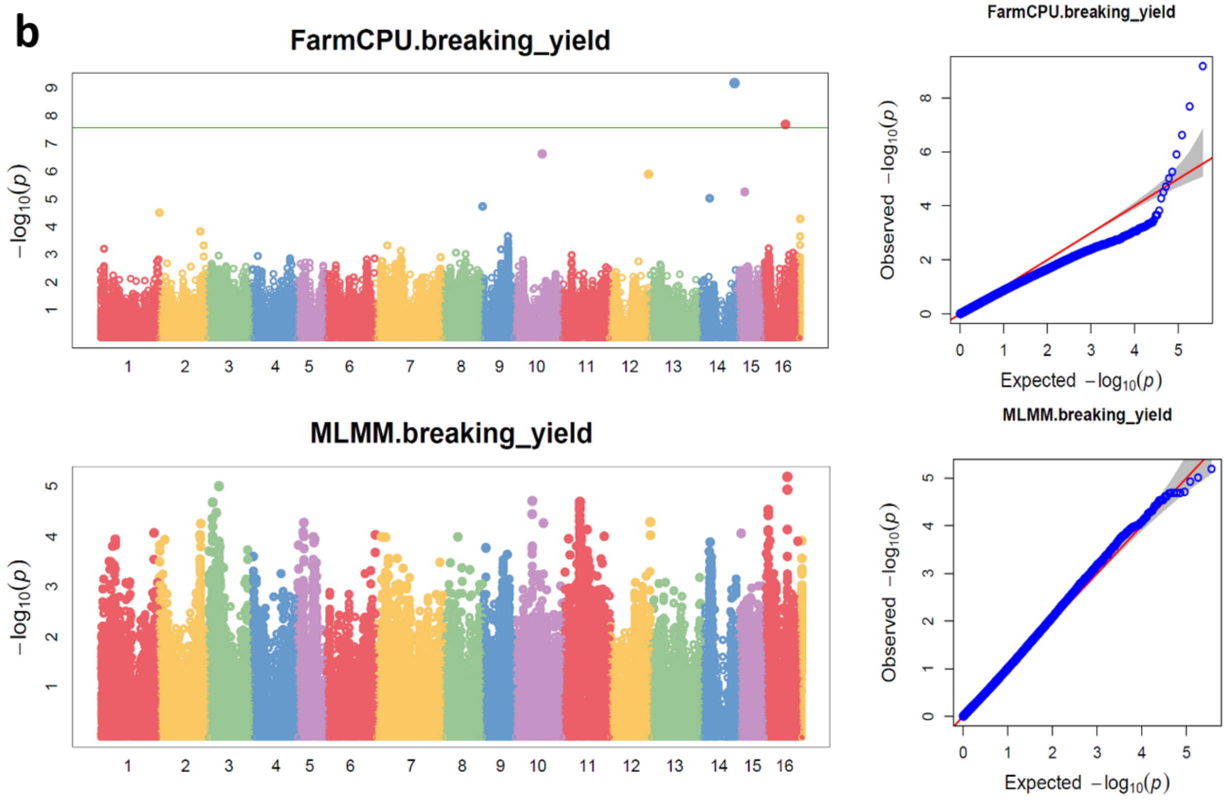


Fig 5. GWAS results for **a.** Face strength and **b.** Breaking yield traits, using FarmCPU and MLMM models

Considering both models, a total of 17 associations were kept for all the traits studied, distributed over almost all the walnut chromosomes. We obtained five SNPs associated for the nut weight, three SNPs associated for the extreme groups and for the kernel weight, and two for the compression force on the face as well as on the suture and for the breaking yield (Table 2). For nut weight, extreme groups, kernel weight and breaking yield, we focused on significant marker-trait associations that were obtained using FarmCPU model. However, for both compression forces, we decided to keep marker-trait associations having the higher p-value, knowing that we did not obtain significance.

For each association, the R^2 value was calculated, representing the percentage of phenotypic variation explained by this association. The traits studied are quantitative, therefore several genes can control them, and the associations detected are responsible for only a small percentage of the phenotypic variation measured. For instance, the SNP “AX-171175345” on chromosome 8 would explain 0.59% of the variation in nut weight. SNP “AX-170573680” on chromosome 3 would explain 2.10%.

However, two associations stand out particularly because of their R^2 value greater than 20%, on chromosomes 14 and 16, respectively for the extreme groups (21.70) and the breaking yield (27.23).

