Évaluation d'une stratégie de sélection génomique chez la tomate cultivée : prise en compte d'informations extérieures et des interactions GxE

Ange Fouabi

▶ **To cite this version:**

Ange Fouabi. Évaluation d'une stratégie de sélection génomique chez la tomate cultivée : prise en compte d'informations extérieures et des interactions GxE. Agronomie. 2020. dumas-03032941

HAL Id: dumas-03032941

https://dumas.ccsd.cnrs.fr/dumas-03032941v1

Submitted on 1 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Mémoire de fin d'études

Présenté pour l'obtention du Diplôme Ingénieur Agronome
Spécialité: Amélioration des Plantes et Ingénierie Végétale Méditerranéennes et Tropicales (APIMET)

Evaluation d'une stratégie de sélection génomique chez la tomate cultivée : prise en compte d'informations extérieures et des interactions GxE

Par **Ange FOUABI**

Année de soutenance : 2020

Organisme d'accueil
Institut National de la Recherche Agronomique, Avignon,
Unité GAFL (Génétique et Amélioration des Fruits et Légumes)

Mémoire de fin d'études

Présenté pour l'obtention du Diplôme Ingénieur Agronome
Spécialité: Amélioration des Plantes et Ingénierie Végétale Méditerranéennes et Tropicales (APIMET)

Evaluation d'une stratégie de sélection génomique chez la tomate cultivée : prise en compte d'informations extérieures et des interactions GxE

Par Ange FOUABI

Année de soutenance : 2020

Mémoire préparé sous la direction **de :**
Jacques DAVID

Présenté le : **17/09/2020**

Devant le jury :

Dominique THIS

Lydie GUILIONI

Vincent SEGURA

Organisme d'accueil **: INRA d'Avignon**

Maître de stage **: Mathilde CAUSSE**

REMERCIEMENTS

Je tiens tout d'abord à remercier mon maitre de stage Madame Mathilde CAUSSE pour m'avoir acceptée dans son équipe du GAFL et pour son encadrement tout au long du stage. Un grand merci pour m'avoir permis de travailler en autonomie et surtout d'avoir le goût de la recherche d'information. Aussi, je tiens à lui dire merci pour ses nombreux retours sur mon rapport. Elle m'a permis d'en apprendre d'avantage et d'approfondir mes connaissances acquises au cours de ma formation.

Je remercie aussi l'unité GAFL (Génétique et Amélioration des Fruits et Légumes) pour son accueil. Je remercie en particulier Frédérique Bitton pour sa disponibilité et pour son aide dans la mise en œuvre de mes scripts. Aussi, Morgane Roth pour sa disponibilité et ses conseils sur la sélection génomique. Enfin Emmanuel Le Calonnec pour sa disponibilité.

Merci à Vincent SEGURA pour sa disponibilité et ses conseils sur la sélection génomique.

Je remercie M. Jacques DAVID pour son encadrement et ses retours sur mon rapport de stage.

Un grand merci à toute l'équipe pédagogique de l'option APIMET pour son encadrement tout au long de la formation et pour m'avoir permis d'avoir des connaissances qui m'ont été utile dans l'exécution de ce stage. Mention spécial à Isabel MARTIN-GRANDE, Dominique THIS, Jacques DAVID et Pierre BERTHOMIEU.

Enfin, Mention spéciale pour mes camarades de promotion APIMET-SEPMET, pour cette belle année académique passée en leur compagnie.

TABLE DES MATIERES

SIGLES ET ACRONYMES

BLUP : *Best linear unbiased prediction* – Meilleur prédicteur linéaire non-biaisé

BGGE: *Bayesian Genomic Genotype x Environment Interaction*

BGLR: (*Bayesian Genomic Linear Regression*), prédire sans GxE

BL: *Bayesian LASSO*

DL : Déséquilibre de liaison

G-BLUP : Méthode BLUP utilisant la matrice d'apparentement G

GEBV : *Genomic estimated Breeding Value* – Estimation génomique de la Valeur génétique additive

GWAS : *Genome Wide Association Study*

Interaction G×E : Interaction entre le génotype et l'environnement

MAGIC: *Multi-parent Advanced Generation InterCross*

PE : Population d'entraînement

PredGen : Prédiction Génomique

PV : Population de validation

QTL : *Quantitative trait locus*

SAM : Sélection assistée par marqueurs

SelGen : Sélection génomique

SNP : Marqueur « *Single-nucleotide polymorphism* »

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

La tomate est l'un des légumes les plus consommés au monde. La superficie et la production de la tomate ont beaucoup augmenté (figure 1) ces 20 dernières années en raison de la demande de plus en plus croissante des consommateurs. Parallèlement, les critères de choix des consommateurs sont de plus en plus contraignants et précis. Pour répondre à ces critères, la recherche sur la tomate a connu un véritable essor avec la découverte de nombreux gènes responsables de la conservation et de la coloration du fruit et d'autres affectant la croissance de la plante. Les techniques de détection de QTLs (*Quantitative trait locus*) et de génétique d'association (GWAS : *Genome Wide Association Study*) ont permis la découverte de plusieurs gènes et QTLs intervenant dans la qualité du fruit. L'utilisation de ces résultats en sélection a été d'abord envisagée par sélection assistée par marqueurs (SAM). Cependant, la SAM présente des insuffisances en utilisant une part infime de la variation, et en se basant sur des effets souvent sur-estimés. La sélection génomique apparaît comme une alternative. Elle permet à l'aide de la prédiction génomique de faire un gain considérable de temps et de moyens et de combler les insuffisances des méthodes précédentes. La prédiction génomique est nouvelle dans la recherche sur la tomate. Des études précédentes réalisées dans l'équipe INRA de l'UR GAFL ont permis d'évaluer l'impact de certains paramètres de la prédiction génomique tels que le déséquilibre de liaison, la taille de la population d'entrainement, le nombre de marqueurs, etc. sur la précision de la prédiction génomique. Ce stage s'inscrit dans la continuité de ces projets en vue d'améliorer la prédiction génomique sur la tomate.

Notre étude porte sur l'évaluation d'une stratégie de prédiction génomique chez la tomate cultivée avec la prise en compte d'informations extérieures (cofacteurs environnementaux) et des interactions GxE. Ce stage s'inscrit dans le cadre du méta-programme de l'INRA SelGen qui vise à tester la sélection génomique chez plusieurs espèces pour améliorer les productions végétales vis-à-vis des facteurs biotiques et abiotiques. Il s'agit d'explorer le bénéfice de la sélection génomique pour accélérer le progrès génétique et permettre une amélioration de la qualité interne des fruits de tomate, caractère complexe et coûteux à mesurer, pour répondre à l'attente des consommateurs. L'objectif de ce stage est d'abord de tester l'impact des effets de stress environnementaux sur la qualité des prédictions, puis d'inclure dans les équations de prédiction l'effet de l'environnement l'interaction GxE ou des données de cofacteurs environnementaux, afin d'estimer si on accroît la précision des prédictions. Il est question d'évaluer les modèles de prédiction génomique qui ont permis de mieux connaître l'impact de plusieurs paramètres tels que le déséquilibre de liaison, la taille de la population d'entraînement et bien d'autres, évalués dans les études précédentes, en y ajoutant les effets de l'environnement et de l'interaction GxE.

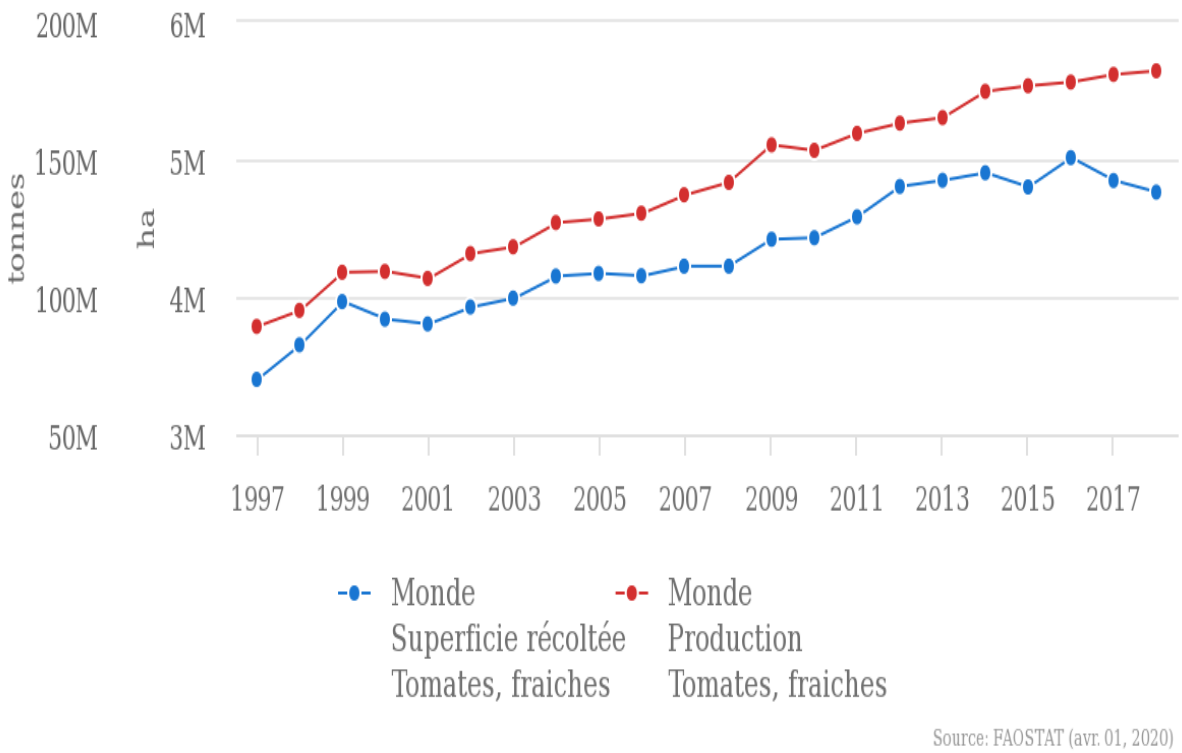


Figure 1 : Production et superficie dédiées à la culture de la tomate dans le monde de 1997 à 2018.

1. SYNTHESE BIBLIOGRAPHIQUE

1.1 Origine et importance de la tomate

La tomate, originaire d'Amérique du sud, plus précisément du nord-est de la Cordillère des Andes et du Pérou, était déjà cultivée par les Incas et par les Aztèques en Amérique centrale (Mattoo & Razdan, 2007). Elle a été domestiquée en deux temps, d'abord, à partir de l'espèce sauvage *Solanum pimpinellifolium*, au Perou où elle a donné les tomates de la taille d'une cerise *Solanum lycopersicum var. cerasiforme,* puis au Mexique où des variétés à gros fruits sont apparues (Blanca *et al.,* 2015). Ce sont les conquérants espagnols qui ont rapporté cette plante en Europe au XVI^{ème} siècle. Elle a d'abord été introduite en Italie et en Espagne comme plante ornementale. Le fruit de cette plante était considéré comme non comestible. C'est à partir du XVIII^{ème} siècle que les plants de tomates sont cultivés pour leurs fruits. On trouve encore actuellement à l'état sauvage, au Pérou et aux Antilles, la "tomate cerise" d'où dérivent probablement par améliorations culturales et hybridations successives, les nombreuses variétés utilisées actuellement.

La tomate cultivée est une espèce diploïde de 24 chromosomes, autogame avec de rares cas d'allogamie. L'espèce cultivée *Solanum lycopersicum* compte de nombreux mutants monogéniques naturels (Grassely *et al.,* 2000). Elle compte des dizaines de milliers de variétés conservées dans les banques de gènes, classées en différents types, *Solanum lycopersicum var. lycopersicum* à gros fruits et *Solanum lycopersicum var. cerasiforme* à petit fruit (tomate cerise) (Mattoo & Razdan, 2007).

La tomate est l'un des légumes les plus consommés dans le monde. La production mondiale de tomate a connu un essor particulier à partir des années 2000, et notamment en 2016, en battant un record de production jamais atteint auparavant avec 177 042 000 T produits (FAO, 2016). La Chine est le principal pays producteur de tomate avec une production moyenne en 2018 s'élevant à 38 637 970 T (voir Figure 2). L'Europe a produit 15,5 % de la production mondiale en 2018, avec à sa tête l'Italie suivie par l'Espagne qui approvisionnent d'autres pays d'Europe (FAO, 2018). La France a connu une baisse de sa production au cours des 20 dernières années (voir Figure 3) avec une réduction de sa surface au champ au profit de production sous serre et hors sol (Grassely *et al.*, 2000). Avec une production de 526 845 tonnes (Agreste, 2019), la France est le 7ème pays producteur de tomate en Europe.

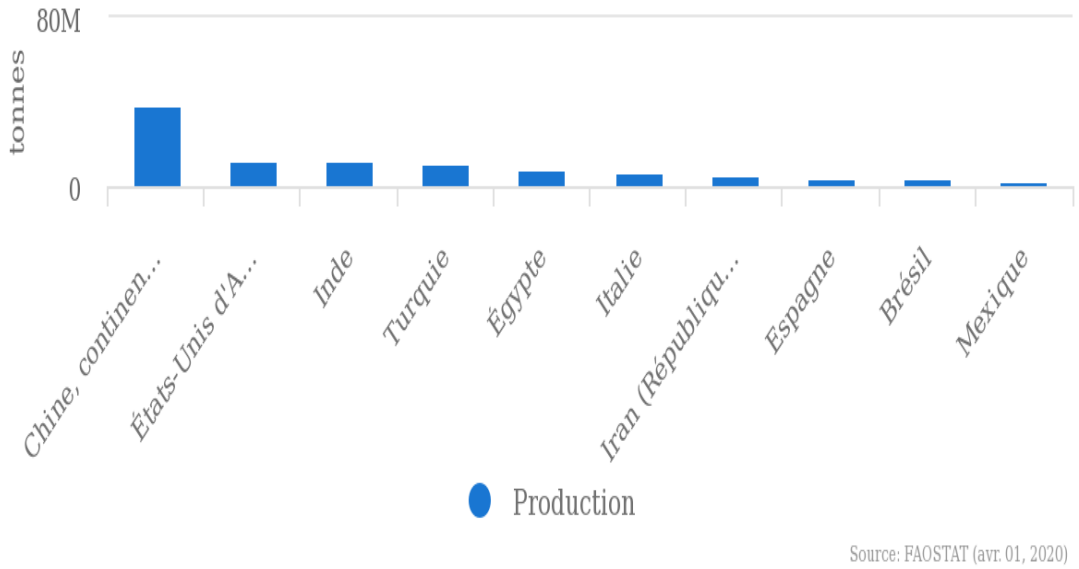


Figure 2 : Les principaux pays producteurs de tomate dans le monde

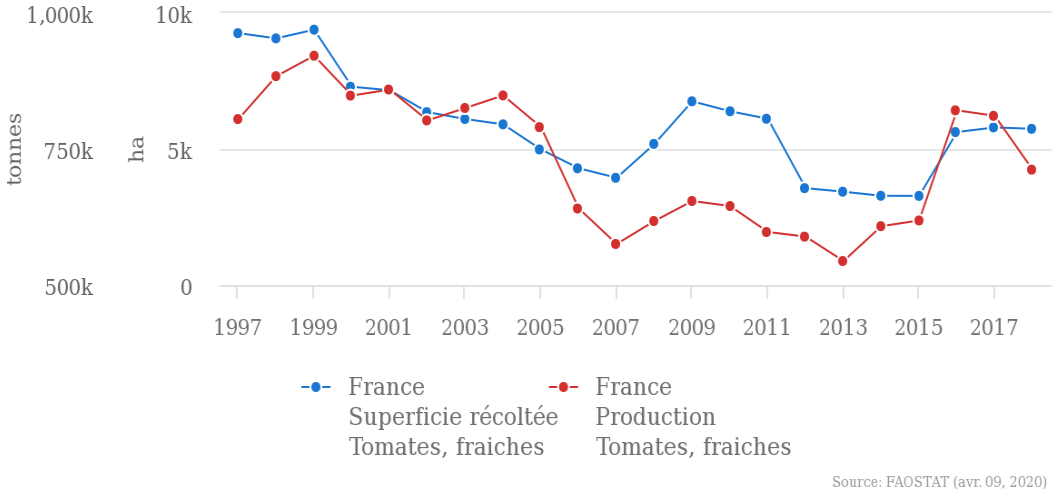


Figure 3 : Evolution de la production et superficie en tomate de la France de 1997 à 2018.

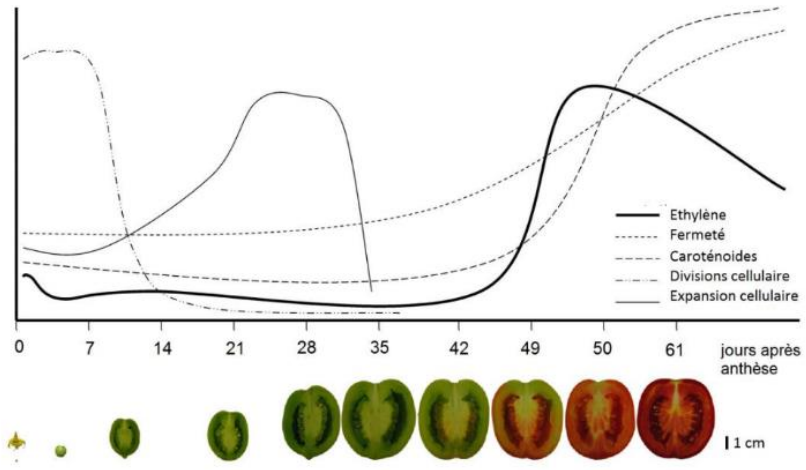


Figure 4 : Cycle de développement du fruit de la tomate (Giovannoni, 2004).

1.2 Qualité du fruit de la tomate

La tomate est consommée sous deux formes à savoir la tomate d'industrie et la tomate de frais, qui correspondent à des types variétaux très différents. Une grande proportion des tomates produites dans le monde est dédiée à l'industrie, avec une production très mécanisée uniquement en plein champ et une récolte unique. La tomate de frais est celle qu'on retrouve sur les étals ou dans les grandes surfaces qui est directement utilisée pour cuisiner. En France elle est majoritairement cultivée sous serre, avec une production qui peut durer jusque 11 mois, et une récolte continue à partir des premiers fruits qui apparaissent au bout de 90-120 jours (Figure 4). La plante est une liane qui produit environ un bouquet de fleurs et de fruits par semaine. Les consommateurs choisissent les tomates suivant différents critères qui sont liés à la manière dont ils les consomment (cuites ou crues). Les principaux critères de consommation, la fermeté, la fraicheur, le parfum en bouche, la jutosité ainsi que les critères d'achat, la fraicheur, l'aspect, la fermeté, le prix et la couleur, établis suite à une enquête réalisée par la CTIFL en 2000 en France, permettent de connaître les attentes des populations et d'y répondre au mieux (Grassely *et al.,* 2000). La qualité interne du fruit est définie par plusieurs composantes qui lui confèrent une qualité gustative qui répond aux besoins des populations. Elle est fonction de l'arôme (avec plusieurs dizaines de composés volatils responsables de l'arôme), des saveurs (acidité et sucrosité) et de la texture (fermeté, jutosité, farinosité etc.). Ces différentes composantes de la qualité du fruit sont à la fois influencées par l'expression de gènes et l'environnement.