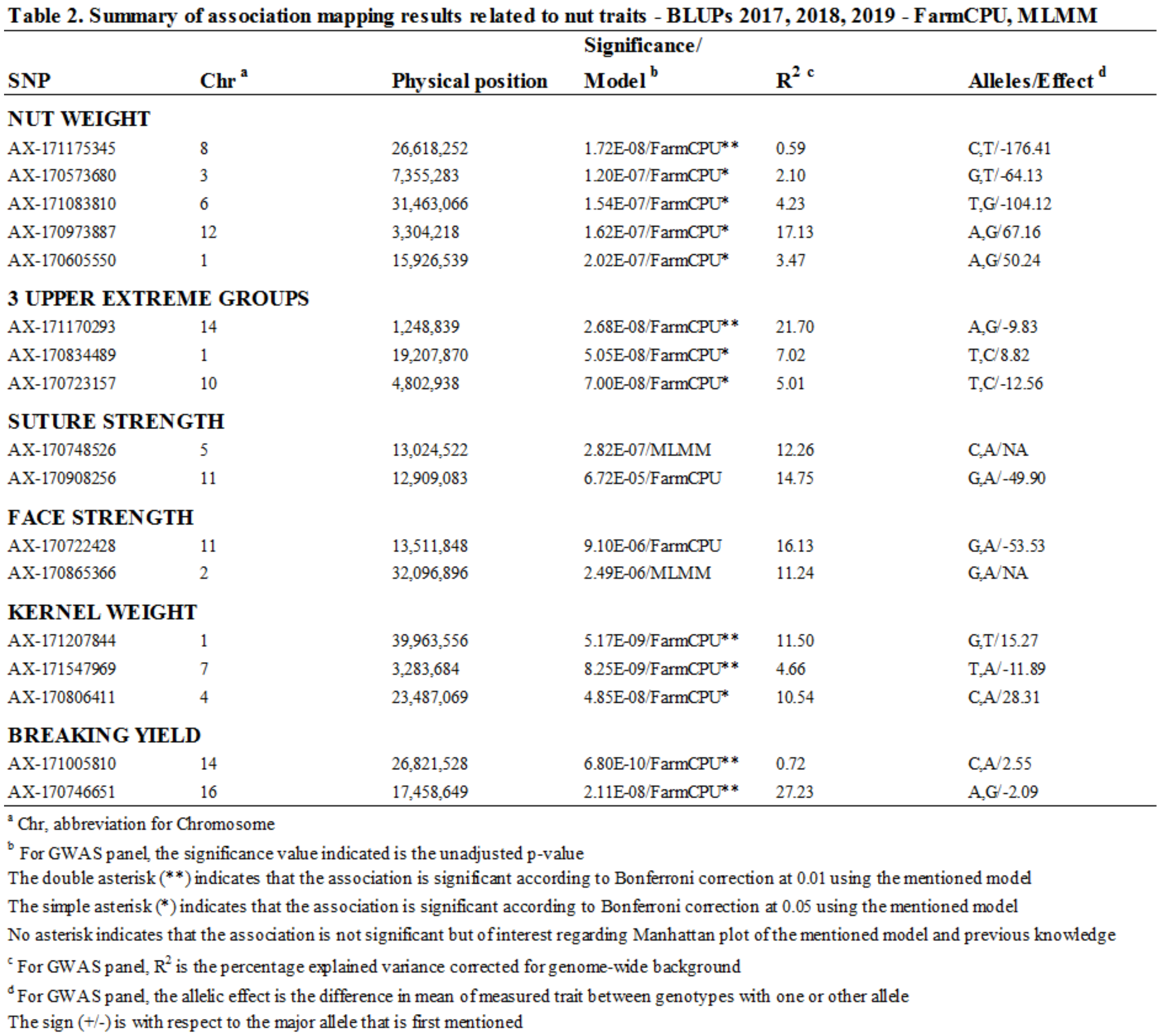


Table 2. Summary of association mapping results related to nut traits - BLUPs 2017, 2018, 2019 - FarmCPU, MLMM

SNP	Chr [a]	Physical position	Significance/ Model [b]	R^2 [c]	Alleles/Effect [d]
NUT WEIGHT					
AX-171175345	8	26,618,252	1.72E-08/FarmCPU**	0.59	C,T/-176.41
AX-170573680	3	7,355,283	1.20E-07/FarmCPU*	2.10	G,T/-64.13
AX-171083810	6	31,463,066	1.54E-07/FarmCPU*	4.23	T,G/-104.12
AX-170973887	12	3,304,218	1.62E-07/FarmCPU*	17.13	A,G/67.16
AX-170605550	1	15,926,539	2.02E-07/FarmCPU*	3.47	A,G/50.24
3 UPPER EXTREME GROUPS					
AX-171170293	14	1,248,839	2.68E-08/FarmCPU**	21.70	A,G/-9.83
AX-170834489	1	19,207,870	5.05E-08/FarmCPU*	7.02	T,C/8.82
AX-170723157	10	4,802,938	7.00E-08/FarmCPU*	5.01	T,C/-12.56
SUTURE STRENGTH					
AX-170748526	5	13,024,522	2.82E-07/MLMM	12.26	C,A/NA
AX-170908256	11	12,909,083	6.72E-05/FarmCPU	14.75	G,A/-49.90
FACE STRENGTH					
AX-170722428	11	13,511,848	9.10E-06/FarmCPU	16.13	G,A/-53.53
AX-170865366	2	32,096,896	2.49E-06/MLMM	11.24	G,A/NA
KERNEL WEIGHT					
AX-171207844	1	39,963,556	5.17E-09/FarmCPU**	11.50	G,T/15.27
AX-171547969	7	3,283,684	8.25E-09/FarmCPU**	4.66	T,A/-11.89
AX-170806411	4	23,487,069	4.85E-08/FarmCPU*	10.54	C,A/28.31
BREAKING YIELD					
AX-171005810	14	26,821,528	6.80E-10/FarmCPU**	0.72	C,A/2.55
AX-170746651	16	17,458,649	2.11E-08/FarmCPU**	27.23	A,G/-2.09

[a] Chr, abbreviation for Chromosome

[b] For GWAS panel, the significance value indicated is the unadjusted p-value
The double asterisk (**) indicates that the association is significant according to Bonferroni correction at 0.01 using the mentioned model
The simple asterisk (*) indicates that the association is significant according to Bonferroni correction at 0.05 using the mentioned model
No asterisk indicates that the association is not significant but of interest regarding Manhattan plot of the mentioned model and previous knowledge

[c] For GWAS panel, R^2 is the percentage explained variance corrected for genome-wide background

[d] For GWAS panel, the allelic effect is the difference in mean of measured trait between genotypes with one or other allele
The sign (+/-) is with respect to the major allele that is first mentioned

3. LD blocks and candidate genes

Once the GWAS analysis has been carried out and the SNPs associated with the studied traits found, we are interested in the LD blocks to which they belong using the Haploview software. This will subsequently enable us to find candidate genes potentially linked to these traits. These LD blocks were obtained using two analysis methods: Confidence Interval and Solid Spine of LD. However, the first approach being too strict and not giving results for all SNPs, the second one was used to be sure that a maximum of SNPs belong to an LD block for further research. Among all the positions detected, only the SNP "AX-170746651" located on chromosome 16, linked to the breaking yield, is in total linkage equilibrium and does not belong to any LD block and no gene was found in a window of 50 kb.