1.3 Impact de l'environnement sur la qualité du fruit de la tomate

La qualité gustative de la tomate résulte de la combinaison ou de l'équilibre entre les différentes composantes de la qualité interne du fruit de la tomate. Ces composantes sont développées tout au long du cycle de développement du fruit. En effet, ce cycle se subdivise en 3 phases, à savoir la phase de division cellulaire qui dure 10 à 15 jours après la floraison, la phase de croissance rapide qui dure jusqu'au stade de fruit vert mature (30 à 35 jours) et la phase de maturation qui dure 2 semaines, au bout de laquelle le fruit devient consommable (Grassely *et al.,* 2000). La qualité du fruit est donc fonction du bon déroulement de ces phases. Elles sont toutes impactées par l'environnement, notamment par la température, la lumière, l'irrigation et la nutrition minérale (en éléments majeurs N, P, K et autres élément minéraux). La phase de floraison est fonction des temps thermiques, elle peut être précoce ou tardive suivant la température. De même la phase de fructification répond à la satisfaction des besoins de la plante en ce qui concerne les temps thermiques ou encore les besoins en eau et en nutriments de la plante. Ainsi, l'analyse de QTL détectés dans différents environnements montre que suivant les stress environnementaux (stress hydrique, température ou stress salin),

certaines composantes de la qualité sont plus modifiées que d'autres (Diouf *et al.*, 2018b).

1.4 Amélioration génétique de la tomate

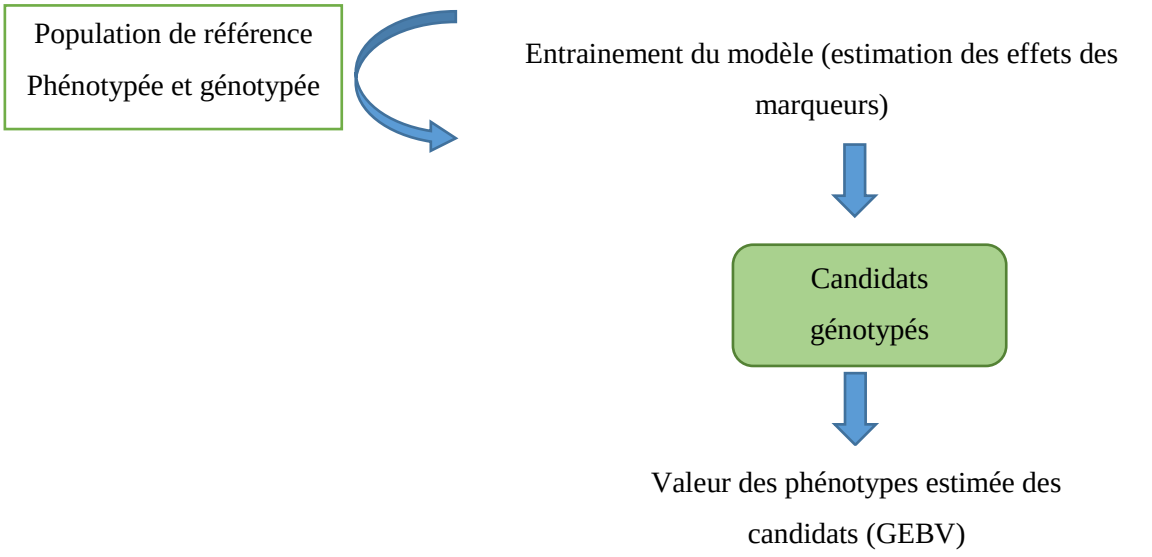
Le séquençage du génome de la tomate a été initié en 2005 par un consortium de 14 pays. Il a débuté par le génome d'une variété de tomate cultivée (Heinz 1706) qui a été séquencé et assemblé grâce à la technologie Sanger puis aux technologies de séquençage de nouvelle génération. De même, le génome de l'espèce sauvage *S. pimpinellifolium* a été séquencé et assemblé avec la technologie Illumina de séquençage de *short reads*. La comparaison de ces génomes (sauvage et cultivée) a montré qu'il existe une faible différence entre eux (0,6 % de SNP divergent) (Tomato Genome Consortium, 2012).

La tomate est une plante modèle, qui présente une forte synténie avec les autres Solanacées cultivées. En effet, son génome présente une forte similarité avec ceux des autres Solanacées (pomme de terre, piment et aubergine notamment) quand on les compare (Ranjan *et al.*, 2012a). Ce projet de séquençage a été suivi par plusieurs autres projets de séquençage de plus de variétés de tomate ou d'espèces apparentées. Le séquençage a permis de mieux connaître le génome de la tomate. En effet, l'identification plus précise des modifications qui interviennent lors de mutations naturelles a pu être menée avec succès. Différentes mutations naturelles sont survenues dans des gènes d'intérêt au cours de l'évolution de la tomate. Il s'agit notamment de mutations liées à la croissance de la plante (sp : self pruning) qui entraîne une croissance déterminée de la plante ; les mutations liées à la coloration du fruit (r : yellow flesh, t : tangerine et de nombreuses autres) et les mutations de maturation des fruits telles que rin (ripening inhibitor), nor (non ripening) et alc (alcobaca) qui bloquent ou ralentissent le murissement du fruit (Grassely *et al.*, 2000). Le séquençage a facilité l'identification des marqueurs moléculaires polymorphes de type SNP. Les marqueurs rendent possible l'identification des gènes et régions du génome qui sont responsables des phénotypes observés. De nombreux gènes d'intérêt ont été clonés (notamment ceux cités plus haut, mais aussi plus de 20 gènes de résistance à des pathogènes), et un certain nombre de QTL à effets majeurs. Un très grand nombre d'études de QTLs (*Quantitative Trait Loci*) responsables de la variation de différents traits ont été réalisées (Mattoo & Razdan, 2007), en particulier en réponse à des stress abiotiques. Par exemple, 54 QTLs ont été détectés dans le cadre de l'étude de (Diouf *et al.*, 2018b), dont certains sont responsables de la croissance de la plante, de la qualité du fruit tandis que d'autres résultent de l'interaction de la plante avec son environnement. La précision de cartographie des QTL peut être améliorée par le type de population utilisée. En effet, l'utilisation d'une population MAGIC, population issue de croisement mixte entre huit parents, a permis la détection plus facile des SNP causaux ou liés aux QTLs et la réduction du nombre des gènes et de SNP candidats (Pascual *et al.*, 2015b). Les marqueurs de gènes et de QTL sont utilisés en sélection assistée par marqueurs (Mattoo

& Razdan, 2007). Dans l'optique de rendre la sélection plus précise, plusieurs technologies ont vu le jour, il s'agit notamment des technologies d'édition du génome. En effet, la technologie d'édition du génome CRISPR/cas9 (*clustered regularly interspaced short palindromic repeats*), permettrait une meilleure précision et un gain de temps dans le processus de sélection de la tomate (Rothan *et al.*, 2019a). Cependant, la réglementation européenne constitue un frein à sa pleine utilisation par les sélectionneurs de la tomate.

1.5 Sélection génomique

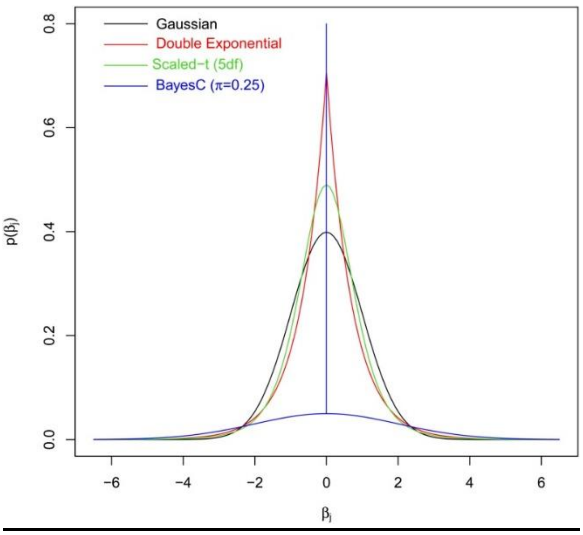
La dissection des bases génétiques de caractères d'intérêt, notamment la qualité du fruit, a été faite durant de nombreuses années avec la détection de QTLs puis la GWAS (*Genome Wide Association Study*) basées sur les informations des marqueurs (Duangjit *et al.*, 2016a). Cependant, les approches de QTL et GWAS ne permettent pas de détecter les effets des allèles mineurs, en raison des seuils stricts utilisés pour éviter la détection des faux positifs (*Finding the missing heritability of complex diseases*, s. d.) . L'utilisation des marqueurs de QTL en SAM ne permet donc de n'utiliser qu'une part réduite de la variation génétique et les effets des QTL sont en général surestimés. Contrairement à la SAM, la sélection génomique (SelGen) utilise les marqueurs de l'ensemble du génome pour la prédiction génomique de la valeur des individus à sélectionner. La sélection génomique est un outil de sélection d'animaux et de plantes visant à prédire les performances des individus pour des caractères d'intérêt suivant des modèles statistiques (Duangjit *et al.*, 2016a). Elle utilise les données moléculaires et phénotypiques d'une population d'entraînement pour obtenir la GEBV (*genomic estimated breeding value*) des individus de la population test qui a été génotypée mais n'a pas été phénotypée (Crossa *et al.*, 2017).



La sélection génomique comprend les différentes phases dont les croisements, les génotypages et phénotypages, la prédiction génomique, etc. La prédiction génomique (PredGen) est couramment utilisée dans l'amélioration génétique de plusieurs espèces sans faire intervenir les autres étapes de la sélection génomique. Elle vise à prédire les phénotypes d'individus génotypés mais dont le phénotype est inconnu. La PredGen a été introduite pour améliorer la précision de la sélection et en réduire les coûts et accélérer le progrès génétique (Lan *et al.,* 2020). La sélection génomique est influencée par deux types de facteurs. Il s'agit des facteurs intrinsèques à la population analysée (déséquilibre de liaison, diversité génétique, structure de la population et héritabilité des caractères) qui ne peuvent pas être modifiés par le sélectionneur tandis que les facteurs modifiables (la taille de la population d'entraînement, le nombre de marqueurs, les modèles statistiques, etc.) peuvent être modifiés par le sélectionneur. Les deux principaux facteurs qui influencent la précision de la prédiction de la SelGen, sont le déséquilibre de liaison et la taille de la population d'entraînement. En effet, il est important de connaître l'étendue du déséquilibre de liaison au niveau des différentes populations pour mieux estimer la densité de marqueurs nécessaire pour mettre en place la PredGen (Denis, 2016). A ces facteurs s'ajoutent la diversité de la population d'entraînement et l'apparentement entre les deux populations (d'entraînement et de validation) (Lan *et al.,* 2020), le modèle statistique et l'héritabilité des caractères sélectionnés. En effet, pour un caractère complexe, à faible héritabilité, de nombreux marqueurs à faibles effets sont nécessaires pour faire la PredGen, tandis que pour un caractère moins complexe à forte héritabilité quelques marqueurs sont appropriés (Crossa *et al.,* 2017). Plusieurs modèles statistiques (Tableau 1, figure 5) ont été proposés et certains d'entre eux sont intégrés dans des packages du logiciel R. Il s'agit des packages Synbreed, rrBLUP et BLR dans lesquels les modèles P-BLUP, G-BLUP, RR-BLUP, Bayesian ridge regression, Bayesian Lasso regression, BGLR, BWGS et BGGE (Annexe 1) sont implémentés (Endelman, 2011; Legarra *et al.,* 2014; Pérez & de los Campos, 2014a; Bastien *et al.,* 2016; Duangjit *et al.,* 2016a; Granato *et al.,* 2018a; Charmet *et al.,* 2019).

Tableau 1 : Exemple de modèles statistiques utilisé en SelGen.

Espèces	Modèles	Sources
Tomate Et Maïs	$y = 1_n\mu + X\beta + e$ y : variable à prédire ; X : matrice n × m des génotypes SNP, dont les entrées codent 0, 1 ou 2 copies des allèles de référence; β : vecteur m × 1 des effets des SNP ; e : effet résiduel.	(Duangjit *et al.*, 2016a) (Millet *et al.*, 2019a)
Manioc	$y = X\beta + Zu + e$ y : vecteur de réponse des traits ; X : matrice des effets fixes β ; Z : matrice des effets génétiques aléatoires u ; e : effet résiduel.	(Okeke *et al.*, 2017)



Model (prior density)	Hyperparameters	Treatment in BGLR[a]
Flat (FIXED)	Mean (μ_β)	$\mu_\beta = 0$
	Variance (σ^2_β)	$\sigma^2_\beta = 1 \times 1^{10}$
Gaussian (BRR)	Mean (μ_β)	$\mu_\beta = 0$
	Variance (σ^2_β)	$\sigma^2_\beta \sim \chi^{-2}$
Scaled-t (BayesA)	Degrees of freedom (df_β)	User specified (default value, 5)
	Scale (S_β)	$S_\beta \sim Gamma$
Double exponential (BL)	λ^2	λ fixed, user specified, or $\lambda^2 \sim Gamma$, or $\lambda/\max \sim Beta^b$
Gaussian mixture (BayesC)	π (prop. of nonnull effects)	$\pi \sim Beta$
	df_β	User specified (default value, 5)
	S_β	$S_\beta \sim Gamma$
Scaled-t mixture (BayesB)	π (prop. of nonnull effects)	$\pi \sim Beta$
	df_β	User specified (default value, 5)
	S_β	$S_\beta \sim Gamma$

Figure 5 : Densités des coefficients de régression des différents modèles mis en œuvre dans le package BGLR

Les populations biparentales et multi-parentales sont utilisées pour la SelGen, auxquelles s'ajoute l'utilisation fréquente de collections d'accessions non ou peu apparentées à large base génétique. La validation croisée (*cross-validation*) a montré son efficacité quant à la mise en œuvre de la SelGen. La validation croisée requière 4 étapes principales : la première est l'assignation des individus à la population d'entraînement (PE) et de validation (PV); la deuxième est l'estimation des effets des marqueurs basés sur les génotypes et les phénotypes dans la population d'entraînement ; la troisième consiste à calculer la GEBV en utilisant les informations des génotypes de la population de validation ajustée dans un modèle de prédiction et la dernière étape consiste à estimer la précision (*accuracy*)

de la prédiction à partir de la corrélation entre la *breeding value* prédite et les données phénotypique de la population de validation (Duangjit *et al.,* 2016). Cette méthode a été utilisée pour la sélection génomique de plusieurs espèces animales et végétales (voir Tableau 2). La SelGen a tout d'abord été utilisée dans la filière bovine laitière où elle a apporté un plus dans la mesure où elle permet un progrès génétique plus élevé, plus équilibré et une meilleure précision des index femelles (Bastien *et al.,* 2016, Croiseau et Boichard, 2016). Certaines études essaient d'unifier les processus de sélection génomique chez les plantes et chez les animaux en vue d'améliorer le gain génétique (Hickey *et al.,* 2017). La sélection génomique est aujourd'hui appliquée à plusieurs espèces végétales. Elle permet une amélioration génétique plus rapide pour des espèces à cycle long comme le manioc (Okeke *et al.,* 2017) ou les espèces pérennes (Cros *et al.,* 2015a). Plusieurs améliorations ont été recherchées. La comparaison de modèles de SelGen a montré qu'un modèle multi-traits a une meilleure précision de prédiction (corrélation entre les phénotypes observés et prédits) qu'un modèle avec un seul trait quand les deux sont corrélés et évalués dans un même environnement, tandis qu'un modèle multi-environnement a une précision de prédiction meilleure que celle d'un modèle uni-environnement (Okeke *et al.,* 2017). Les modèles multivariés se sont révélés plus appropriés pour l'amélioration génétique du manioc. Une étude sur le lin montre que l'utilisation des marqueurs de QTLs en lieu et place des SNP améliore la précision de la prédiction de la SelGen. Cependant, l'élimination des QTLs redondants et des faux positifs maximise la précision de la prédiction (Lan *et al.,* 2020). La précision de la prédiction dépend aussi de l'interaction génotype x environnement. En effet, l'étude de Millet *et al.,* (2019) sur le maïs a montré que la prise en compte des variations de l'environnement et des interactions génotype x environnement (GxE) améliore la précision de la prédiction génomique.

La tomate a fait aussi l'objet de recherches sur la sélection génomique. La prédiction génomique a permis de simuler l'amélioration génétique du rendement et de la saveur de la tomate. On peut ainsi mettre en place des modèles de prédiction du phénotype à partir des données génotypiques et phénotypiques (Yamamoto *et al.,* 2016). Elle s'applique aussi dans la sélection de résistances à des agents pathogènes de la tomate. En effet, la comparaison de la sélection basée sur GEBV et la sélection phénotypique, dans la recherche de résistance à une bactérie, a montré que la précision de modèles avec GEBV est supérieure à celle de la sélection phénotypique (Liabeuf *et al.,* 2017). La précision de la prédiction génomique est dépendante des facteurs de la sélection génomique. En effet, Duangjit *et al.,* (2016a) a montré dans son étude sur la qualité des fruits de tomate, que l'héritabilité et la précision de la prédiction sont positivement corrélées. De même que pour la majeure partie des traits, l'augmentation du nombre de marqueurs permet une meilleure précision de la prédiction comme la taille de la population d'entraînement. L'étude menée par Picard, (2015) fait partie d'un ensemble d'études menées par l'unité GAFL (Génétique et Amélioration des

Fruits et Légumes) dans l'optique d'amélioration génétique de la tomate. Cette étude menée sur les populations de GWAS (utilisée par Pascual *et al.,* 2015), MAGIC et RILs, met en évidence l'avantage de la sélection génomique et évalue les impacts de plusieurs facteurs sur sa précision. Elle montre que cette précision est d'autant plus élevée que la population d'entraînement est grande et contient des individus génétiquement proches et phénotypés dans les mêmes environnements. De même, les caractères de la qualité du fruit ayant une forte héritabilité ont une meilleure précision de prédiction. Après évaluation de l'impact des facteurs de la sélection génomique sur la précision, il convient d'évaluer l'effet de l'interaction génotype x environnement sur la précision de la prédiction génomique des traits de qualité du fruit de la tomate.

Tableau 2: Exemples de quelques espèces végétales sur lesquelles la sélection génomique est appliquée.

Espèces	intitulé	Sources
Avoine	Précision et formation de la population d'entraînement pour la sélection génomique sur les caractères quantitatifs de l'avoine d'élite nord-américaine	(Asoro *et al.,* 2011)
Maïs	Une fondation pour la biofortification de la provitamine A de Maïs : Modèles d'association à l'échelle du génome et de prédiction génomique des niveaux de caroténoïdes	(Owens *et al.,* 2014)
Blé	Précision de la sélection génomique à l'aide de données historiques générées dans le cadre d'un programme de sélection du blé en France	(Storlie & Charmet, 2013)
Canne à sucre	Évaluation expérimentale de la précision de la sélection génomique de la canne à sucre	(Gouy *et al.,* 2013)
Palmier à huile	Précision de la prédiction de la sélection génomique dans une culture pérenne : étude de cas du palmier à huile (Elaeis guineensis Jacq.)	(Cros *et al.,* 2015a)

1.6 Prise en compte de l'interaction Génotype x Environnement (GxE) et modèles multi-traits dans la Sélection Génomique

La Sélection génomique appliquée aux plantes implique la prise en compte de plusieurs facteurs liés à la plante et à son environnement. Il convient de prendre en compte ces différents facteurs dans les modèles statistiques de prédiction, pour garantir une bonne précision de la prédiction génomique. Ces facteurs sont les génotypes (apparentement ou génotypes aux marqueurs), les différents

environnements d'essai, les interactions potentielles entre génotypes, les interactions génotypes x environnement et les effets non additifs (Oakey *et al.,* 2016). Deux types de prédiction peuvent être envisagés : (1) de nouvelles lignées dans un nouvel environnement ou (2) de nouvelles lignées testées seulement dans certains environnements. Oakey *et al.,* (2016) étudie la prise en compte de la variation spatiale à travers les composantes environnementales, les interactions entre marqueurs et les interactions entre marqueurs et environnement par l'utilisation de *ridge regression* BLUP dans un modèle linéaire mixte. Ce modèle basé sur l'analyse multi-environnement (voir Tableau 3) permet une amélioration des prédictions génomiques (amélioration de la capacité de prédiction de 11,4 %) par rapport à celles du modèle basé sur un seul environnement. Cette augmentation a été observée pour les essais dont les héritabilités sont élevées (supérieure à 0,70). L'efficacité dépend des corrélations entre environnements. L'utilisation des données d'essais multi-environnementaux et de modèles linéaires mixtes en comparaison avec des modèles mixtes à environnement unique, pour la sélection génomique avec une structure d'analyse factorielle, augmente le pouvoir prédictif d'environ 6%. Seul le modèle à 3 facteurs analytiques qui considère les sites comme des effets fixes et qui modélise l'interaction GxE permet cette augmentation (Burgueño *et al.,* 2011). Dans son étude sur le manioc, Okeke *et al.,* (2017) fait la comparaison de modèles mixtes à un trait et multiple traits dans un environnement, à ceux effectués dans plusieurs environnements qui tiennent compte des interactions GxE. Il a ainsi montré que les modèles multi-traits et multi- environnements ont une précision meilleure que celle des modèles avec un trait ou un environnement. L'utilisation de modèles mixtes avec multiples environnements (Tableau 3), qui intègrent les interactions GxE, permet une amélioration de 40% de la précision de la prédiction par rapport au modèle avec un seul environnement. De même, le modèle multi-trait améliore la précision de la prédiction génomique de 12% par rapport au modèle de trait unique. L'intégration de données issues de pedigree et/ou de marqueurs permet une amélioration de la précision de la prédiction génomique pour des traits en condition multi-environnement. En effet, les modèles (voir Tableau 3) utilisés par Burgueño *et al.,* (2012) montrent que l'intégration de données de pedigree et de marqueurs dans un même modèle améliore la précision de la prédiction par rapport à un modèle qui n'intègre que les données de pedigree ou de marqueurs seuls. Dans une étude, Fodor *et al.,* (2014) montrent que l'utilisation de marqueurs génomiques dans une combinaison de la GWAS et la SelGen améliore la précision de la prédiction jusqu'à 90 % par rapport aux modèles de GWAS ou de la SelGen. En effet, ils utilisent les marqueurs SNP issus d'une étude GWAS en cofacteurs pour effectuer la prédiction génomique. Il en ressort que l'utilisation d'un modèle de prédiction combiné et de core collection comme population d'entraînement est approprié pour la Selgen chez la vigne. Liu *et al.*, (2019) montrent que l'intégration d'informations issues de marqueurs cartographiés dans un modèle statistique améliore la prédiction de traits agronomiques. La précision est maximale quand un total de 500 à 1000

marqueurs pertinents est utilisé. On observe une meilleure précision de la prédiction en incluant les effets non additifs en plus des marqueurs pertinents dans le modèle quand les données génotypiques contiennent une grande proportion d'hétérozygotes et pour des caractères complexes avec une grande proportion de variance non additive dans la variance phénotypique.

Tableau 3: Différents modèles qui intègrent l'interaction GxE et des cofacteurs.

Modèles et paramètres	Sources
Maïs : $Y = 1\mu + X_sS + Z_r r + Z_g g + \varepsilon$ Y : vecteur de la variable de réponse ; 1 : vecteur colonne de uns ; μ : moyenne générale ; X_s : matrice des effets fixes des sites S: vecteur des effets fixes des sites Z_r : matrices des effets aléatoires des répétitions dans les sites; Zg : matrice des effets aléatoires des génotypes dans les sites individuels.	(J. F. Burgueño *et al.,* 2011) (Oakey *et al.,* 2016) (Okeke *et al.,* 2017)
$y = X\beta + Zg_p + Zg_M + \varepsilon$ y : vecteur de la variable phénotypique ; X et Z : matrices des effets des génotypes; β: effets génétiques aléatoires ; g_p : régression sur le pedigree ; g_M : régression sur des marqueurs.	(J. Burgueño *et al.,* 2012)

Dans notre étude il est question d'évaluer l'impact de la prise en compte de l'interaction GxE sur la précision de la prédiction génomique. Cette évaluation est faite sur un ensemble de données issues d'expérimentations mises en œuvres dans plusieurs environnements plus ou moins stressés, différents traits tels que la date de floraison, le poids du fruit, etc. Ces données proviennent du phénotypage des individus de deux populations différentes (décrites dans la partie matériel ci-dessous). Il s'agit d'une population multi-parentale (MAGIC) non structurée et d'un panel de GWAS diversifié et partiellement structuré. Il est question d'estimer l'apport de la prédiction génomique dans l'étude des traits de telles populations, évaluées dans les conditions énoncées. Dans ce rapport, nous présenterons d'abord une analyse descriptive des données dont nous disposons, puis la mise en œuvre de la prédiction à l'aide de modèles (1) simple environnement (G), (2) prenant en compte l'effet de l'environnement (G+E), (3) intégrant les interactions GxE. Pour finir, on intégrera (4) les cofacteurs environnementaux dans le modèle (3) qui prend en compte GxE. Le matériel et les méthodes utilisés sont décrits dans la suite du rapport, puis les résultats et la discussion.

2. MATERIEL ET METHODES

Les analyses porteront sur une population multi-parentale et un panel d'associations préalablement phénotypés et génotypés.

2.1 Matériel

2.1.1 Jeu de données de la population MAGIC

La population MAGIC (*Multiparent advanced generation intercross*) décrite par Pascual *et al.,* (2015) a été obtenue à partir de multiples croisements. Elle est constituée de 397 lignées (250 étudiées ici). Elle a été obtenue à partir de plusieurs croisements réalisés entre 8 lignées. Il s'agit de 4 lignées de tomate à gros fruit et 4 lignées à petit fruit (figure 6). Cette population a été obtenue après 4 générations de croisements et 3 autofécondations (Pascual *et al.,* 2015b). Ce type de plan de croisements favorise une grande diversité génétique et réduit la structure au sein de la population. Il a été vérifié qu'il n'y a pas de structure dans cette population car il n'y a qu'un seul groupe selon l'étude d'Evanno *et al.*, (2005). Aussi, le déséquilibre de liaison (DL) dans les chromosomes diminue quand la distance génétique augmente. Il est inférieur à 0,70 pour une distance génétique inférieure à 5 cM et est inférieur à 0,3 quand la distance est de 25 cM. De même, le DL diminue quand la distance physique augmente en passant de 0,45 à 1 kb à moins de 0,2 à 2 Mb en moyenne (Pascual *et al.,* 2015b).

Le reséquençage des lignées fondatrices de la population MAGIC a permis d'identifier 4 millions de polymorphismes dont 1345 SNP qui ont été utilisés pour analyser la population MAGIC (Causse *et al.,* 2013). Cette population a été phénotypée pour 6 caractères concernant la croissance de la plante et la qualité du fruit. Il s'agit de : la date de floraison (flw), le nombre de fleurs (nflw), le taux de fructification qui correspond au rapport du nombre de fruits sur le nombre de fleurs (fset), la longueur de la feuille (leaf), le poids du fruit (fw), le nombre de fruits (nfr). Le phénotypage de ces différents caractères s'est fait dans deux pays (France et Maroc) dans des essais soumis à différentes conditions de stress, qui représentent 7 environnements différents. Il s'agit de 3 environnements en France, condition favorable à Avignon en 2012 (Avi12), condition contrôle (optimale) à Avignon en 2017 (Avi17), condition de stress thermique à Avignon en 2017 (HAvi17 ; en moyenne +5°C par rapport au contrôle), et de 4 environnements au Maroc, condition de contrôle en 2015 (Mor15), stress hydrique en 2015 (WDMor15, 50% de réduction de l'apport d'eau), fort stress salin en 2016 (HSMor16, EC de 6,5dS/m), faible stress salin en 2016 (LSMor16, EC de 3,5dS/m). L'impact de ces différents environnements sur la détection de QTL a été décrit par Diouf *et al.,* (2018, 2020).

2.1.2 Jeu de données de la population GWAS

Cette population est constituée de 140 accessions de tomate de petite taille utilisée pour faire une étude de GWAS. Elle comprend 11 accessions de *S. pimpinellifolium* (SP, espèce sauvage)

originaires du Pérou et de l'Équateur, de 108 accessions de *S. lycopersicum* var. *cerasiforme* (SLC) originaires d'Amérique du Sud et 21 accessions de tomates cerise commerciales. Ces accessions constituent une core-collection décrite par Albert *et al.,* (2016). Cette core-collection a pour but de maximiser la diversité génétique et réduire la structure intra-population. Cette population a été génotypée à l'aide de la puce Infinium de la tomate mise en œuvre dans le projet SolCAP (http://solcap.msu.edu/) et comprenant 8000 SNP utiles. Le seuil des fréquences des allèles mineurs est de 0.04. Les données manquantes après purification ont été estimées à 10% pour les SNPs. Le phénotypage a été fait pour 5 traits dans deux endroits à savoir Avignon en France (de mars à juillet 2014) et Agadir au Maroc (de décembre 2013 à mars 2014). Il s'agit de traits liés à la croissance de la plante et aux fruits à savoir la date de floraison (FLW), la longueur des feuilles (Leaf), le nombre de fruits (nfr), le poids du fruit (FW), le Brix (SSC), etc. Ces plantes ont été phénotypées dans des conditions de stress hydrique et de contrôle décrites par Albert *et al.,* (2016).

Si le COVID n'avait pas stoppé les expérimentations du printemps j'aurais dû contribuer au phénotypage de cette population cultivée en condition contrôle et de faible apport azoté. Malheureusement l'essai a été reporté à l'automne.

Les caractères et environnements des deux populations (GWAS et MAGIC) utilisés pour cette étude sont consignés dans le Tableau 4 ci-dessous.

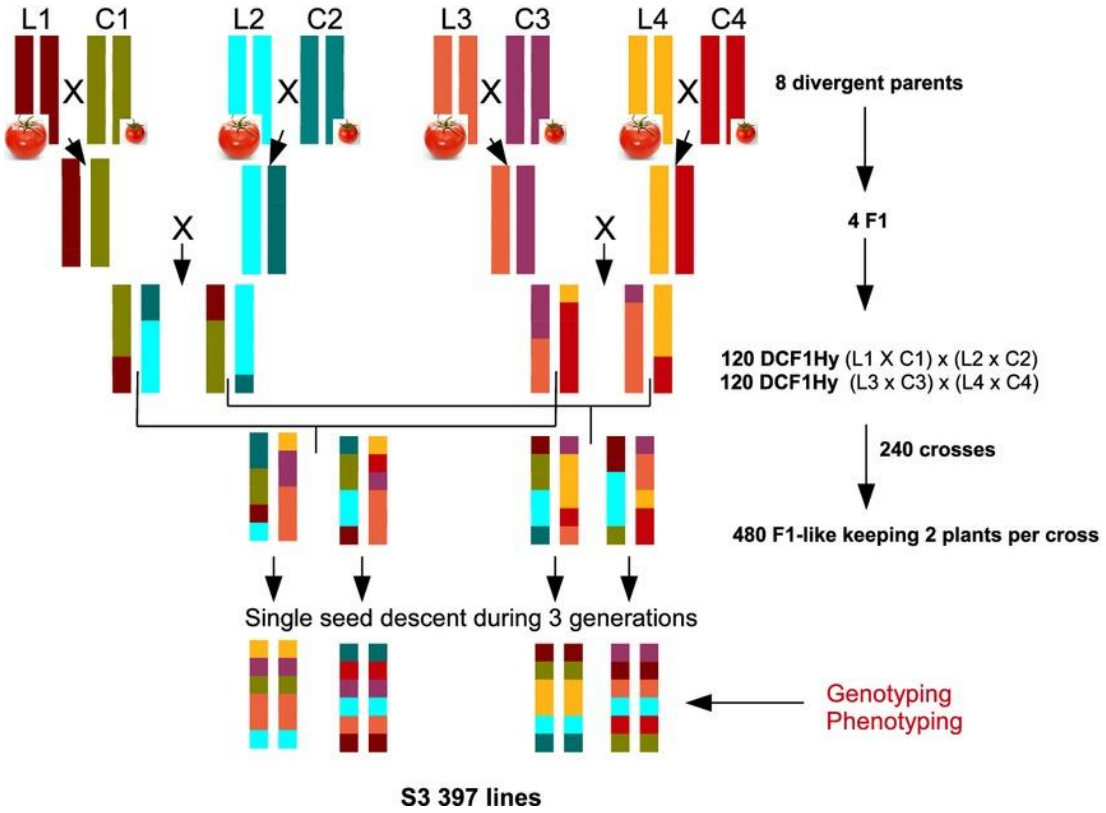


Figure 6 : Construction d'une population de tomates MAGIC à 8 voies. Les tomates à gros fruits sont L1 Levovil, L2 Stupicke PR, L3 LA0147, L4 Ferum. Les tomates à petit fruit sont C1 Cervil, C2 Criollo, C3 Plovdiv24A, C4 LA1420. DCF1Hy: hybride F1 double croisement.

Tableau 4: Informations relatives aux traits des populations MAGIC et GWAS utilisé pour cette étude.

Population	Traits	Nombre d'environnements	Environnements	Individus/Trait
MAGIC	FLW, FW	7	Avi12, Avi17, Havi17, HSMor16, LSMor16,Mor15, WDMor15	250
	NFLW, LEAF, NFR, FSET	6	Avi17, Havi17, HSMor16, LSMor16,Mor15, WDMor15	250
GWAS	FLW,FW,LEAF	4	AviS, AviT, AgaS, AgaT	141
	NFR, SSC	2	AviS, AviT	141

Tableau 5 : Données moyennes des variables environnementales dans la serre; Moyenne sur les 20 1ers jours (=P1) les 20 jours suivant P1 (=P2) et les 20 jours suivant P2 (=P3) de la population MAGIC (cofacteurs environnementaux) (Diouf *et al.*, 2020).

cofacteur	Avi12	Avi17	HAvi17	HSMor16	LSMor16	Mor15	WDMor15
P1_Tmin	14,79	15,46	17,21	10,10	10,10	7,54	7,54
P1_Tmean	20,43	20,77	24,86	21,03	21,03	16,82	16,82
P1_Tmax	29,06	28,63	34,32	42,61	42,61	32,52	32,52
P1_RH	57,86	61,90	62,20	54,81	54,81	58,32	58,32
P1_Vpd	1,38	1,50	1,95	1,37	1,37	1,11	1,11
P1_Th.Amp	14,27	13,18	17,12	32,52	32,52	24,99	24,99
P1_Ec	1,33	1,50	1,50	3,73	3,73	2,67	2,67
P1_SDD	127,62	121,83	162,93	171,61	171,61	116,22	116,22
P1_WD	0,00	0,00	0,00	0,00	0,00	0,00	0,25
P2_Tmin	15,98	15,69	19,65	10,34	10,34	7,61	7,61
P2_Tmean	20,26	20,90	26,72	19,37	19,37	17,09	17,09
P2_Tmax	26,19	28,53	34,52	34,13	34,13	34,47	34,47
P2_RH	71,05	57,95	69,67	58,31	58,31	64,86	64,86
P2_Vpd	1,69	1,43	2,46	1,31	1,31	1,27	1,27
P2_Th.Amp	10,21	12,84	14,87	23,78	23,78	26,87	26,87
P2_Ec	1,33	1,50	1,50	3,57	3,47	1,91	1,91
P2_SDD	354,47	381,12	483,60	456,44	456,44	355,97	355,97
P2_WD	0,00	0,00	0,00	0,00	0,00	0,00	0,50
P3_Tmin	14,98	16,41	21,06	12,73	12,73	8,61	8,61
P3_Tmean	20,74	21,99	27,10	20,61	20,61	19,24	19,24
P3_Tmax	26,61	29,50	34,17	34,08	34,08	37,36	37,36
P3_RH	70,61	69,52	71,77	60,80	60,80	59,98	59,98
P3_Vpd	1,73	1,82	2,58	1,47	1,47	1,33	1,33
P3_Th.Amp	11,63	13,09	13,12	21,35	21,35	28,74	28,74
P3_Ec	1,33	1,50	1,50	5,58	4,97	1,72	1,72
P3_SDD	579,41	611,69	844,52	709,69	709,69	613,83	613,83
P3_WD	0,00	0,00	0,00	0,00	0,00	0,00	0,50

Tableau 6: Liste des caractères des population MAGIC et GWAS étudié par les différents modèles.