A total of 29 candidate genes were found associated with the markers detected (Table 3), four of which stand out particularly because of their location exactly at the position of the SNP marker. The

SNP "AX-170865366" on chromosome 2 linked to the compression force on the face is in the coding sequence for the TPX2-like protein.

Similarly, the SNP "AX-171547969" on chromosome 7, linked to the kernel weight, is included in the coding sequence of the gene for the BEL1-like homeodomain 4 protein.

There is also the SNP “AX-171170293” present on chromosome 14 and linked to the extreme groups, located in the coding sequence for the beta-galactosidase enzyme.

Finally, we found other molecules of interest linked to the traits studied, such as the lamin-like protein on chromosome 5 linked to the compression force on the suture.

Table 3. Summary of candidate genes within LD blocks defined by the "Solid spine of LD" method

Trait	SNP associated	Chr [a]	Physical position [b]	LD block interval [b]	Gene ID	Gene interval [b]	Functional annotation
3 upper extrem groups	AX-170834489	1	19,207,870	19,202,146 - 19,218,695	108991966	19,217,134 - 19,217,736	uncharacterized LOC108991966
3 upper extrem groups	AX-171170293	14	1,248,839	1,248,139 - 1,248,953	109012316	1,248,505 - 1,248,934	**beta-galactosidase**
Face strength	AX-170865366	2	32,096,896	32,095,993 - 32,104,213	109010833	32,096,186 - 32,100,753	**protein TPX2-like, transcript variants X1 to X3**
Face strength	AX-170722428	11	13,511,848	13,503,419 - 13,528,275	109000368	13,524,920 - 13,527,615	putative GPI-anchor transamidase
Kernel weight	AX-171207844	1	39,963,556	39,934,883 - 39,963,556	109003196	39,944,022 - 39,944,119	endochitinase 2-like
					109003484	39,947,326 - 39,948,386	uncharacterized mitochondrial protein AtMg00810-like, transcript variants X1 to X3
					109003196	39,948,866 - 39,949,391	endochitinase 2-like
					109003737	39,952,078 - 39,957,851	uncharacterized LOC109003737, transcript variants X1 and X2
					109006297	39,960,250 - 39,962,011	uncharacterized LOC109006297
Kernel weight	AX-171547969	7	3,283,684	3,279,632 - 3,290,086	108995343	3,280,546 - 3,286,458	**BEL1-like homeodomain protein 4**
Nut weight	AX-170605550	1	15,926,539	15,898,312 - 15,966,124	109011118	15,898,355 - 15,906,694	ER lumen protein-retaining receptor
					109006016	15,940,330 - 15,943,514	protein XRI1
					109011150	15,943,581 - 15,943,832	protein XRI1-like
					109011117	15,947,604 - 15,951,327	translation initiation factor IF-2-like
					109006017	15,953,917 - 15,961,294	purple acid phosphatase 15-like, transcript variants X1 to X3
					109006018	15,965,164 - 15,966,365	uncharacterized LOC109006018, transcript variants X1 and X2
Nut weight	AX-170573680	3	7,355,283	7,338,541 - 7,356,913	109003589	7,339,267 - 7,342,354	14 kDa zinc-binding protein
					109003588	7,343,193 - 7,344,784	sulfredoxin, chloroplastic/mitochondrial, transcript variants X1 and X2
					109003587	7,345,910 - 7,351,460	phosphoglycerate kinase, cytosolic
					109003586	7,354,174 - 7,354,840	chaperonin 60 subunit alpha 2, chloroplastic, transcript variants X1 and X2
Nut weight	AX-171083810	6	31,463,066	31,364,479 - 31,464,703	108985243	31,384,084 - 31,385,250	myb-related protein 308-like
					108985242	31,425,116 - 31,426,216	transcription repressor MYB6-like
Nut weight	AX-171175345	8	26,618,252	26,611,559 - 26,618,252	108985800	26,611,675 - 26,613,601	phosphoenolpyruvate carboxykinase [ATP]-like, transcript variants X1 to X5
					108985801	26,615,053 - 26,618,277	**uncharacterized LOC108985801, transcript variants X1 to X3**
Nut weight	AX-170973887	12	3,304,218	3,296,379 - 3,352,059	108998258	3,304,683 - 3,341,673	peroxisome biogenesis protein 1, transcript variants X1 and X2
					108998259	3,344,378 - 3,345,799	FHA domain-containing protein FHA2
Suture strength	AX-170748526	5	13,024,522	13,017,020 - 13,109,943	108986888	13,045,730 - 13,046,379	lamin-like protein
					108986895	13,046,658 - 13,056,810	protein argonaute 4A-like, transcript variants X1 to X4
					108986890	13,104,784 - 13,106,382	uncharacterized LOC108986890

[a] Chr, abbreviation for Chromosome

[b] Physical position given in bp

The candidate genes in **bold** overlap the physical position of the associated SNP

IV. Discussion and perspectives

1. Plant material

The plant material used during this internship comes from the INRAE collection of genetic resources of the genus *Juglans*. From this collection, which includes more than 400 individuals, a subset of 170 *J. regia* accessions was selected, representing the greatest diversity of the collection, to perform the GWAS analysis. However, of these 170 individuals, only 150 were used for the analysis because 20 trees with missing data due to lack of fruit production were discarded. The size and structure of the population used for a GWAS have an influence on the reliability of the results obtained: it is generally preferable to work with a large population to obtain statistically robust results. When working on plants, usually a panel of several hundred individuals is used (Korte and Farlow, 2013). Therefore, a population of 150 accessions may seem too small to obtain correct results. However, several studies have succeeded in detecting associations with a panel of less than 200 accessions, notably in apples (McClure et al. 2018) and peaches (Elsadr et al. 2019). Another GWAS analysis was also carried out on this same panel of 170 accessions, studying different traits, and gave several interesting results (Bernard et al. 2020).