Modèles	Population	Caractères
Simple environnement (G) : BL et BayesC	MAGIC	FLW, FW, NFLW, LEAF, NFR, FSET
	GWAS	FLW, FW, LEAF, NFR, SSC
Inter-environnement (G+E) : RKHS	MAGIC	FLW, FW, LEAF, NFR
	GWAS	FLW, FW, LEAF, NFR, SSC
Interaction GxE : RKHS	MAGIC	FLW, FW, LEAF, NFR
	GWAS	FLW, FW, LEAF, NFR, SSC
GxE + Cofacteurs environnementaux	MAGIC	FLW, FW, LEAF

2.2 Méthodes

Les populations MAGIC et GWAS seront toutes les deux utilisées pour la mise en œuvre des différentes méthodes. L'objectif étant de prédire les valeurs phénotypiques des caractères présentés dans le matériel dans différentes conditions. Cette prédiction est faite en prenant en compte plusieurs paramètres tels que le nombre de cycles, les modèles, la composition des populations d'entrainement, les interactions GxE, l'intégration de cofacteurs environnementaux. Les caractères étudiés dans chacun des modèles sont consignés dans le tableau 6 ci-dessus. Cependant, seuls les traits FLW, FW et LEAF sont présentés dans les résultats, pour des contraintes de pages et parce que ces trois caractères sont représentatifs de l'héritabilité des caractères. Les scripts utilisés pour les modèles simple environnement et GxE sont proposés en annexes 13 et 14.

2.2.1 Prédire les caractères sans les interactions GxE : package BGLR avec les modèles BL et BayesC

Le package BGLR (*Bayesian Genomic Linear Regression*) disponible sur R, décrit par Pérez & de los Campos, (2014) est utilisé pour la prédiction des caractères de la plante et du fruit à l'aide des modèles BL et BayesC, qui avaient donné les résultats les plus intéressants dans l'étude de Picard (2015). Dans un premier temps, ces modèles sont utilisés pour prédire les caractères à partir de données phénotypiques mesurées dans chaque environnement indépendamment. Le modèle linéaire utilisé est le suivant :

$$y_j = \mu 1 + X_j\beta_j + \epsilon_j \qquad (1)$$

Où y est le vecteur des valeurs phénotypiques enregistrés, μ est la moyenne. X_j est la matrice des génotypes, $X_j = \{x_{ijk}\}$, $\beta_j = \{\beta_{jk}\}$ le vecteur des effets associés aux génotypes et $\beta_j \sim N(0, I\sigma^2_{\beta j})$. Enfin, ϵ_j sont les résidus, indépendants et $\epsilon_j \sim N(0, I\sigma^2_{\epsilon j})$, (k =1,2, 3, …, p marqueurs et i= 1,2,3, …, n individus, j=1,2,3, …, s environnements).

L'utilisation du package BGLR exige de fixer un certain nombre de paramètres, tels que le nombre de cycles testés et la composition de la population d'entrainement (PE) et de validation (PV).

Une étude précédente (Picard, 2015) a montré que l'utilisation de 100 cycles permet d'optimiser le temps de calcul des modèles pour les différents caractères et populations, bien que dans la littérature ce sont plutôt 1000 à 10000 cycles qui sont testés. Aussi, la composition de la population d'entrainement et de validation est faite à deux niveaux, permettant ainsi de voir son effet sur la précision de la PG. Le premier choix de composition est de faire les prédictions sur une PE de 75% avec une PV de 25% et le second est de faire une PE de 50% de même que la PV. La PE de 25% a été testée mais n'est pas utilisée pour les prédictions car la précision de prédiction est trop faible. Les modèles BL et BayesC sont utilisés et appliqués sur les deux populations GWAS et MAGIC. Une comparaison de la précision de ces deux modèles est faite. Elle permet d'évaluer lequel de ces modèles prédit mieux les caractères.

Les deux modèles sont testés avec des BLUPs issus des données de MAGIC et GWAS estimés sur l'ensemble des environnements afin de voir leur effet sur la précision de la prédiction. Les BLUPs sont obtenus à l'aide du modèle dans lequel les effets des génotypes et des environnements sont considérés comme aléatoires (voir modèle ci-dessous).

$$P_{ij} = \mu + G_i + E_j + R_{ij} \qquad (2)$$

Où P_{ij} représente les BLUPs, G_i représente les effets des génotypes et $G_i \sim N(0, \sigma^2_{Gi})$. E_j représente les effets des environnements et $E_j \sim N(0, \sigma^2_{Ej})$. R_{ij} représente la résiduelle et $R_{ij} \sim N(\mu, \sigma^2_{Rij})$.

De plus, des prédictions sont faites en considérant les effets des marqueurs constants entre les environnements (**modèle inter-environnement**) en vue d'évaluer l'impact de cette approche sur la précision de la prédiction. Pour ce faire le modèle RKHS (Bayesian Reproducing Kernel Hilbert Spaces regressions) est utilisé. (1) est modifié en considérant que β est constante quel que soit l'environnement.

$$yj = 1\mu_j + X_j\beta + \epsilon_j \qquad (3a)$$

(j=1,2,3,…, s environnements)

$y_j = \{y_{ij}\}$ est le vecteur des phénotypes pour les différents environnements, μ_j est la moyenne des phénotypes dans les différents environnements. $X_j = \{x_{ij}\}$ est la matrice des génotypes, β matrice des effets des marqueurs $\beta \sim N(0, I\sigma^2_{\beta})$, ϵ_j est le vecteur de la résiduelle du modèle, $\epsilon_j \sim N(0, I\sigma^2_{\epsilon_j})$.

L'écriture matricielle du modèle (3a) pour J=3 se présente comme suit :

$$\begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ 1\mu_2 \\ 1\mu_3 \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \qquad (3b)$$

La validation croisée est utilisée pour la composition de la PE et PV. Deux schémas de validation croisée sont utilisés. Le premier schéma (CV1) vise à évaluer la capacité du modèle à prédire les performances des génotypes qui n'ont été phénotypés dans aucun environnement (données manquantes). Ces génotypes seront prédits à partir des informations phénotypiques des autres lignées en assignant aléatoirement 70% des individus à PE et les 30% restant à PV. Le deuxième schéma (CV2) consiste à évaluer les performances de lignées présentes dans certains environnements mais absentes dans les autres, ce dernier est détaillé dans Lopez-Cruz *et al.,* (2015).

2.2.2 Prédire les caractères en prenant en compte les multiples environnements et les interactions GxE : package BGLR avec le modèle RKHS

Le package BGLR a été utilisé pour évaluer l'influence de GxE sur la précision de prédiction, avec le modèle RKHS, utilisé pour les prédictions du modèle inter-environnement et avec les mêmes paramètres de modèle. A savoir la composition de la population d'entrainement et de validation, le nombre d'itérations, etc. Le modèle utilisé est basé sur le modèle multi-environnement proposé par Burgueño *et al.,* (2012); Lopez-Cruz *et al.,* (2015).

Le modèle qui prend en compte les interactions GxE considère que l'effet de chaque génotype dans

chaque environnement j est la somme de l'effet commun à tous les environnements (b_{0k}) et de la variation aléatoire propre à chaque environnement (b_{jk}). L'équation se présente comme suit :

$y_{ij}= 1\mu + \sum_{k=1}^{p} x_{ijk}\,(b_{0k}+ b_{jk}) + \epsilon_{ij}$, (4a) avec $\beta_{jk} = b_{0k}+ b_{jk}$, (k =1,2,3,…,p marqueurs et i= 1,2,3,…, n individus).

L'écriture matricielle de cette équation se présente comme suit :

$$\begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ 1\mu_2 \\ 1\mu_3 \end{bmatrix} + \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \mathbf{b_0} + \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \mathbf{b_3} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix} \quad (4b)$$

Avec b_0 qui correspond au vecteur des effets des génotypes communs à tous les environnements, avec $b_0 \sim N(0, I\sigma^2_{b0})$, b_j est l'effet de l'interaction GxE avec $b_j \sim N(0, I\sigma^2_{bj})$ et $\epsilon_j \sim N(0, I\sigma^2\epsilon_j)$.

La composition de la PE et PV est faite suivant les schémas décrits dans le modèle inter-environnement ci-dessus (voir 2.2.1).

Une prédiction est faite pour le caractère flw de la population GWAS avec 50000 itérations et 5000 burnIn. Cette dernière a mis environ 30h à tourner sur le serveur de l'INRAE contre 3 à 4h pour celle que nous avons effectué sur le même caractère avec 1200 Itérations et 200 burnIn. Afin de gagner du temps, toutes les analyses sont faites avec 1200 Itérations et 200 burnIn.

2.2.3 Prédire les caractères en prenant en compte les cofacteurs environnementaux en plus des interactions GxE.

Après la prise en compte des interactions GxE, on a intégré des cofacteurs environnementaux (de la population MAGIC) au modèle afin d'évaluer leur impact sur la précision de la prédiction. Ces cofacteurs représentent les moyennes de variables environnementales enregistrées pour décrire les conditions de stress ou contrôle de la population MAGIC. Ces variables ont été séparées suivant 3 périodes, celles recueillies sur les 20 premiers jours après plantation (=P1), les 20 jours suivant P1 (=P2) et les 20 jours suivant P2 (=P3). Il s'agit de valeur des paramètres de la température (min, moy, max, amplitude thermale), de l'humidité relative (RH), de la somme des degrés jours (SSD), etc. Ces paramètres ont été détaillés par Diouf *et al.*, (2020) et sont présentés dans le tableau 5.

En ce qui concerne leur intégration dans le modèle de prédiction qui prend en compte l'interaction GxE dans tous les environnements, en CV1 et CV2, ils ont été rajoutés en effet fixe. Il est donc question d'évaluer leur impact potentiel sur la précision de la prédiction génomique. Ils seront évalués dans les mêmes conditions, avec les mêmes paramètres que les modèles précédents.

3. RESULTATS

3.1 Analyse descriptive des données

3.1.1 Distribution des caractères

La distribution des caractères dans la population MAGIC (250 lignées) de la Figure 7 ci-dessous montre une variation au sein des données de chaque caractère. Les dates de floraison (flw) varient d'un environnement à l'autre. En moyenne, les tomates des essais réalisés au Maroc fleurissent plus tardivement que celles des essais réalisés à Avignon (tous environnements confondus). De plus, à Avignon, les tomates de l'environnement Havi17 (stress thermique) fleurissent plus précocement. On constate que les tomates soumises au stress salin fleurissent plus tardivement que celles du stress hydrique. Cependant, il y a très peu de variation entre les dates de floraison des tomates mises en stress salin fort et faible pour les mêmes températures. De même pour les tomates soumises au stress hydrique (tableau 5). Le stress salin et le stress hydrique affectent peu la date de floraison. La différence observée entre le stress thermique, le stress salin et le stress hydrique est due à l'écart de température entre Tmin et Tmax respectivement (14,8 °C, 23,7 °C et 26,8) et aux autres cofacteurs environnementaux. Enfin, la différence de date de floraison est fonction du lieu de phénotypage.

Le poids du fruit (fw) varie d'un environnement à l'autre. Globalement, en condition de stress les fruits sont plus petits. En effet, il y a des plus gros fruits dans l'environnement Avi12 et Avi17 tandis qu'en stress thermique (Havi17) le poids des fruits diminue passant respectivement de 58 g à 48 g puis à 38g environ. Cependant, le stress salin (fort et faible) entraine une plus forte diminution du poids des fruits par rapport aux stress thermique et hydrique. Le fort stress salin est celui qui affecte le plus le poids du fruit avec la majorité des fruits pesant moins de 20 g. Cela s'explique par le fait qu'en présence d'une forte concentration de sel dans le substrat, la plante absorbe peu d'eau et de nutriments indispensables au bon développement des fruits.

La longueur des feuilles (leaf) varie également d'un environnement à l'autre. Globalement, les environnements de contrôle dans tous les sites présentent des feuilles plus longues (avec Av17 ou la majorité des feuilles a une longueur comprise entre 30 cm et 40 cm ; figure 7) que les environnements de stress. Le stress thermique affecte la longueur de la feuille plus que les autres stress, avec des longueurs comprises entre 15 et 25 cm.

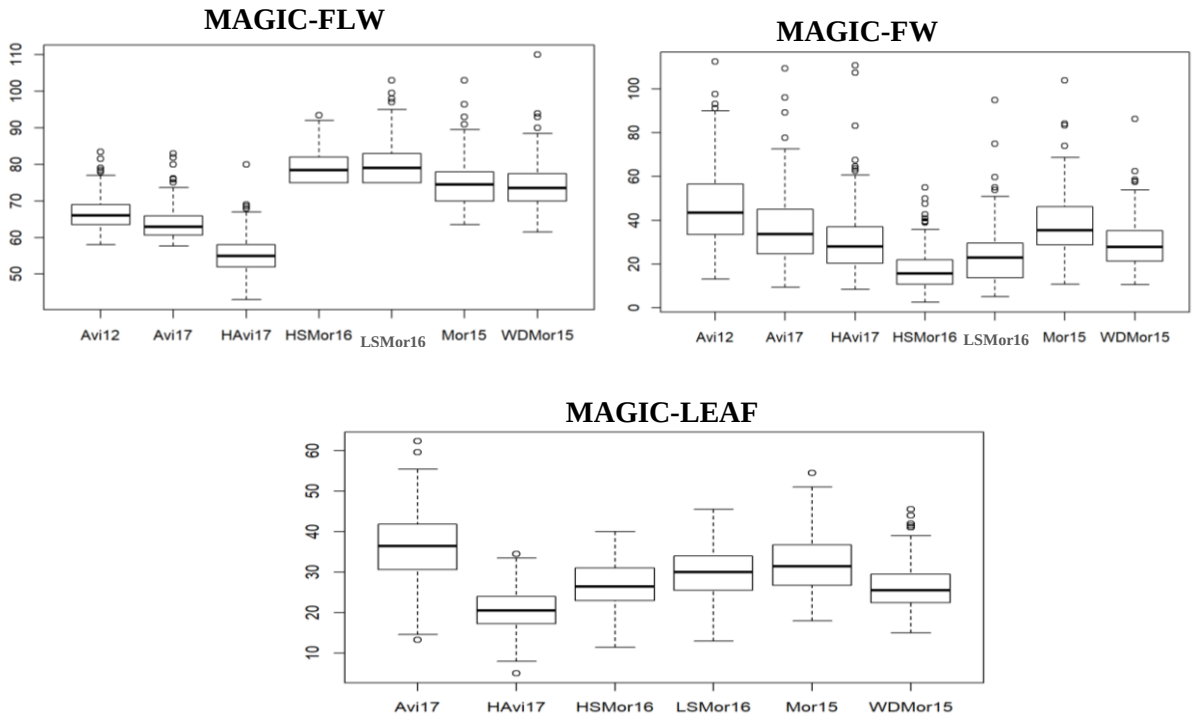


Figure 7: Boxplot de la distribution des caractères floraison (FLW), poids du fruit (FW) et longueur des feuilles (LEAF) par environnements à Avignon (Avi12: condition favorable, Avi17: condition de contrôle, HAvi17: condition de stress thermique) et au Maroc (HSMor16 et LSMor16: condition de fort et faible stress salin, Mor15 et WDMor15 : condition contrôle et stress hydrique) de la population MAGIC.

La distribution des caractères dans le panel de GWAS suivant les environnements de la Figure 8 ci-dessous montre une variabilité suivant les caractères et les environnements. Le caractère date de floraison (flw) présente peu de variabilité au sein des différents environnements. Les dates de floraison sont plus précoces à Agadir par rapport à Avignon. Cependant, pour un même lieu les dates de floraison sont les mêmes en présence ou absence de stress hydrique. Ce résultat confirme le fait que le stress hydrique n'affecte pas la date de floraison. La différence de date de floraison peut s'expliquer par d'autres facteurs environnementaux puisque les phénotypage ont été faits à deux périodes différentes hiver-début printemps (décembre 2013 à mars 2014) à Agadir et printemps-été (mars à juillet 2014) à Avignon.

En ce qui concerne le poids du fruit (fw), il y a une variation inter-environnement et intra-site. En effet, la différence de poids du fruit entre les sites est faible. Aussi, les fruits en condition de stress sont plus petit dans leur majorité (environ 5 g de moins) que ceux des conditions contrôle. Le stress hydrique entraine la diminution de la taille du fruit.

Enfin, la longueur de la feuille (leaf) varie d'un environnement à un autre et d'un site à un autre. En condition contrôle, la longueur de la feuille est la même quel que soit le site. Cependant, en condition

de stress hydrique la longueur des feuilles varie d'un site à l'autre. En effet, les feuilles sont plus longues à AviS que à AgaS. Le stress hydrique entraine une réduction de la surface foliaire.

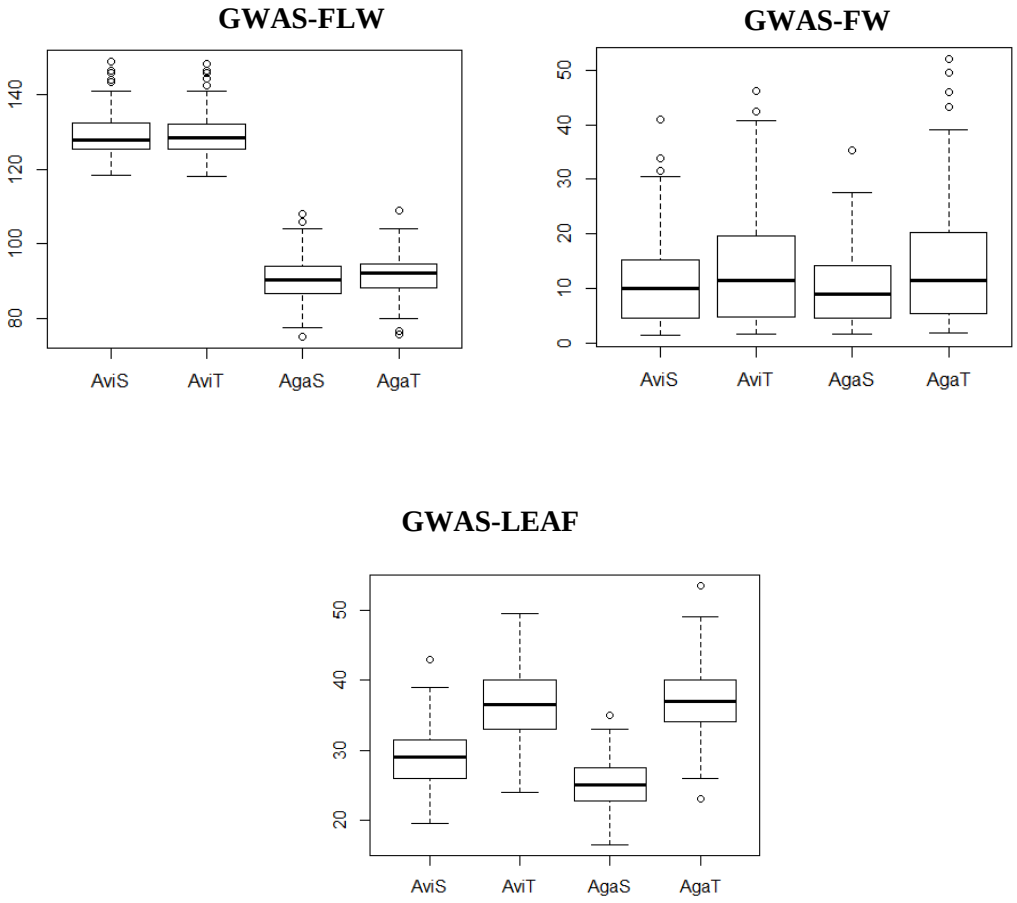


Figure 8: Boxplot de la distribution des caractères floraison, poids du fruit et longueur des feuilles par environnement à Avignon en 2014 (AviT :condition contrôle, AviS: condition de stress hydrique) et Agadir en 2014 (AgaT :condition contrôle, AgaS: condition stress hydrique) de la population GWAS.

3.1.2 Corrélations entre les environnements d'un caractère et entre les environnements de plusieurs caractères

Les corrélations entre les environnements des caractères de la population MAGIC (figure 9) varient suivant les caractères et sont positives. En ce qui concerne le caractère date de floraison (flw), tous les environnements sont positivement corrélés entre eux. Tous les essais du Maroc sont fortement corrélés entre eux (supérieure à 0,50, avec 0,86 pour Mor15 et WDMor15). Les environnements favorables et de contrôle (Avi12 et Avi17) d'Avignon sont fortement corrélés (0,64). Havi17 est moins bien corrélé avec Avi17 (0,54) comme tous les environnements du Maroc de faible et fort stress salin (respectivement 0,45 et 0,47).

Pour le poids du fruit, les environnements sont moins corrélés dans l'ensemble (corrélation inférieure à 0,30). Cependant, les environnements Mor15, WDMor15 et Avi12 sont très corrélés (avec 0,82 pour Mor15-WDMor15, 0,64 pour Mor15-Avi12 et 0,65 pour WDMor15-Avi12). Les environnements de Avignon sont bien corrélés entre eux. Les environnements de stress salin et thermique sont peu corrélés entre eux avec 0,24 pour Havi17-HSMor16 et 0,20 pour Havi17-LSMor16.

Enfin, en ce qui concerne la longueur de la feuille, tous les environnements sont très faiblement corrélés quels que soit le site (Avignon, Maroc). Cependant, les environnements de faible et de fort stress salin sont bien corrélés entre eux (0,45). Il en est de même pour la condition de contrôle et de stress thermique à Avignon 2017 (0,53).

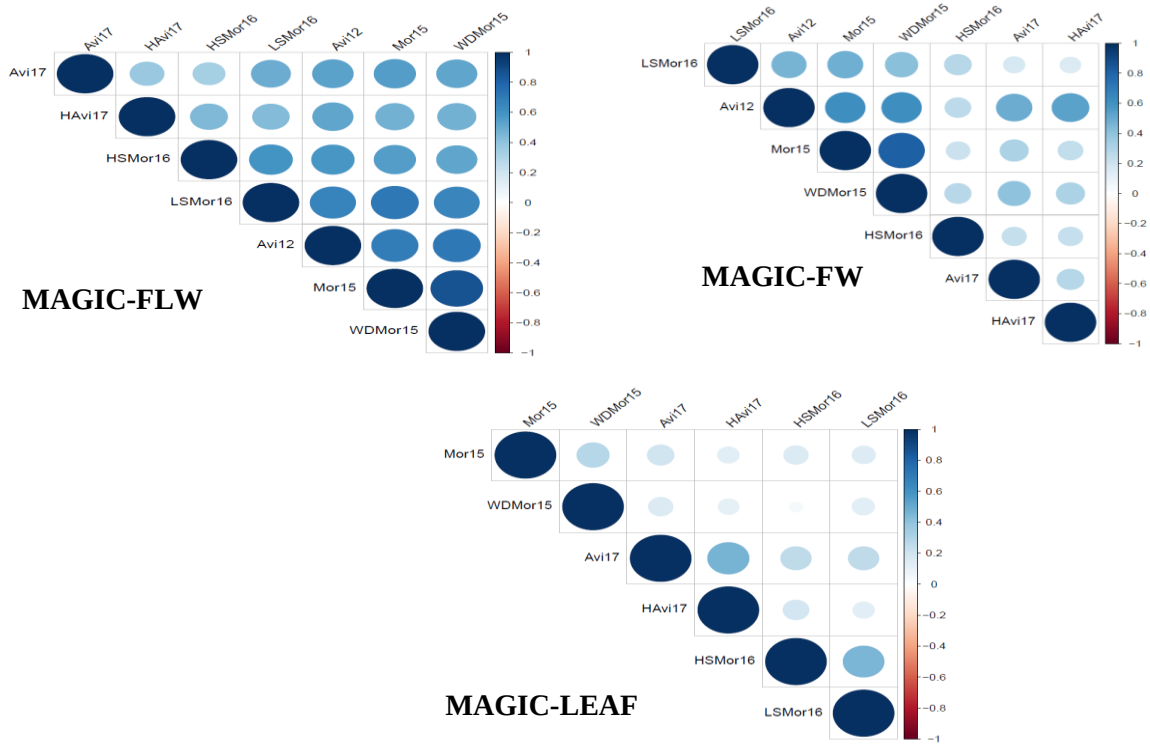
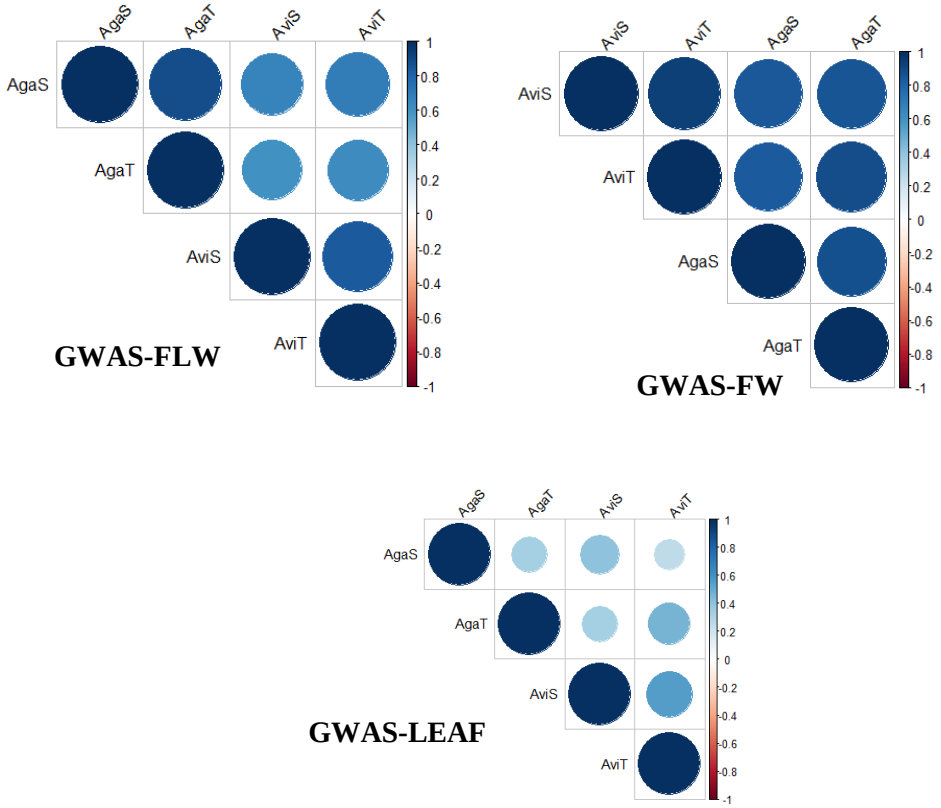


Figure 9: Corrélation entre les différents environnements de chaque caractère date de floraison (flw), poids du fruit (fw) et longueur de la feuille de la population MAGIC.

La figure 10 ci-dessous montre les corrélations des environnements des différents caractères pour le panel de GWAS. Les dates de floraison dans les 4 environnements sont corrélées (0,62 à 0,88) entre eux en absence ou en présence du stress hydrique (tous sites confondus). Cependant, AviS et AviT sont moins corrélés (entre 0,60 et 0,69) aux environnements de Agadir. Pour le poids du fruit, les environnements sont fortement corrélés (0,83 à 0,93) par site (AviT avec AviS et AgaT avec AgaS), tandis que AviS et AgaT ne sont pas corrélés (moins de 0,10). Les environnements de Avignon et Agadir sont peu corrélés entre eux quel que soit le niveau de stress. Enfin, les environnements du caractère leaf sont peu corrélés entre eux (corrélation comprise entre 0,25 et

0,55). Les environnements de stress sont mieux corrélés entre eux qu'avec les conditions de contrôle du site différent du leur. C'est le cas pour AviT et AgaS qui sont faiblement corrélés (0,25) tandis que AviS et AgaS sont mieux corrélés (0,40).



<u>Figure 10</u>: Corrélations entre les différents environnements de chaque caractère date de floraison (flw), poids du fruit (fw) et longueur de la feuille (LEAF) de la population GWAS.

3.1.3 Composantes de la variance pour les différents modèles (simple environnement, G+E et GxE)

Les héritabilités des caractères dans chaque dispositif ont été calculées préalablement. Le caractère flw est plus héritable avec une héritabilité moyenne de 0,72, tandis que fw a une héritabilité moyenne de 0,63 ; le moins héritable est le caractère leaf avec une héritabilité moyenne de 0,31 pour la population MAGIC (Diouf *et al.*, 2020). Les héritabilités suivant les sites et les conditions de présence ou d'absence de stress ont été obtenues pour la population GWAS par Albert *et al.*, (2016). L'héritabilité de flw est de 0,73 pour AviT et 0,71 pour AviS, tandis qu'elle est de 0,81 à AgaT et 0,82 à AgaS. Celle de fw est de 0,92 à AviT et de 0,90 à AviS, tandis qu'elle est de 0,95 à AgaT et de 0,93 à AgaS. Enfin, l'héritabilité de leaf est de 0,52 pour AviT et 0,55 pour AviS, tandis qu'elle est de 0,53 à AgaT et 0,71 à AgaS. La date de floraison est le caractère le plus héritable dans cette population et Leaf est le moins héritable.

Les composantes de la variance ont été estimées par le package BGLR pour les différentes

populations, les caractères et les modèles (G, G+E, GxE). Les résultats se présentent comme suit :

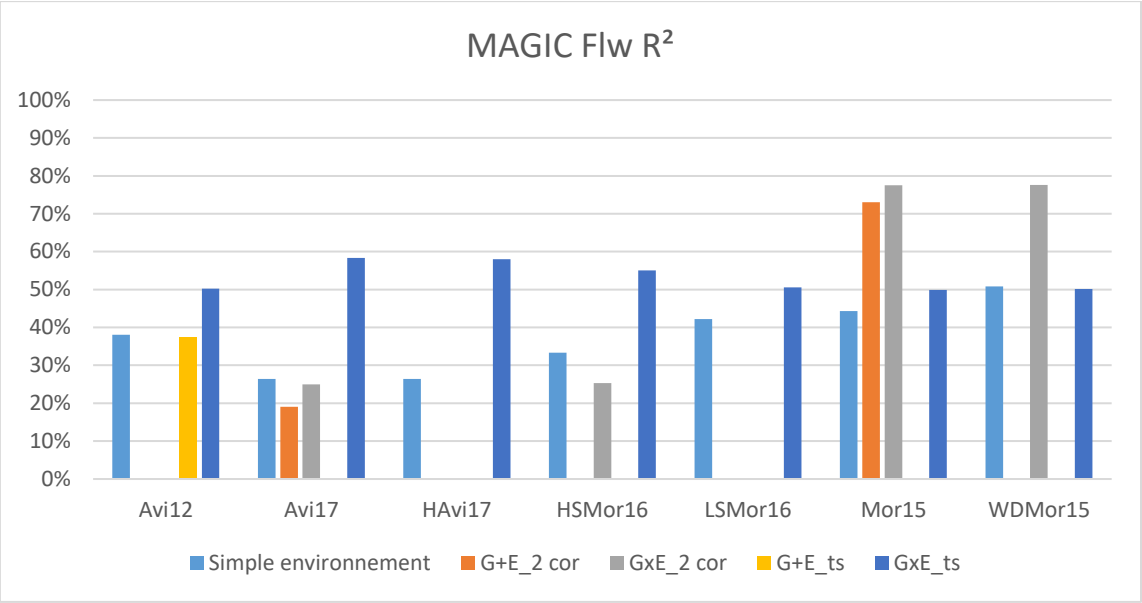
- **Modèle simple environnement (G)**

Population MAGIC : La proportion de la variance expliquée par les génotypes, représenté par le R^2, varie entre 19% et 51% pour tous les caractères confondus. Le R^2 diminue quand on passe d'un caractère très héritable tel que flw (Figure 11) à un caractère peu héritable leaf (Annexe 4, Magic).

Population GWAS : La proportion de la variance expliquée par les génotypes varie entre 27% et 72% suivant les environnements (Figure 11 et Annexe 4, Gwas).

- **Modèle inter-environnement (G+E)**

La décomposition des variances des caractères suivant la corrélation entre les environnements par paire (Figure 11 ; Annexe 3 et 4) montre que la proportion de variance génétique varie suivant la corrélation des environnements quels que soit la population et le caractère. En effet, pour des environnements très corrélés (Mor15-WDMor15 avec 0,86, Figure 11 ou AviS-AviT avec 0,93, annexe 3, Gwas) le R^2 est plus élevé (respectivement 73% et 82%). Tandis que pour les environnements peu corrélés (Havi17-LSMor16 avec 0,20, Annexe 3 Magic, ou HSMor16-WDMor15 avec 0,04, Annexe 3 Magic) le R^2 est plus faible (respectivement 11% et 7%). Les part de variance génétique estimées sur l'ensemble des environnements présentent une variance de l'erreur plus grande que celle de la variance génétique pour la population Magic. Avec des valeurs qui évoluent dans le sens contraire de l'héritabilité des caractères, voir figure 11, Annexes 3 et 4 (Magic), la variance de l'erreur est plus élevée quand l'héritabilité est faible. Tandis que la variance de l'erreur est plus faible que la variance génétique pour la population GWAS (Figure 11, Annexe 3) que la variance de l'erreur quel que soit le caractère. La variance de l'erreur est d'autant plus faible quand les environnements du caractère sont très corrélés entre eux. C'est le cas pour fw dont les environnements sont très corrélés. La différence observée au niveau des composantes de la variance entre la population MAGIC et GWAS est due au fait que les environnements des caractères de la population GWAS sont plus corrélés entre eux que ceux des caractères de la population MAGIC.



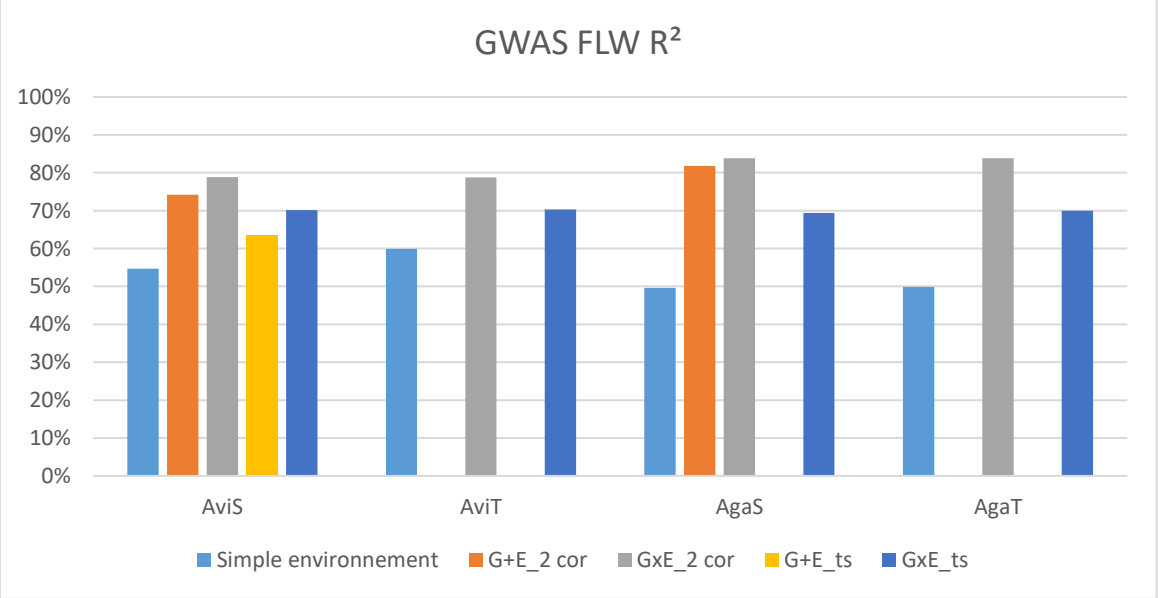


Figure 11: Proportion de variance expliquée par la variance génétique suivant les différents modèles et environnements pour le caractère flw dans la population MAGIC et GWAS. R^2 est estimé à partir du ratio de la variance (effet génétique + interaction GxE) par rapport au total de la variance (résiduelle + effet génétique+ interaction GxE). Les valeurs numériques des variances sont consignées dans les annexes 2 à 4.

- **Modèle avec interaction (GxE)**

Dans ce modèle, la variance génétique totale est la somme de la variance due aux génotypes et la variance de l'interaction GxE. La proportion de la variance due aux génotypes évolue de la même manière que celle du modèle inter-environnement pour les prédictions des caractères avec des environnements par paire. En effet, elle est élevée quand les environnements sont très corrélés. De plus, plus la corrélation entre les environnements est faible plus la variance de l'interaction GxE est élevée. La proportion expliquée par la variance génétique totale demeure plus élevée quand les environnements sont très corrélés avec un R^2 qui atteint 78% pour la MAGIC et 85% pour la GWAS (respectivement figure 11, Annexe 3 et 4 Gwas). En ce qui concerne les prédictions sur tous les environnements, la variance de l'erreur est supérieure à celle des génotypes dans la population MAGIC quel que soit le caractère. Elle est d'autant plus élevée que les environnements sont peu corrélés (figure 11 Magic, Annexe 3 et 4 Magic). Elle est d'autant plus faible quand les environnements sont très corrélés. Les valeurs de R^2 dans la GWAS sont largement supérieures à celles de la MAGIC. Globalement, les R^2 en GxE sont supérieurs à ceux du modèle G+E. La variance due aux génotypes dans le modèle GxE est meilleure que celle du modèle G+E quel que soit la population et le caractère.