2. Phenotyping data

The comparison of our phenotyping data over the three years shows consistency across the means and standard deviations for the seven traits studied. This is in agreement with the high heritability of the traits, an important characteristic for carrying out a GWAS analysis on a population. It means that the effect of the genome is more important than the effect of the environment.

After calculating this broad-sense heritability (H^2), large values were obtained with a lower H^2 for the compression forces on the suture and the face, suggesting less inheritable traits. It can also be observed in Table 2 that only two associations were detected for these two traits, compared to a maximum of five associations for the nut weight which also has a higher H^2 value. The results obtained in GWAS partly depend on the inheritance of a character. The more this character is inheritable, the easier it would be to analyze in GWAS.

The relationships between the seven measured traits were also analyzed using the correlation matrices. Thus, we were able to highlight a strong positive correlation between nuts weight, their size, and the kernel weight. Likewise, we have seen a correlation between the two compressive forces. A positive correlation between two traits could suggest that they are genetically linked. However, after our GWAS analysis, no position was detected in common between these traits.

3. Linear models

In a GWAS analysis, the choice of linear models to use for the study is an important parameter. There are many, ranging from the simple linear model to the most recent multi-locus mixed model, each with its advantages and disadvantages. For this study, two mixed multi-locus linear models were used: MLMM and FarmCPU. As described above, MLMM is a mixed model which will integrate the kinship matrix as well as several cofactors. It replaces simple linear models by reducing the probability of false positives (detecting an association that is not one) by considering the structure of the population. Its association detection power is more important than a simple model.

Meanwhile, FarmCPU is a multi-locus mixed model developed in 2015 and based on MLMM which it divides into two parts: one with fixed effect and another with random effect.

The comparison between these two models shows that FarmCPU seems more precise in detecting associations with a trait. It would also better control the appearance of false positives and negatives during the analysis (Kaler et al. 2020).

Applied to our data, by comparing the Q-Q plots (Fig 5) we notice that with the MLMM model the adjustment of the data follows more the line y = x than with FarmCPU on most characters. The MLMM model could seem the most suitable. With FarmCPU, the curve goes below this line y = x in certain cases, which means that the distribution is too adjusted compared to reality; too much information has been considered. However, this does not invalidate the results.

Most of the significant associations were detected with the FarmCPU model on the six traits studied, which suggests that the MLMM model may be too strict, or not so powerful. Fortunately, by comparing the Manhattan plots, both models give similar results. For instance, the significant signals on chromosomes 1 and 4 for kernel weight found using FarmCPU are also found remarkably close, but not significant, using MLMM. The same results can be observed for the association at the end of chromosome 16 for breaking yield. These findings tend to give weight to FarmCPU results that we focused on, even if the QQ-plots are not perfect.

4. Detected associations

GWAS analysis found 17 SNPs associated with the traits studied. Among them, three were found on chromosome 1, at positions 15,926,539 bp, 19,207,870 bp and 39,963,556 bp, respectively associated with nuts weight, extrem groups and kernel weight. Other major Quantitative Trait Locus (QTLs) have also been detected on this chromosome in previous studies, notably in the regions located at the start of the chromosome, around 6,000,000 bp and 9,000,000 bp, linked to walnut phenology (Marrano et al. 2019b; Bernard et al. 2020). We also detected two associations on chromosome 11, linked to the two compression forces, at 12,909,083 bp and 13,511,848 bp. It was in this region that another team found two other SNPs associated with the compressive force on the suture (Sideli et al.

2020). Another more distant region is associated with bearing habit, with a major SNP detected at position 20,831,267 bp responsible for 34.3% of the phenotypic variations observed (Bernard et al. 2020).

The R^2 values calculated for each association vary between 0.59% for the lowest (SNP "AX-171175345" in position 26,618,252 bp associated with nut weight) and 27.23% for the highest (SNP "AX-170746651" in position 17,458,649 bp associated with the breaking yield). Thus, we have two major SNPs on chromosomes 14 and 16, responsible for a large part of the phenotypic variation observed (respectively 21.70% and 27.23%) for nut size and breaking yield. There is also an association on chromosome 12 responsible for an important part of the variation observed in nut weight (17.13%) compared to the other SNPs detected.

5. Candidate genes function

Twenty-nine candidate genes were found associated with the markers detected during the GWAS analysis. Among them, four include the associated marker in their coding sequence: the TPX2-like protein, the BEL1-like homeodomain 4 protein, the β-galactosidase enzyme, and a protein unknown to date.

SNP "AX-170865366" is included in the coding sequence of the TPX2-like protein on chromosome 2 and linked to the compression force on the face. It is a Microtubule-Associated Protein (MAP) involved in the formation of the mitotic spindle (Boruc et al. 2019) and the regulation and organization of microtubules (Lei et al. 2019). It is therefore necessary for cell division but also for cell integrity, which makes it an ideal candidate for this trait. Related to the same trait, a GPI-anchor transamidase has been detected on chromosome 11. It is a protein involved in many metabolic and developmental processes such as plant growth and embryogenesis (Bundy et al. 2016).

Related to the kernel weight, we also found the BEL1-like homeodomain 4 protein which is part of the Three Amino-acid Loop Extension (TALE) class proteins. It interacts with Knotted-like homeobox proteins (KNOX) to form a heterodimer and regulate the transcription of genes involved in the development of the shoot apical meristem (Bhatt et al. 2004) and the ovum (Reiser et al. 1995).