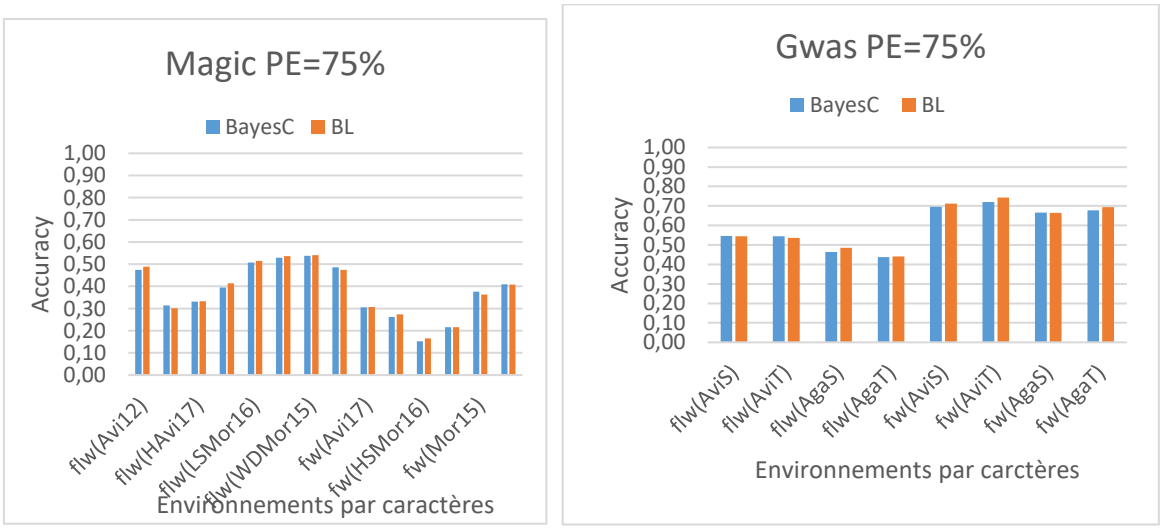


Figure 12: *Accuracy* moyenne dans chaque environnement et caractère suivant les modèles BL et BayesC, avec PE=75%, des population MAGIC et GWAS.

3.2 Résultats des prédictions génomiques

3.2.1 Résultats de prédictions avec les modèles Simple environnement, G+E et GxE

Les résultats concernant la précision des prédictions génomiques (*accuracy*) faites sur les caractères flw, fw et leaf des individus issus des populations GWAS et MAGIC sont consignés dans les figures ci-dessous (figures 12, 13, 15, 16; Annexes 6, 7, 9et 10). En moyenne, les prédictions sont supérieures dans la population GWAS par rapport à la population MAGIC, bien que cette dernière soit de taille de population plus grande (250 individus contre 141 pour la Gwas) mais la part variance génétique est plus importante en GWAS. On rappelle que le schéma de composition de la PE, CV1 vise à évaluer la capacité du modèle à prédire les performances des génotypes qui n'ont été phénotypés dans aucun environnement. Tandis que le schéma (CV2) consiste à évaluer les performances de lignées présentes dans certains environnements mais absentes dans les autres.

a. Comparaison des résultats de prédiction avec 1200 itérations avec ceux avec 50000 itérations.

Les prédictions avec un grand nombre d'itérations (50000 iter et 5000 burnIn) comme c'est le cas dans les publications, améliorent très peu la précision de la prédiction (figure 13 et 14) par rapport à celles qui sont faites avec moins d'itérations (1200 iter et 200 burnIn). Cependant, un grand nombre d'itération permet de resserrer la valeur autour de la moyenne de l'*accuracy*, d'où les boxplots de plus petite taille sont observés (figure 13). Pour les prédictions GxE avec 50000 itérations, le temps d'exécution sur le serveur de l'INRAE est d'environ 30 heures contre au plus 3 heures pour 1200 Itérations, pour le même caractère flw de la population Gwas. Nous avons donc décidé de faire toutes les prédictions avec 1200 Itérations et 200 burnIn.

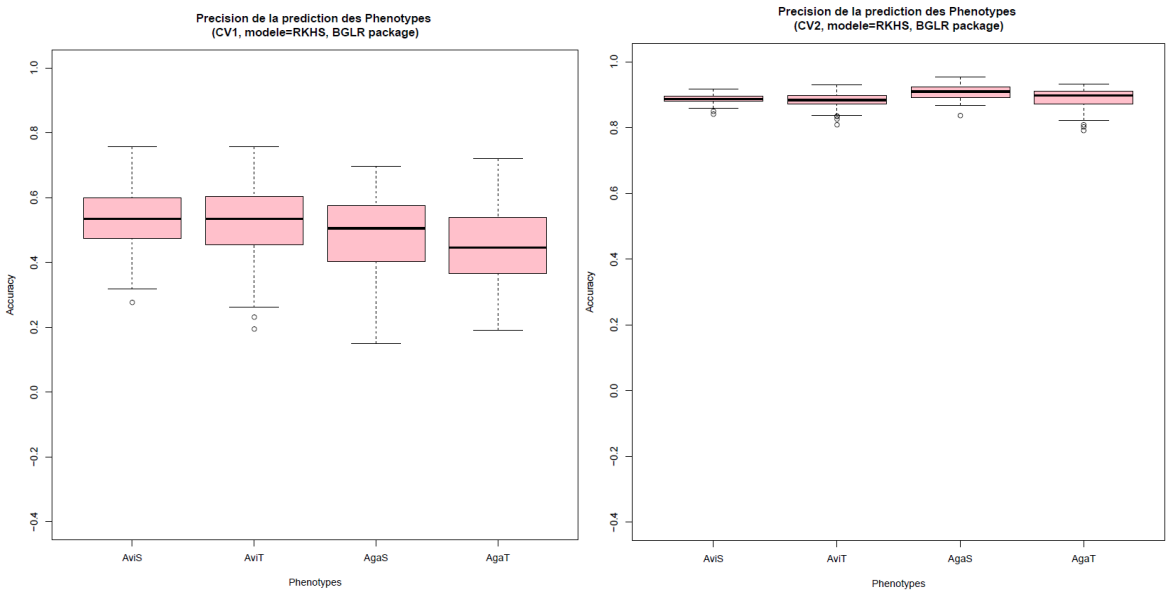


Figure 13 : Boxplot de *Accuracy* du caractère flw de la population GWAS obtenus à partir du modèle GxE dans CV1 et CV2. Prédiction fait avec 1200 itérations et 200 burnIn.

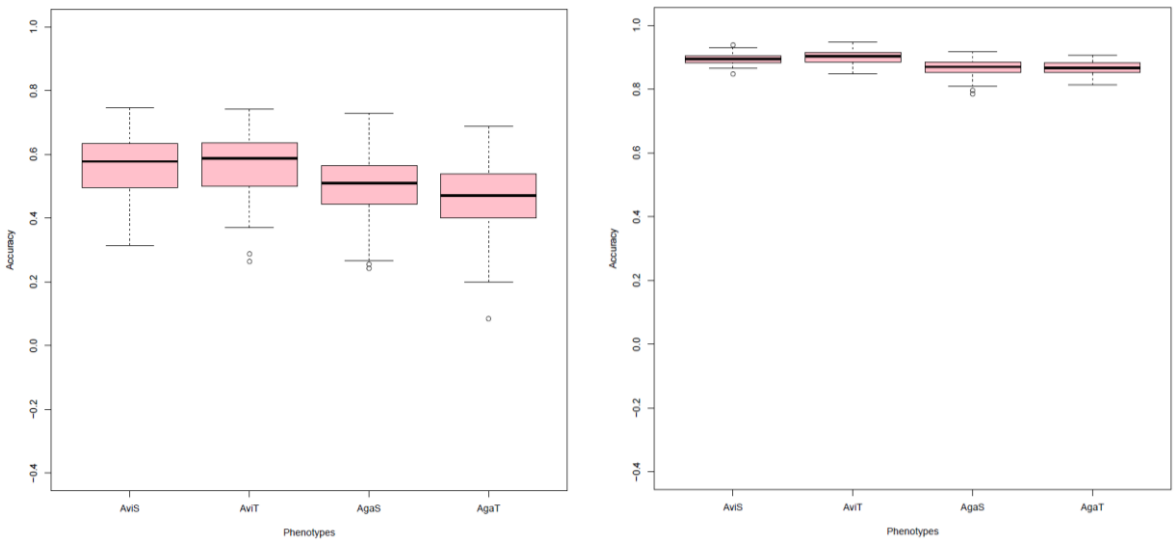


Figure 14 : Boxplot de l'*Accuracy* du caractère flw de la population GWAS obtenus à partir du modèle GxE dans CV1 et CV2. Prédictions faites avec 50000 itérations et 5000 burnIn.

b. Modèle simple environnement

Population MAGIC : La précision de la prédiction du caractère flw varie entre 0,27 (Avi17, PE=50%, modèle BL) et 0,54 (Mor15, PE=75%, modèle BL). En ce qui concerne le caractère fw, la précision de prédiction varie entre 0,14 (HSMor16, PE=50%, modèle BL) et 0,47 (Avi12, PE=75%, modèle BL ; Annexes 6 et 9). Enfin, le caractère leaf quant à lui a une précision de prédiction comprise entre 0,08 (Mor15, PE=50%, modèle BL) et 0,30 (HAvi17, PE=75%, modèle BL ; annexes 7 et 10). Le modèle BayesC prédit mieux les caractères les moins héritables par rapport aux modèle BL.

Population GWAS : Pour le caractère flw, la précision de prédiction varie entre 0,43 (AgaT, PE=50%, modèle BL) et 0,54 (AviS, PE=75%, modèle BL). Elle est comprise entre 0,65 (AgaS, PE=50%, modèle BL) et 0,74 (AviT, PE=75%, modèle BL) pour fw. Enfin, elle varie entre 0,32 (AgaT, PE=50%, modèle BL) et 0,53 (AviT, PE=75%, modèle BL). La prédiction évolue en fonction de l'héritabilité des caractères.

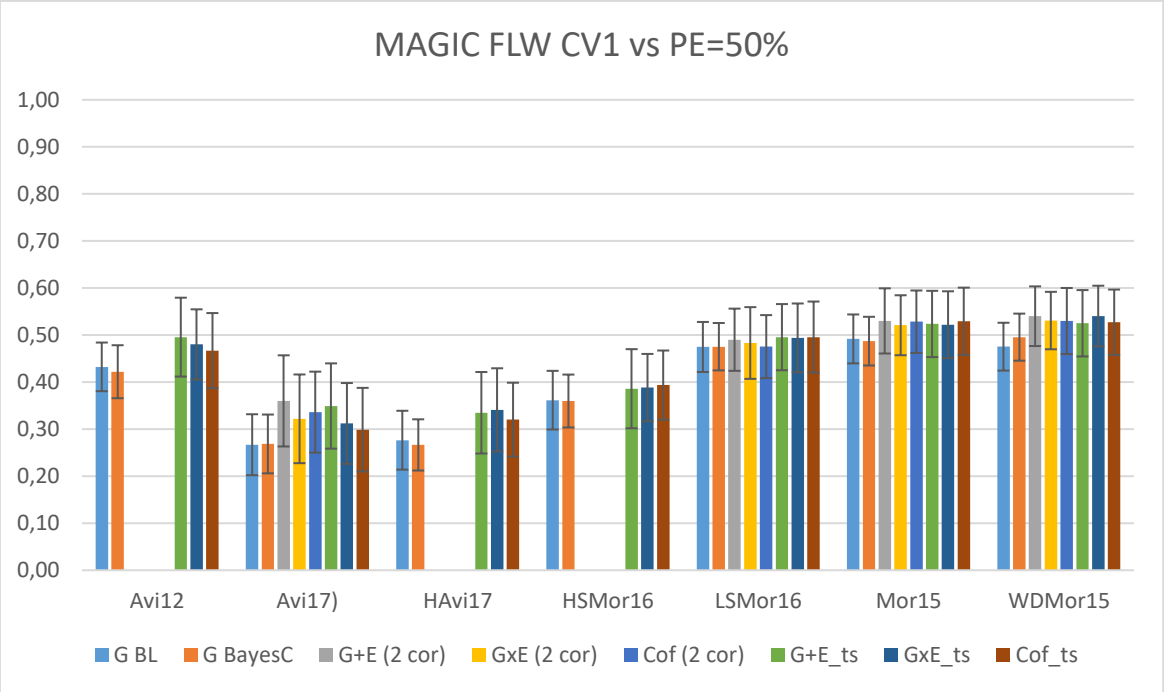
c. Modèle inter-environnement (G+E)

Cette approche est meilleure que la précédente quels que soit le caractère et la population. Globalement, la prédiction en CV2 est meilleure à celle faite en CV1. En CV1, elle prédit mieux les caractères les plus héritables et dont les environnements sont très corrélés entre eux par rapport à l'approche qui intègre les interactions GxE. En effet, les % change (GxE par rapport à G+E) sont négatifs pour les caractères flw et fw, tandis qu'ils sont positifs pour le caractère leaf, moins héritable et dont les environnements sont peu corrélés entre eux (Annexes 5 et 6). Aussi, les prédictions faites par paire d'environnements confirment que plus les environnements sont corrélés, plus leur précision de prédiction est élevée quelle que soit la population. C'est le cas pour les caractères tels que flw où Mor15 et WDMor15 sont très corrélés (cor=0,86) (figure 15), dont les %change sont négatifs

(Annexe 5, Magic) alors qu'ils sont positifs (Annexe 6, Magic) pour Havi17 et LSMor16 de fw qui sont très peu corrélés (cor=0,2). Enfin, cette approche prédit mieux en CV1 que l'approche de simple environnement quand PE=50%, tandis qu'elle prédit moins bien que l'approche simple environnement quand PE=75%. Ceci confirme l'hypothèse selon laquelle la précision de la prédiction diminue quand PE diminue, puisse qu'en CV1 la composition de la PE se fait comme en simple environnement.

Dans CV2, le modèle G+E prédit mieux qu'en CV1 quels que soit le caractère et la population. La précision de prédiction est meilleure (%change de 2 à 3% de plus) pour les environnements de Maroc avec un faible stress (Mor15, LSMor16) de flw par rapport au modèle GxE, pour les prédictions faites sur tous les environnements (figure 16, Annexe 8 Magic). Cependant, il prédit moins bien par rapport à GxE pour le même caractère dans la population GWAS quels que soient le site et les conditions de présence ou absence de stress (5 à 12% de moins, Annexe 8, GWAS). Pour les prédictions par paire d'environnements, la précision de prédiction est bonne avec des valeurs comprises entre 0,79 et 0,95 pour la GWAS et les environnements d'Agadir sont mieux prédits, car ils sont plus corrélés entre eux.

Aussi, ce modèle prédit mieux le caractère fw dans la population GWAS que le modèle GxE. En effet, pour tous les environnements, la précision de prédiction est supérieure de 2% à celle de GxE, à l'exception d'AgaS qui est mieux prédit par le modèle GxE, supérieure de 7%. Néanmoins les environnements du Maroc demeurent les mieux prédits par G+E pour GWAS et MAGIC, en prédiction par paire d'environnements, quelle que soit les corrélations entre ces environnements (Annexes 6 et 9). En ce qui concerne le caractère leaf, la précision de prédiction est inférieure à celle de flw et fw quelle que soit la population. Avec le modèle G+E, la précision de prédiction est largement inférieure à celle de GxE, 1 à 35% de différence suivant les environnements et populations (Annexes 7 et 10).



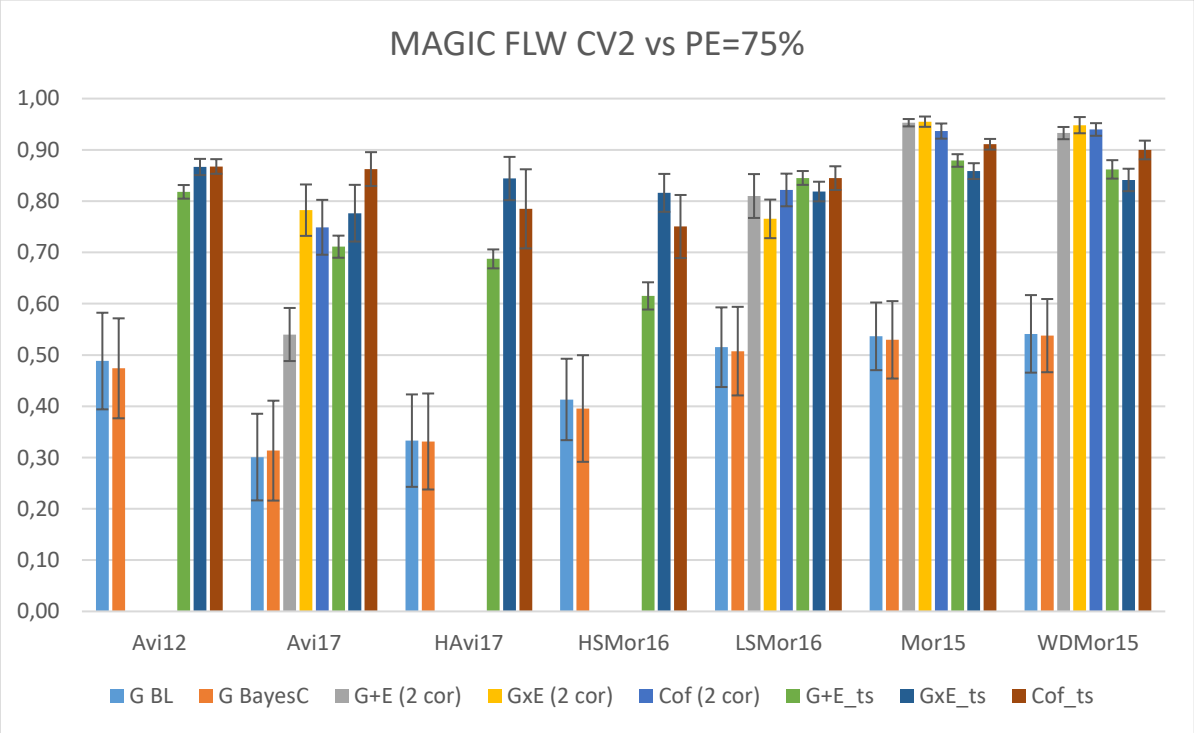


Figure 15 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles (G, G+E, GxE, Cofacteur) du caractère flw de la population Magic, CV1 et PE=50. La notation « 2 cor » correspond aux prédictions par paire d'environnements les plus corrélés. Les barres noires représentent les écartypes observés pour les différents modèles dans chaque environnement.