Nut size and weight are two other complex traits involving several proteins and different metabolic pathways. In our case, we find the β-galactosidase enzyme linked to the size of the nut, which is responsible for the structure of cell walls in many fruits, such as peach and lemon, as well as their biogenesis (Wu and Burns 2004; Guo et al. 2018). Indeed, this enzyme can hydrolyse non-reducing terminal residues of β-D-galactosyl from polymers of β-D-galactoside (Wu and Burns, 2004 ; Guo et al., 2018). Nut weight is the trait with the most associations with markers, as well as a significant number of molecules, most of which are involved in the growth of the plant and its defences, such as purple acid phosphatase, zinc-binding. protein or phosphoglycerate kinase (Nishimura et al. 2013; Antonyuk et al. 2014; Rosa-Téllez et al. 2018).

Finally, we found three genes linked to the compressive force necessary on the suture to break the shell, one of which corresponds to the gene coding for a lamin-like protein. The team at University of Davis in California performed a similar GWAS analysis on the compressive force on the suture, using different techniques: manual force and texturometer (Sideli et al. 2020). Their results also highlight, with the FarmCPU model, an SNP "AX-170748528" on chromosome 5, at 13,023,760 bp, linked to a lamin-like protein. This protein appears to be involved in the formation of the nuclear lamina.

V. Conclusion

The creation of new varieties of nuts adapted to climate change and of quality requires better knowledge of the genetic origin of these criteria and the identification of new markers linked to them. The GWAS analysis carried out here made it possible to better understand this genetic determinism linked to the quality of the nut such as its weight, its size, and the strength of the shell, on a collection of INRAE walnut trees. This work provided tools for the implementation of a future selection, thanks to the phenotyping carried out on these accessions which bring new information on the characteristics of interest. New markers have also been detected as well as candidate genes for certain traits. Among these new candidate genes, we were able to find the coding sequence of a protein in common with the team at University of Davis in California, in connection with the compression force applied to the suture of the shell, a very promising result. Therefore, we have identified new SNPs associated with these quality traits and their position in the genome. This information will then be able to give rise to the development of markers usable in MAS, such as KASP markers. After their development, these markers must be tested on other trees phenotyped for the same traits, to be validated for use in MAS.

VI. References

Antonyuk SV, Olczak M, Olczak T, et al (2014) The structure of a purple acid phosphatase involved in plant growth and pathogen defence exhibits a novel immunoglobulin-like fold. IUCrJ 1:101–109. https://doi.org/10.1107/S205225251400400X

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265. https://doi.org/10.1093/bioinformatics/bth457

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting Linear Mixed-Effects Models Using {lme4}. Journal of Statistical Software 67:1–48. https://doi.org/10.18637/jss.v067.i01

Bernard A, Barreneche T, Lheureux F, Dirlewanger E (2018a) Analysis of genetic diversity and structure in a worldwide walnut (Juglans regia L.) germplasm using SSR markers. PLoS ONE 13:e0208021. https://doi.org/10.1371/journal.pone.0208021

Bernard A, Lheureux F, Dirlewanger E (2018b) Walnut: past and future of genetic improvement. Tree Genetics & Genomes 14:1. https://doi.org/10.1007/s11295-017-1214-0

Bernard A, Marrano A, Donkpegan A, et al (2020) Association and linkage mapping to unravel genetic architecture of phenological traits and lateral bearing in Persian walnut (Juglans regia L.). BMC Genomics 21:203. https://doi.org/10.1186/s12864-020-6616-y

Bhatt AM, Etchells JP, Canales C, et al (2004) VAAMANA—a BEL1-like homeodomain protein, interacts with KNOX proteins BP and STM and regulates inflorescence stem growth in Arabidopsis. Gene 328:103–111. https://doi.org/10.1016/j.gene.2003.12.033

Boruc J, Deng X, Mylle E, et al (2019) TPX2-LIKE PROTEIN3 Is the Primary Activator of α-Aurora Kinases and Is Essential for Embryogenesis. Plant Physiol 180:1389–1405. https://doi.org/10.1104/pp.18.01515

Bundy MGR, Kosentka PZ, Willet AH, et al (2016) A mutation in the catalytic subunit of the glycosylphosphatidylinositol transamidase disrupts growth, fertility and stomata formation in Arabidopsis. Plant Physiol pp.00339.2016. https://doi.org/10.1104/pp.16.00339

Elsadr H, Sherif S, Banks T, et al (2019) Refining the Genomic Region Containing a Major Locus Controlling Fruit Maturity in Peach. Sci Rep 9:7522. https://doi.org/10.1038/s41598-019-44042-4

FAO (2019) FAOSTAT. In: Food and Agriculture Organization of the United Nations. http://www.fao.org/faostat/fr/#data/QC. Accessed 26 Feb 2020

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol Evol 6:925–929. https://doi.org/10.1111/2041-210X.12382

Germain E, Prunet J-P, Garcin A (1999) Le noyer, monographie, Ctifl

Guo S, Song J, Zhang B, et al (2018) Genome-wide identification and expression analysis of beta-galactosidase family members during fruit softening of peach [Prunus persica (L.) Batsch]. Postharvest Biology and Technology 136:111–123. https://doi.org/10.1016/j.postharvbio.2017.10.005

Kaler AS, Gillman JD, Beissinger T, Purcell LC (2020) Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. Front Plant Sci 10:1794. https://doi.org/10.3389/fpls.2019.01794

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. https://doi.org/10.1186/1746-4811-9-29

Lei K, Liu A, Fan S, et al (2019) Identification of TPX2 Gene Family in Upland Cotton and Its Functional Analysis in Cotton Fiber Development. Genes 10:508. https://doi.org/10.3390/genes10070508

Lipka A, Tian F, Wang Q, et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu X, Huang M, Fan B, et al (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLoS Genet 12:e1005767. https://doi.org/10.1371/journal.pgen.1005767

Luedeling E, Steinmann KP, Zhang M, et al (2011) Climate change effects on walnut pests in California: CLIMATE CHANGE EFFECTS ON WALNUT PESTS. Global Change Biology 17:228–238. https://doi.org/10.1111/j.1365-2486.2010.02227.x