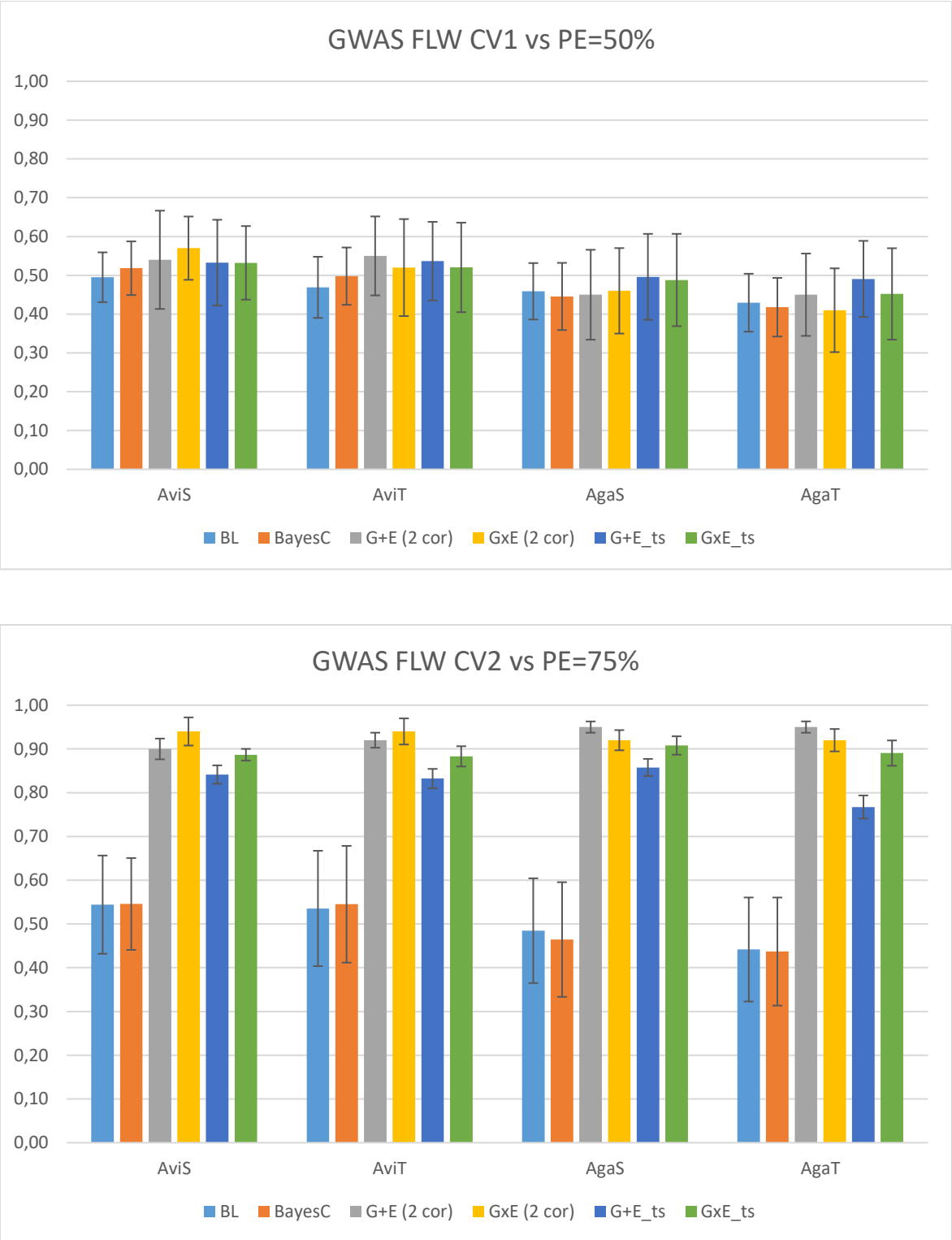


Figure 16 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles (G, G+E, GxE) du caractère flw de la population GWAS, CV1 et PE=50. La notation « 2 cor » correspond aux prédictions par paire d'environnements les plus corrélés. Les barres noires représentent les écartypes observés pour les différents modèles dans chaque environnement.

d. Modèle interaction (GxE)

Globalement, ce modèle améliore la précision de prédiction par rapport aux deux modèles précédents. En CV1, la précision de prédiction est améliorée par rapport au modèle simple environnement dans la population MAGIC pour les 3 caractères. Cependant, dans la population GWAS, il n'améliore que peu voire pas du tout la précision de prédiction par rapport au modèle simple environnement (Figure 16, Annexes 6, 7 Gwas). En comparaison avec le modèle G+E, il prédit mieux leaf par rapport à fw ou encore flw, avec un gain de précision de 2 à 6%.

Dans CV2, le modèle GxE améliore fortement la précision de prédiction par rapport au modèle simple environnement avec un gain de précision de 20 à 52% (PE=75%) pour les prédictions sur tous les environnements et jusqu'à 56% pour celle des environnements par paire. Plus les environnements sont corrélés, plus la précision de prédiction est augmentée. C'est le cas pour les différents caractères, en effet, les *accuracy* de Mor15 et WDmor15 très corrélés pour le caractère flw sont élevées (respectivement 0,86 et 0,84) tandis que celles de Avi17 et LSMor16 moins corrélés, sont plus faibles (respectivement 0,78, et 0,82). Le gain de précision avec GxE par rapport à G+E augmente chez les environnements stressés quand ils sont peu corrélés avec les autres. C'est le cas pour WDMor15 qui est mieux prédit (0,86 avec %change = -2%) pour flw par G+E quand il est très corrélé (cor = 0,86 avec Mor15) aux autres (Annexe 8, Magic), alors qu'il est mieux prédit par GxE (0,66, %change = 35%) pour leaf quand il est moins corrélé avec les autres (Annexe 10). Dans la population GWAS, la précision de prédiction de GxE par rapport à G+E augmente quand on passe de fw à flw puis à leaf, avec un % change qui varie de -2 à 7% pour fw, de 5 à 12% pour flw et de 5 à 34% pour leaf (Annexes 8, 9 et 10).

e. Comparaison des partitions CV1 et CV2

La précision de prédiction génomique en CV2 est meilleure que celle de CV1 quel que soit le modèle, le caractère et la population. En effet, la figure 17 montre que la moyenne par environnement de l'*accuracy* en CV2 suivant les caractères et les environnements est largement supérieure à celle de CV1.

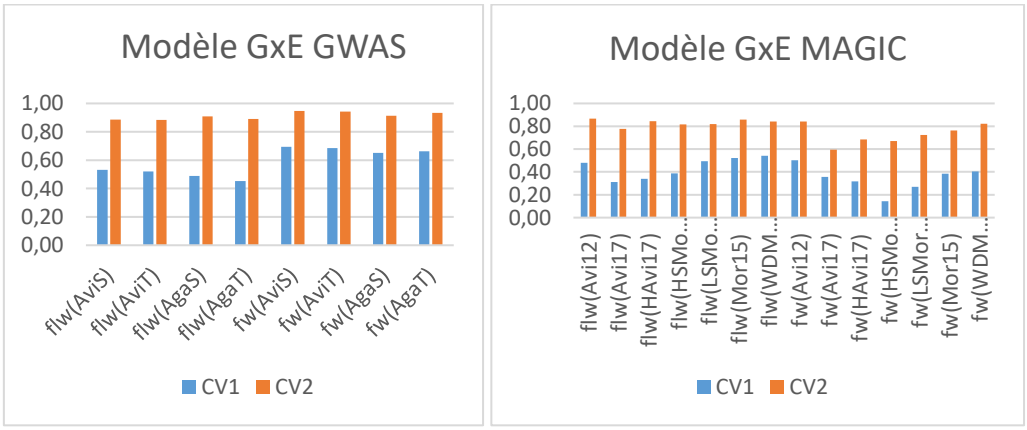


Figure 17: Comparaison de l'*accuracy* moyenne des caractères flw et fw des populations Gwas et

Magic obtenue avec le modèle GxE dans CV1 et CV2.

3.2.2 Intégration de cofacteur environnementaux

De même que le modèle GxE, le modèle avec les cofacteurs environnementaux prédit mieux en CV2 qu'en CV1 (figure 15). Quand on passe de CV1 à CV2, les environnements les moins bien prédits tels que Avi17 (flw, fw) et HAvi17 (flw, fw) par rapport aux modèle G+E, ont leur précision de prédiction qui devient nettement supérieure à celle des modèles G+E et GxE.

L'intégration des cofacteurs dans le modèle de prédiction GxE, entraîne globalement une amélioration de la précision de prédiction par rapport au modèle GxE dans la plupart des cas. Ce modèle prédit mieux le trait flw, dans les conditions de contrôle (Avi12, Avi17, Mor15), de faible stress salin (LSMor16) et de stress hydrique (WDMor15) par rapport à GxE. Il prédit moins bien les environnements de stress thermique et salin par rapport à GxE seul (figure 15, CV2). En ce qui concerne fw, ce modèle prédit mieux tous les environnements (toutes conditions confondus, CV2) par rapport aux autres modèles, à l'exception de LSMor16. Enfin, le modèle avec les cofacteurs, prédit mieux les caractères les plus héritables (Annexe 12).

4. DISCUSSION

En résumé, nous avons montré que la précision de prédiction diminue quand on passe d'une PE=75% à une PE=50% quels que soient le caractère et la population (figure 15 et 16 ; Annexe 6, 7 9 et 10). De même la précision de prédiction en CV2 est supérieure à celle observée avec CV1 quels que soit le caractère et la population. Globalement, les modèles G+E et GxE prédisent mieux que le modèle simple environnement, avec une *accuracy* maximale respectivement de 0,96 (AviS et AviT, GWAS, fw), 0,95 (AviS, GWAS, fw) et 0,74 (AviT, GWAS, fw) en CV2. Suivant la population, plus les environnements où les caractères sont mesurés sont corrélés, meilleure est la précision de prédiction de ces caractères. L'utilisation des Blups améliore la précision de prédiction des caractères de la population Magic (16% par rapport à la moyenne des *accuracy*, PE=75%, flw, annexe 8). Les modèles BL et BayesC ne sont pas très différents, leurs précisions de prédiction sont proches voire égales suivant les caractères et environnements (Figures 15 et 16). Cependant, pour la majorité des caractères et environnements, le modèle BL a une précision de prédiction 1 à 2% meilleure que celle de BayesC. Le modèle BL sera donc utilisé pour la comparaison de l'approche simple environnement aux autres.

La qualité du fruit de la tomate a fait l'objet d'études précédentes, ayant pour but d'évaluer l'impact des plusieurs facteurs de la sélection génomique sur la précision de la prédiction génomique. En effet, Picard, (2015) et Duangjit *et al.,* (2016) ont évalué l'impact de la taille de la population, du nombre et de la densité de marqueurs sur la précision de la prédiction génomique. Ces études ont révélé que quand la taille de PE augmente l'*accuracy* augmente également. Aussi, elles ont révélé que la réduction du nombre de marqueurs entraine la diminution de l'*accuracy*. De plus, peu de différences entre les modèles BL, BayesC et RKHS ont été observées. Enfin, l'*accuracy* est fortement liée à l'héritabilité des caractères. Plus le caractère est héritable plus la précision de prédiction est élevée. Toutes ces évaluations ont été faites sur le modèle simple environnement qui ne prend pas en compte les interactions GxE.

Les études sur les modèles multi-environnements ont montré que ces modèles permettent d'obtenir des résultats en général meilleurs que ceux du modèle simple environnement. En effet, Burgueño *et al.,* (2012) montrent que l'utilisation de modèles multi-environnement avec des facteurs qui intègrent l'interaction GxE permet d'améliorer la précision de la prédiction par rapport aux modèles simple environnement. De même, Lopez-Cruz *et al.,* (2015) montrent que les modèles multi-environnements, G+E et avec interaction GxE prédisent mieux que les modèles simple environnement. Nous avons donc comparé ces modèles.

4.1 Différence entre les populations MAGIC ET GWAS : Précision de prédiction

Les prédictions des traits dans la population GWAS sont meilleures que celles de la population MAGIC, pour les mêmes traits. Quels que soient le modèle ou l'approche de constitution de la

population d'entrainement (PE, CV1 et CV2), la précision de prédiction est plus élevée dans la population GWAS par rapport à MAGIC. Cette différence provient du fait que la population MAGIC est obtenue à partir de 8 parents (gros et petit fruits), il n'y a donc au maximum que 8 allèles qui ségrégent. Dans la population MAGIC, on étudie le contraste entre allèles de tomates à gros fruits et petit fruits, issus de la sélection récente des tomates. Tandis que dans la core-collection GWAS, qui est constitué de petit fruits, on explore plutôt le contraste des allèles issus de la domestication vs ceux du parent sauvage. Par ailleurs le nombre de marqueurs dans la population MAGIC est plus faible (1350) qu'en GWAS (7000).

4.2 Différence de prédiction entre traits : prédiction et héritabilité

Les résultats obtenus dans les différents modèles montrent que, globalement, les traits les plus héritables sont les mieux prédits en terme d'*accuracy*. Ces résultats sont observés pour les deux populations. L'héritabilité diminue quand on passe du caractère flw à fw puis à leaf pour la population Magic, tandis qu'elle diminue quand on passe de fw à flw puis leaf en ce qui concerne la population GWAS. La précision de la prédiction évolue dans le même sens que l'héritabilité pour les différentes populations. Nos résultats sont en adéquation avec les résultats de Duangjit *et al.,* (2016), qui ont respectivement montré que l'héritabilité et la précision de prédiction sont fortement corrélées (r=0,69). Cette corrélation est due au fait que les caractères héritables ont une grande proportion de variance expliquée par la variance génétique, ce qui les rend donc plus facile à prédire.

4.3 Différence entre les modèles simple environnement, G+E et GxE

Globalement, les modèles G+E et GxE prédisent mieux que le modèle simple environnement (Figures 15 et 16). Ces résultats sont en accord avec ceux de Burgueno (2012) et Lopez Cruz et al. (2015). Le modèle GxE prédit mieux que le modèle G+E dans certains environnements. Notamment lorsque l'analyse des composantes de la variance montre que la part de variance expliquée par le modèle qui intègre l'interaction GxE est supérieure à celle du modèle avec G+E (voir 3.2.3, valeur de R^2).

Les résultats de prédiction faites à partir de modèles multi-environnementaux de Burgueño *et al.,* (2012) et Lopez-Cruz *et al.,* (2015) vont dans le même sens que nos résultats. Les prédictions sont meilleures quand les environnements sont corrélés par paire par rapport à celles des faites sur tous les environnements.

4.4 Différence entre CV1 et CV2 : Précision de prédiction

Quand la taille de PE diminue l'*accuracy* diminue. Nos résultats sont en adéquation avec ceux de Picard, (2015) et Duangjit *et al.,* (2016). Pour le modèle simple environnement (PE=75%), la précision de prédiction est peu différente de celles des modèles G+E et GxE en CV1 (PE=70%).

L'utilisation des BLUPs améliore la précision de prédiction des caractères. La prise en compte de tous les environnements avec le BLUP, donne de meilleurs résultats de prédiction par rapport à aux prédictions faites dans chaque environnement.

Cependant, la précision de prédiction de CV2 est très supérieure à celle de CV1. La précision de prédiction est meilleure en CV2 par rapport à CV1 pour les modèles G+E et GxE. (Voir 3.2.1, résultats de prédiction). Quels que soit le modèle ou la population, CV2 permet de mieux prédire les traits dans les différents environnements par rapport à CV1. De plus, le modèle GxE est le meilleur pour prédire en CV2, cependant en CV1, G+E prédit mieux, même si la différence de prédiction entre les deux modèles n'est pas très grande (Figures 15 et 16, Annexes 5 à 10). Cette bonne précision de prédiction du modèle GxE en CV2 s'explique par le fait que les environnements soient corrélés entre eux. Pour un caractère dont les environnements sont peu corrélés entre eux, le modèles GxE les prédit mieux que le modèle G+E en CV2. Ce gain de prédiction avec GxE en CV2 provient de la capacité de ce modèle à emprunter les informations issues des différents individus. Cela n'est possible qu'avec CV2 et pas avec CV1 comme le souligne Lopez-Cruz *et al.,* (2015). La prédiction des individus ayant été phénotypés dans certains environnements et pas dans d'autres (CV2) est donc préférable à celle des individus n'ayant pas du tout été phénotypés (CV1). Cela peut orienter le design expérimental : si on doit tester plusieurs environnements, il vaut mieux un design de type CV2.

4.5 Limite du package BGLR et autres points à étudier

BGLR est un package très utilisé pour la prédiction génomique. Il existe néanmoins plusieurs limites à ce package, à savoir le temps d'exécution des taches et le mode de calcul de la variance de l'erreur. En effet, avec BGLR on considère que dans les modèles G+E et GxE, la variance de l'erreur est homogène entre les environnements. Le package BGGE est une alternative pour la résolution de ce problème, puisqu'il considère la variance différente dans les différents environnement et modèles (Lopez-Cruz *et al.,* 2015; Granato *et al.,* 2018).

Le temps d'exécution de taches des modèles augmente en fonction du nombre d'environnements et du modèle choisi. En effet, la durée d'exécution des modèles sur le serveur de l'INRA d'Avignon prend entre 2h et 4 jours, quand on passe du modèle simple environnement au modèle GxE puis au modèle avec les cofacteurs environnementaux. Ce temps d'exécution des modèles par BGLR est beaucoup plus long que celui de BGGE. Aussi, la variance de l'erreur avec BGLR est légèrement plus grande qu'avec BGGE. BGGE apparait donc comme une alternative pour répondre aux insuffisances de BGLR (Lopez-Cruz *et al.,* 2015; Granato *et al.,* 2018a). J'ai commencé à l'utiliser mais n'ai pas eu le temps de synthétiser les résultats.

Il conviendrait également de prendre en compte un certain nombre de points qui pourraient potentiellement améliorer la prédiction, à la suite de notre étude. IL s'agit d'abord d'optimiser la population d'entraînement à l'aide de la méthode CD*means*, utilisée par Picard, (2015). Elle permet

d’améliorer la prédiction par rapport au choix de la PE de manière aléatoire. Aussi, l’imputation des marqueurs permet d’améliorer la précision de la prédiction. Cependant, quand le nombre de marqueurs est très grand, cela peut entraîner une diminution de l’*accuracy*. Il semble que l’analyse d’haplotypes permet de réduire ce biais. Par ailleurs on pourrait envisager l’utilisation des informations des QTLs détectés dans les études précédentes (pris en cofacteurs), et évaluer leur impact sur la précision de prédiction.

CONCLUSION ET PERSPECTIVES

En conclusion, cette étude a permis d'évaluer différentes stratégies de prédiction génomique applicables à la tomate. Elle a servi à comparer ces stratégies les unes par rapport aux autres, en vue d'apprécier leur impact sur la précision de prédiction. En tout, trois modèles différents ont été utilisés pour prédire les caractères dans deux populations qui ont été évaluées dans différents environnements.

Globalement, la prédiction à l'aide d'approche multi-environnements est meilleure que celle de l'approche mono-environnement, notamment en CV2. Les prédictions réalisées ont révélé que la prise en compte de l'interaction GxE, pour la prédiction de caractères phénotypés dans des environnements plus ou moins stressés, a une meilleure précision de prédiction suivant les environnements par rapport aux modèles inter-environnements et simple environnement. L'intégration des cofacteurs environnementaux de la population MAGIC dans le modèle GxE, améliore également légèrement la précision de la prédiction.