Manning WE (1978) The classification within the Juglandaceae. 1058–1087

Marrano A, Martínez-García PJ, Bianco L, et al (2019a) A new genomic tool for walnut (*Juglans regia* L.): development and validation of the high-density Axiom^{TM} *J. regia* 700K SNP genotyping array. Plant Biotechnol J 17:1027–1036. https://doi.org/10.1111/pbi.13034

Marrano A, Sideli GM, Leslie CA, et al (2019b) Deciphering of the Genetic Control of Phenology, Yield, and Pellicle Color in Persian Walnut (Juglans regia L.). Front Plant Sci 10:1140. https://doi.org/10.3389/fpls.2019.01140

Martínez-García PJ, Crepeau MW, Puiu D, et al (2016) The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J 87:507–532. https://doi.org/10.1111/tpj.13207

McClure KA, Gardner KM, Douglas GM, et al (2018) A Genome-Wide Association Study of Apple Quality and Scab Resistance. The Plant Genome 11:0. https://doi.org/10.3835/plantgenome2017.08.0075

Nishimura S, Tatano S, Miyamoto Y, et al (2013) A zinc-binding citrus protein metallothionein can act as a plant defense factor by controlling host-selective ACR-toxin production. Plant Mol Biol 11

Peterson BG, Carl P (2020) PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis

Reiser L, Modrusan Z, Margossian L, et al (1995) The BELL1 Gene Encodes a Homeodomain Protein Involved in Pattern Formation in the Arabidopsis Ovule Primordium

Rosa-Téllez S, Anoman AD, Flores-Tornero M, et al (2018) Phosphoglycerate Kinases Are Co-Regulated to Adjust Metabolism and to Optimize Growth. Plant Physiol 176:1182–1198. https://doi.org/10.1104/pp.17.01227

Segura V, Vilhjálmsson BJ, Platt A, et al (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830. https://doi.org/10.1038/ng.2314

Sideli GM, Marrano A, Montanari S, et al (2020) Quantitative phenotyping of shell suture strength in walnut (Juglans regia L.) enhances precision for detection of QTL and genome-wide association mapping. PLoS ONE 15:e0231144. https://doi.org/10.1371/journal.pone.0231144

Stevens KA, Woeste K, Chakraborty S, et al (2018) Genomic Variation Among and Within Six *Juglans* Species. G3 8:2153–2165. https://doi.org/10.1534/g3.118.200030

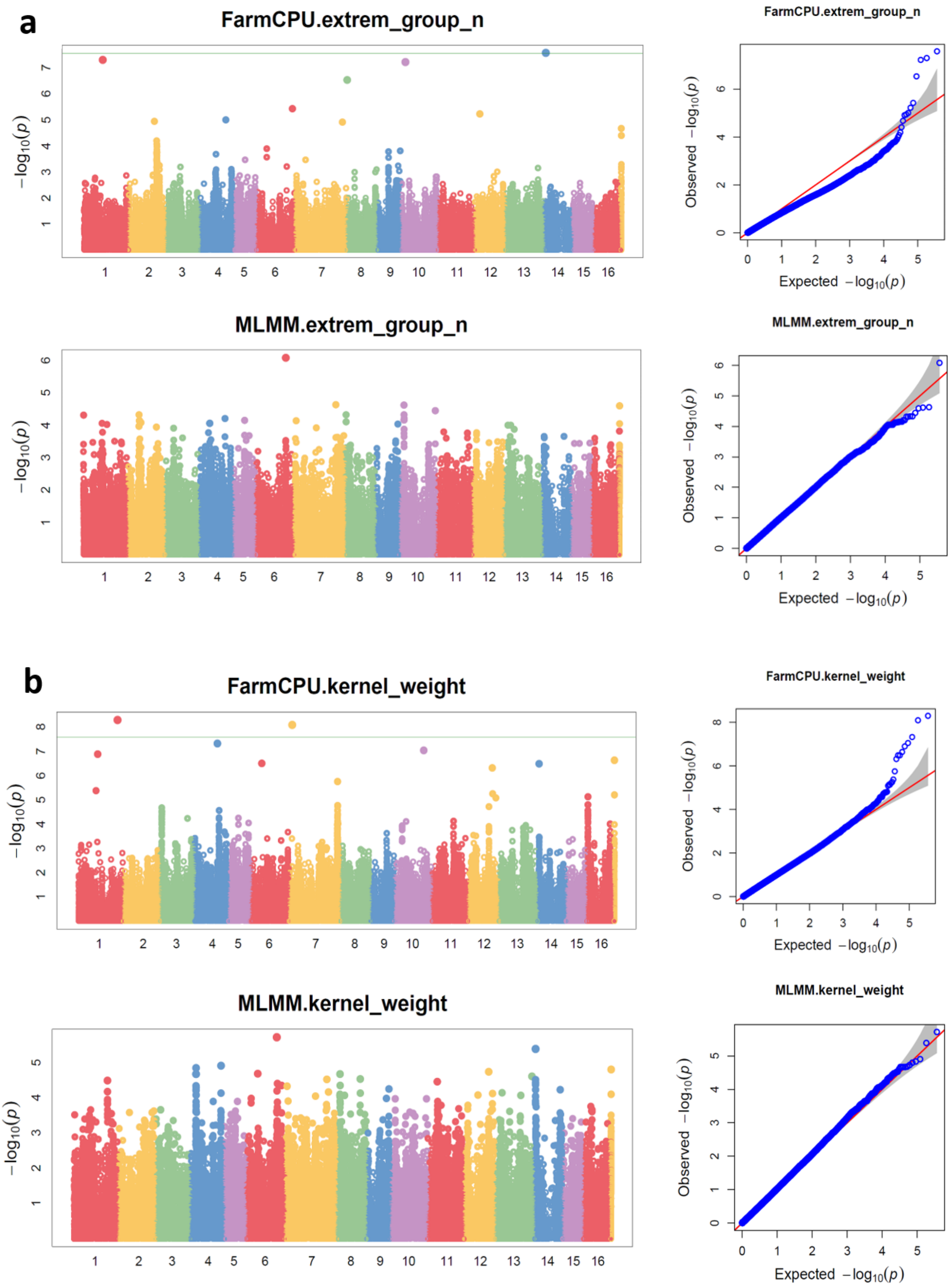
Wei T, Simko V (2017) R package “corrplot”: Visualization of a Correlation Matrix

Wickham H (2019) Tidyverse: Easily Install and Load the “Tidyverse.” Journal of Open Source Software 4:1686. https://doi.org/10.21105/joss.01686

Woodworth RH (1930) Meiosis of micro-sporogenesis in the Juglandaceae. 863–869

Wu Z, Burns JK (2004) A -galactosidase gene is expressed during mature fruit abscission of “Valencia” orange (Citrus sinensis). Journal of Experimental Botany 55:1483–1490. https://doi.org/10.1093/jxb/erh163

VII. Supplemental Data



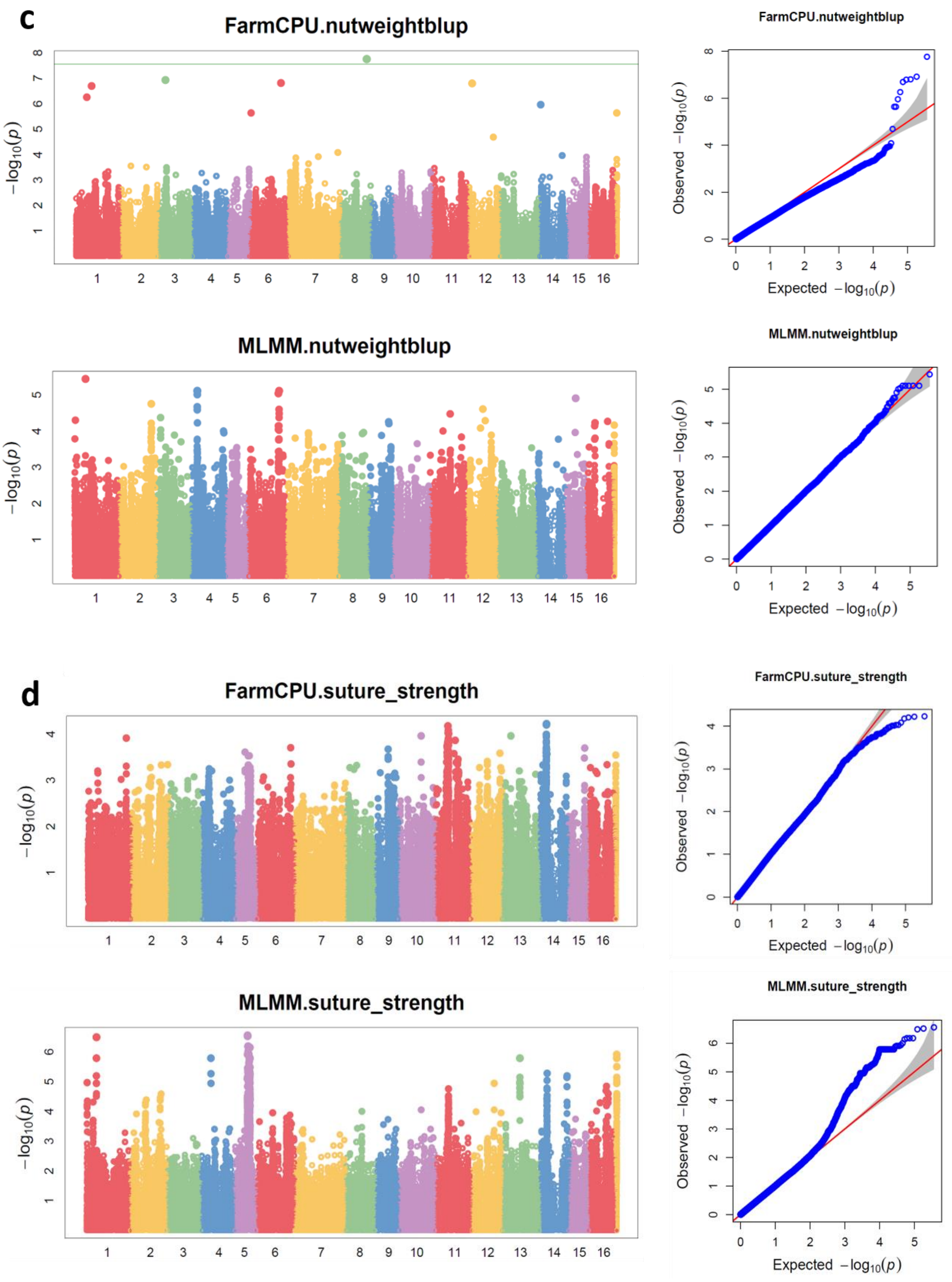


Fig S1. GWAS results for **a.** Extrem groups, **b.** Kernel weight, **c**. Nut weight and **d.** Suture strength traits, using FarmCPU and MLMM models

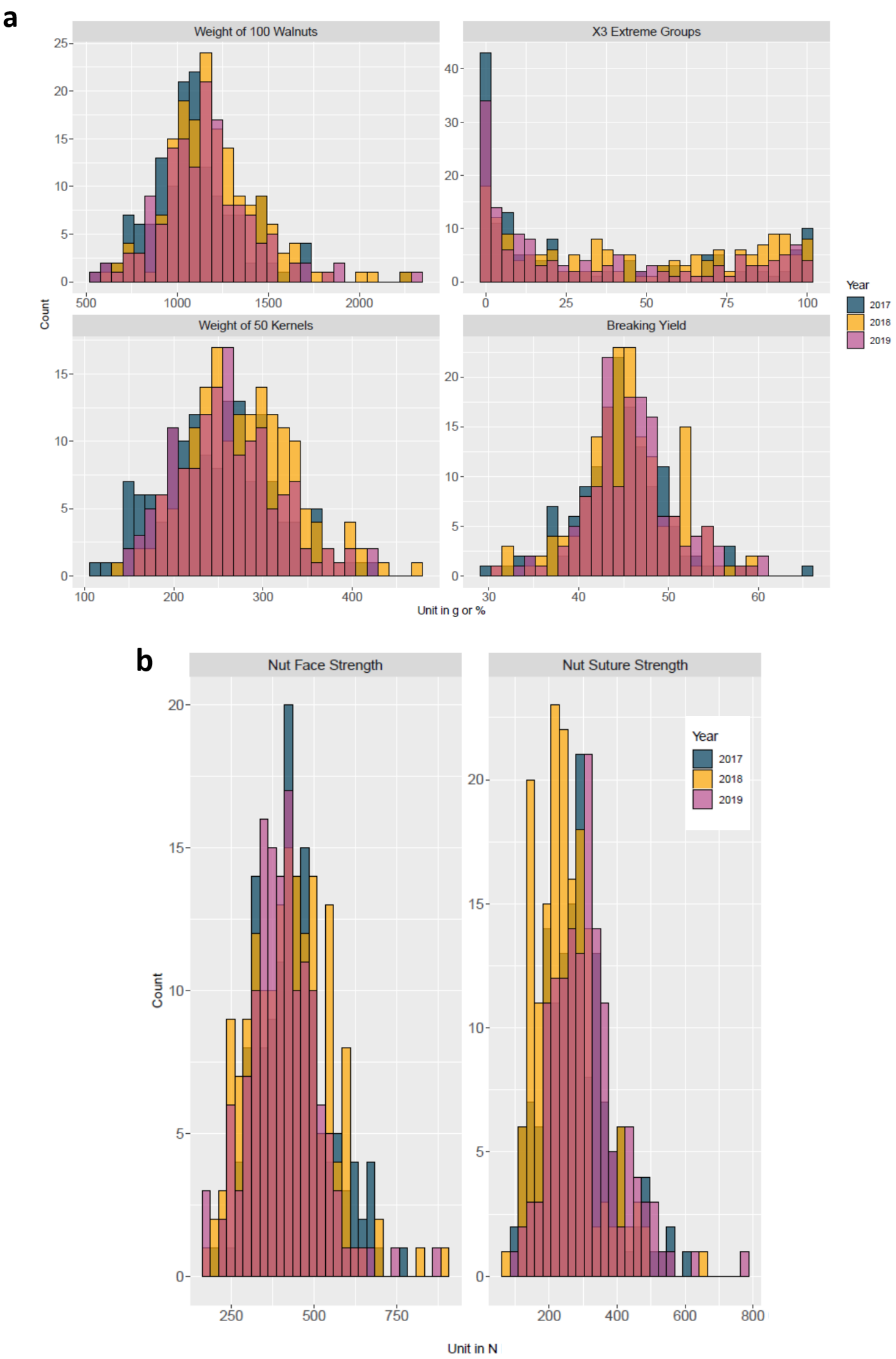


Fig S2. Distribution of phenotyping data over the three years for **a**. Nut weight, Extrem groups, kernel weight, Breaking yield and **b**. Face and Suture strength

Abstract

The “INNOV’noyer” program was set up by the CTIFL of Lanxade in partnership with INRAE Nouvelle-Aquitaine-Bordeaux and University of Davis, California. This project aims to study genetic diversity as well as phenotypic variability within the collection of genetic resources of INRAE, but also to identify the genetic determinism of traits of agronomic interest. My internship is part of this project for the characterization of the fruits in the collection for quality criteria and to identify SNP markers linked to them. The phenotyping of a third harvest year was carried out to complete the two previous years. The comparison of these data shows consistency over the three years, with heritability of traits and strong positive correlations between nuts weight, their size, and kernel weight. A positive correlation was also detected between the strength of the shell at the suture and at the face. A GWAS was carried out and we found 17 associations on the six traits studied, responsible for different percentages of the phenotypic variation observed. We found 29 candidate genes associated with the markers detected, with four including the SNP in their coding sequence.

Key words*: Walnut, GWAS, Association Genetics, Nut quality, Molecular markers*

Résumé

Le projet « INNOV’noyer » a été mis en place par le CTIFL de Lanxade en partenariat avec l'INRAE Nouvelle-Aquitaine-Bordeaux et l'Université de Davis en Californie. Ce projet vise à étudier la diversité génétique ainsi que la variabilité phénotypique au sein de la collection de ressources génétiques de l'INRAE, mais aussi à identifier le déterminisme génétique des caractères d'intérêt agronomique. Mon stage s'inscrit dans ce projet par la caractérisation des fruits de la collection pour des critères de qualité et l’identification de marqueurs SNP qui leur sont liés. Le phénotypage d’une troisième année de récolte a pu être réalisé pour compléter les deux années précédentes. La comparaison de ces données montre une cohérence sur les trois années, avec une héritabilité des traits et de fortes corrélations positives entre le poids des noix, leur taille et le poids du cerneau. Une corrélation positive a également été détectée entre la force nécessaire pour rompre la coque au niveau de la suture et au niveau de la face. La réalisation d’une GWAS a permis de trouver 17 associations sur les six traits étudiés, responsables à différents degrés de la variation phénotypique observée. Nous avons trouvé 29 gènes candidats associés aux marqueurs détectés, dont quatre intégrant le SNP dans leur séquence codante.

Mots-clefs *: Noyer, GWAS, Génétique d’association, Qualité* de la noix, *Marqueurs moléculaires*