Aussi, la précision de prédiction est corrélée positivement à l'héritabilité. De plus, les environnements les plus corrélés sont les mieux prédits. Le modèle GxE améliore la précision de prédiction des caractères les moins corrélés par rapport au modèle G+E et simple environnement.

La composition de la population d'entraînement a un impact important sur la précision de prédiction. En effet, l'utilisation du schéma CV1 prédit moins bien que quand on utilise le schéma CV2. Le schéma CV2, s'avère être la meilleure option de composition de la population d'entraînement. Il est donc plus efficace de prédire les individus qui ont été phénotypés dans certains environnements et pas dans d'autres que des individus qui n'ont pas du tout été phénotypés.

A l'issue de cette étude, qui a montré que la prédiction génomique est un atout pour la sélection de la tomate, il serait intéressant d'étudier :

- A quel moment mettre en place la prédiction génomique dans un schéma de sélection ?
- Comment utilise la prédiction génomique pour prédire les croisements à réaliser pendant la sélection ?
- L'impact de la sélection génomique sur la diversité génétique dans la population de sélection, afin d'éviter la perte de la diversité génétique au fil des générations.

ORGANISATION DE L'ÉTUDE

Les trois premiers mois du stage se sont déroulés pendant la période du confinement. Pendant cette période, la revue bibliographique a été faite, suivie de la mise en forme des données. Ces données ont été utilisées dans différentes études auparavant. J'ai donc dû les mettre au format qui correspondait à nos analyses.

Pendant, cette période j'ai élaboré les scripts pour l'analyse descriptive des données et adapté de scripts issus d'études précédentes à mes données et à mes objectifs de prédiction. Le modèle simple environnement (modèle BL et BayesC) a été mis en œuvre pour 6 caractères de la population MAGIC. Cependant, vu que l'exécution des modèles prend beaucoup de temps, j'ai décidé de faire ces prédictions avec 5 cycles puisque je travaillais avec mon ordinateur personnel. Dans le même temps, j'ai fait la rédaction du matériel et méthodes en ce qui concerne le modèle simple environnement.

Après le confinement, les 3 derniers mois du stage se sont déroulés au sein de l'INRAE d'Avignon (au GAFL). Pendant, cette période, j'ai terminé la rédaction des méthodes et fait l'exécution des prédictions avec le modèle simple environnement. Dans le même temps, les scripts des modèles inter-environnements (GxE), interactions GxE et avec les cofacteurs ont été mis en œuvre. Il s'en est suivi l'exécution des scripts sur le serveur de L'INRAE, puis le formatage des résultats et la rédaction des autres parties du rapport.

Je devais participer à une expérimentation d'analyse du panel GWAS dans de nouvelles conditions, mais en raison du COVID 19, elle a été déplacée pour cet automne, j'ai juste participé aux semis avec certains membres de l'équipe.

REFERENCES BIBLIOGRAPHIQUES

Albert E, Gricourt J, Bertin N, Bonnefoi J, Pateyron S, Tamby J-P, Bitton F, Causse M. **2016a**. Genotype by watering regime interaction in cultivated tomato: lessons from linkage mapping and gene expression. *Theoretical and Applied Genetics* **129**: 395–418.

Albert E, Segura V, Gricourt J, Bonnefoi J, Derivot L, Causse M. **2016b**. Association mapping reveals the genetic architecture of tomato response to water deficit: focus on major fruit quality traits. *Journal of Experimental Botany* **67**: 6413–6430.

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L. **2011**. Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. *The Plant Genome* **4**: 132–144.

Bastien C, Cros D, This P. **2016**. Quelle place pour la sélection génomique chez les plantes pérennes ? *Sélection génomique : théorie et mise en oeuvre en relation avec les programmes d'amélioration*.

Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Díez MJ, Francis D, Causse M, van der Knaap E, Cañizares J. **2015**. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics* **16**: 257.

Burgueño J, de los Campos G, Weigel K, Crossa J. **2012**. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. *Crop Science* **52**: 707–719.

Burgueño J, Crossa J, Cotes JM, Vicente FS, Das B. **2011**. Prediction Assessment of Linear Mixed Models for Multienvironment Trials. *Crop Science* **51**: 944–954.

Causse M, Desplat N, Pascual L, Le Paslier M-C, Sauvage C, Bauchet G, Bérard A, Bounon R, Tchoumakov M, Brunel D, *et al.* 2013. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC genomics* **14**: 791.

Charmet G, Tran LG, Auzanneau J, Rincent R, Bouchet S. **2019**. *BWGS: a R package for genomic selection and its application to a wheat breeding programme*. Genetics.

Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, *et al.* 2015a. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theoretical and Applied Genetics* **128**: 397–410.

Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A, Pomiès V, Riou V, *et al.* 2015b. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theoretical and Applied Genetics* **128**: 397–410.

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, *et al.* 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science* **22**: 961–975.

Denis M. **2016**. Les logiciels pour la sélection génomique. *Sélection génomique : théorie et*

mise en oeuvre en relation avec les programmes d'amélioration. https://agritrop.cirad.fr/582578/

Diouf IA, Derivot L, Bitton F, Pascual L, Causse M. **2018a**. Water Deficit and Salinity Stress Reveal Many Specific QTL for Plant Growth and Fruit Quality Traits in Tomato. *Frontiers in Plant Science* **9**.

Diouf IA, Derivot L, Bitton F, Pascual L, Causse M. **2018b**. Water Deficit and Salinity Stress Reveal Many Specific QTL for Plant Growth and Fruit Quality Traits in Tomato. *Frontiers in Plant Science* **9**: 279.

Diouf I, Derivot L, Koussevitzky S, Carretero Y, Bitton F, Moreau L, Causse M. **2020**. Genetic basis of phenotypic plasticity and genotype x environment interaction in a multi-parental population. *bioRxiv*: 2020.02.07.938456.

Duangjit J, Causse M, Sauvage C. **2016a**. Efficiency of genomic selection for tomato fruit quality. *Molecular Breeding* **36**: 29.

Duangjit J, Causse M, Sauvage C. **2016b**. Efficiency of genomic selection for tomato fruit quality. *Molecular Breeding* **36**: 29.

Endelman J. **2011**. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* **4**: 250–255.

Evanno G, Regnaut S, Goudet J. **2005**. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.

Fodor A, Segura V, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P, Homa FAA, Lacombe T, This P, Le Cunff L. **2014**. Genome-Wide Prediction Methods in Highly Diverse and Heterozygous Species: Proof-of-Concept through Simulation in Grapevine (R Khanin, Ed.). *PLoS ONE* **9**: e110436.

Giovannoni JJ. **2004**. Genetic Regulation of Fruit Development and Ripening. *The Plant Cell* **16**: S170–S180.

Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile J-C, Rocher S, Daugrois J, Toubi L, *et al.* 2013. Experimental assessment of the accuracy of genomic selection in sugarcane. *Theoretical and Applied Genetics* **126**: 2575–2586.

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, Burgueño J, Fritsche-Neto R. **2018a**. BGGE: A New Package for Genomic-Enabled Prediction Incorporating Genotype × Environment Interaction Models. *G3: Genes|Genomes|Genetics* **8**: 3039–3047.

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, Burgueño J, Fritsche-Neto R. **2018b**. BGGE: A New Package for Genomic-Enabled Prediction Incorporating Genotype × Environment Interaction Models. *G3: Genes|Genomes|Genetics* **8**: 3039–3047.

Grassely D, CTIFL Centre Technique Interprofessionnel des Fruits et Légumes P (FRA), Navez B, Letard M. **2000**. *Tomate : pour un produit de qualité*. Paris: CTIFL Centre Technique Interprofessionnel des Fruits et Légumes.

Heslot N, Jannink J-L, Sorrells M. **2013**. Using Genomic Prediction to Characterize Environments and Optimize Prediction Accuracy in Applied Breeding Data. *Crop Science* **53**: 921.

Hickey JM, Chiurugwi T, Mackay I, Powell W. **2017**. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics* **49**: 1297–1303.

Lan S, Zheng C, Hauck K, McCausland M, Duguid SD, Booker HM, Cloutier S, You FM. **2020**. Genomic Prediction Accuracy of Seven Breeding Selection Traits Improved by QTL Identification in Flax. *International Journal of Molecular Sciences* **21**: 1577.

Legarra A, Christensen OF, Aguilar I, Misztal I. **2014**. Single Step, a general approach for genomic selection. *Livestock Science* **166**: 54–65.

Liabeuf D, Sim S-C, Francis DM. **2017**. Comparison of Marker-Based Genomic Estimated Breeding Values and Phenotypic Evaluation for Selection of Bacterial Spot Resistance in Tomato. *Phytopathology*^{TM} **108**: 392–401.

Liu X, Wang H, Hu X, Li K, Liu Z, Wu Y, Huang C. **2019**. Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. *Frontiers in Plant Science* **10**: 1129.

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, Singh RP, Autrique E, de los Campos G. **2015**. Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker × Environment Interaction Genomic Selection Model. *G3: Genes|Genomes|Genetics* **5**: 569–582.

Mattoo A, Razdan MK. **2007**. *Genetic improvement of solanaceous crops. Volume 2. Volume 2*. Enfield, NH: Science Publishers.

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, Eeuwijk F van, Tardieu F. **2019a**. Genomic prediction of maize yield across European environmental conditions. *Nature Genetics* **51**: 952–956.

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, Eeuwijk F van, Tardieu F. **2019b**. Genomic prediction of maize yield across European environmental conditions. *Nature Genetics* **51**: 952–956.

Oakey H, Cullis B, Thompson R, Comadran J, Halpin C, Waugh R. **2016**. Genomic Selection in Multi-environment Crop Trials. *G3: Genes|Genomes|Genetics* **6**: 1313–1326.

Okeke UG, Akdemir D, Rabbi I, Kulakow P, Jannink J-L. **2017**. Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genetics Selection Evolution* **49**: 88.

Owens BF, Lipka AE, Magallanes-Lundback M, Tiede T, Diepenbrock CH, Kandianis CB, Kim E, Cepela J, Mateos-Hernandez M, Buell CR, *et al.* 2014. A Foundation for Provitamin A Biofortification of Maize: Genome-Wide Association and Genomic Prediction Models of Carotenoid Levels. *Genetics* **198**: 1699–1716.

Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet J-P, Le QH, Chauchard B, Verschave P, Causse M. **2015a**. Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnology Journal* **13**: 565–577.

Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet J-P, Le QH, Chauchard B, Verschave P, Causse M. **2015b**. Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnology Journal* **13**: 565–577.

Pérez P, de los Campos G. **2014a**. Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198**: 483–495.

Pérez P, de los Campos G. **2014b**. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**: 483–495.

Pérez P, de los Campos G. **2014c**. Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198**: 483–495.

Picard C. (2015). Évaluation d'une stratégie de sélection génomique dans 3 dispositifs expérimentaux chez la tomate. Ingénieur en Horticulture, APIMET. Montpellier SupAgro : 103.

Ranjan A, Ichihashi Y, Sinha NR. **2012a**. The tomato genome: implications for plant breeding, genomics and evolution. *Genome Biology* **13**: 167.

Ranjan A, Ichihashi Y, Sinha NR. **2012b**. The tomato genome: implications for plant breeding, genomics and evolution. *Genome Biology* **13**: 167.

Robert-Granié C, Legarra A, Ducrocq V. **2011**. Principes de base de la sélection génomique. *INRAE Productions Animales* **24**: 331–340.

Rothan C, Diouf I, Causse M. **2019a**. Trait discovery and editing in tomato. *The Plant Journal* **97**: 73–90.

Rothan C, Diouf I, Causse M. **2019b**. Trait discovery and editing in tomato. *The Plant*

Storlie E, Charmet G. **2013**. Genomic Selection Accuracy using Historical Data Generated in a Wheat Breeding Program. *The Plant Genome* **6**: plantgenome2013.01.0001.

The tomato genome sequence provides insights into fleshy fruit evolution. **2012**. *Nature* **485**: 635–641.

Tomato Genome Consortium. **2012**. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**: 635–641.

Yamamoto E, Matsunaga H, Onogi A, Kajiya-Kanegae H, Minamikawa M, Suzuki A, Shirasawa K, Hirakawa H, Nunome T, Yamaguchi H, *et al.* 2016. A simulation-based breeding design that uses whole-genome prediction in tomato. *Scientific Reports* **6**: 1–11.

SITOGRAPHIE

« Agreste Agriculture, production de tomate en France en 2019 », consulté le 30 mars 2020 https://agreste.agriculture.gouv.fr/agreste-web/download/publication/publie/IraLeg071/2019_071inforaptomate.pdf

FAO stat, Production de tomate dans le monde et en France, consulté le 01 et 09 Avril 2020 http://www.fao.org/faostat/fr/#data/QC

LISTE DES ANNEXES

ANNEXES

Annexe 1 : Packages et modèles implémentés utilisés sur le logiciel R en SelGen (Charmet *et al.,* 2019; Granato *et al.,* 2018; Andres Legarra *et al.,* 2014; Pérez & de los Campos, 2014; Robert-Granié *et al.,* 2011).

Intitulé	Package/ Modèle	Fonction
getK	Modèle	Création de noyaux multi-environnements ou de matrices de covariance connues utilisable dans la fonction BGGE pour s'adapter au modèle.
P-BLUP	Modèle	Obtention de BLUP (best linear unbiased prediction) à partir de matrice d'apparentement issue de pédigrée
G-BLUP	Modèle	Obtention de BLUP à partir de matrice d'apparentement issue de marqueurs
RR-BLUP	Modèle	Utilisation des marqueurs SNP comme effets fixes
Bayesian ridge regression	Modèle	Modèle bayésien ; Prise en compte de tous les marqueurs (même ceux avec de faibles effets)
Bayesian Lasso regression	Modèle	Ajout d'une contrainte ou pénalisation à la méthode des moindres carrés ; Sélection de variable et Régularisation du simulateur par cette pénalisation.
BGGE	Package	Mise en œuvre des prédictions génomiques par le biais d'un modèle linéaire mixte pour des variables continues ; Prise en compte de l'interaction GxE. Il dérive du package BGLR.
BGLR	Package	Mise en œuvre de vaste collection de modèles de régression bayésiens, des méthodes de sélection et de retrait de variables paramétriques et des procédures semi-paramétriques ; Prise en charge de traits continus ainsi que les traits binaires et ordinaux.
BWGS	Package	Facilite le calcul de GEBV ; Exécution de validations croisées aléatoires répliquées dans un ensemble de lignées génotypées et phénotypées d'entraînement ; Prédiction de la GEBV, pour un ensemble de lignées génotypées uniquement.
Synbreed	Package	Intégration des effets aléatoires et fixes pour les BLUP ; Calcul de plusieurs types de matrices d'apparentement (marqueurs et pédigrée) ; Cross-validation.
rr-BLUP	Package	Mise en œuvre de méthodes semi-paramétriques ; Intégration de plusieurs effets fixes mais sur un seul effet aléatoire.
BLR	Package	Intégration des effets fixes, effet polygénique et effets des marqueurs ; Variances hétérogènes.

Annexe 2 : Valeurs estimées des variances résiduelle et génétique par modèle et environnement pour le caractère flw dans la population MAGIC (orange) et GWAS (bleu). Le R^2 est estimé à partir du ratio de la variance (effet génétique + interaction GxE) par rapport au total de la variance (résiduelle + effet génétique+ interaction GxE).

Modèles/environnements			Résiduelle		Effet génétique		Interaction GXE	R^2	
Simple environnement									
Avi12			0,44		0,27		-	38%	
Avi17			0,62		0,22		-	26%	
HAvi17			0,61		0,22		-	26%	
HSMor16			0,51		0,25		-	33%	
LSMor16			0,40		0,29		-	42%	
Mor15			0,36		0,29		-	44%	
WDMor15			0,32		0,33		-	51%	
			G+E	**GxE**	**G+E**	**GxE**	**GxE**	**G+E**	**GxE**
Paire d'environnements									
Avi17			0,66	0,57	0,15	0,11	0,08	19%	25%
HSMor16	cor	0,43					0,08		25%
Mor15			0,16	0,14	0,43	0,44	0,03	73%	78%
WDMor15	cor	0,86					0,03		78%
Tous les environnements									
Avi12			0,45	0,32	0,27	0,30	0,02	37%	50%
Avi17							0,15		58%
HAvi17							0,14		58%
HSMor16							0,09		55%
LSMor16							0,03		51%
Mor15							0,02		50%
WDMor15							0,02		50%

Modèles/environnements			Résiduelle		Effet génétique		Interaction GXE	R^2	
Simple environnement									
AviS			0,37		0,44		-	55%	
AviT			0,32		0,48		-	60%	
AgaS			0,41		0,40		-	50%	
AgaT			0,42		0,42		-	50%	
			G+E	**GxE**	**G+E**	**GxE**	**GxE**	**G+E**	**GxE**
Paire d'environnements									
AviS			0,22	0,18	0,62	0,63	0,06	74%	79%
AviT	cor	0,83					0,05		79%
AviS			0,41	0,35	0,37	0,37	0,08	48%	56%
AgaT	cor	0,60					0,08		56%
AgaS			0,17	0,16	0,78	0,79	0,04	82%	84%
AgaT	cor	0,88					0,05		84%
Tous les environnements									
AviS			0,30	0,25	0,53	0,53	0,06	64%	70%
AviT							0,06		70%
AgaS							0,04		69%
AgaT							0,05		70%

Annexe 3 : Valeurs estimées des variances résiduelle et génétique par modèle et environnement pour le caractère fw dans la population MAGIC (orange) et GWAS (bleu). Le R^2 est estimé à partir du ratio de la variance (effet génétique + interaction GxE) par rapport au total de la variance (résiduelle + effet génétique+ interaction GxE).

Modèles/environnements	Résiduelle	Effet génétique	Interaction GXE	R^2
Simple environnement (G)				
Avi12	0,34	0,33	-	49%
Avi17	0,59	0,22	-	27%
HAvi17	0,61	0,22	-	27%
HSMor16	0,71	0,18	-	20%
LSMor16	0,70	0,18	-	20%
Mor15	0,52	0,26	-	33%
WDMor15	0,50	0,28	-	36%

	G+E	GxE	G+E	GxE	GxE	G+E	GxE
Paire d'environnements							
HAvi17	0,80	0,70	0,10	0,08	0,09	11%	20%
LSMor16 cor 0,20					0,07		17%
Mor15	0,22	0,19	0,49	0,50	0,03	69%	73%
WDMor15 cor 0,82					0,03		73%
Tous les environnements							
Avi12	0,64	0,50	0,19	0,21	0,03	23%	32%
Avi17					0,07		35%
HAvi17					0,10		38%
HSMor16					0,16		42%
LSMor16					0,06		35%
Mor15					0,02		32%
WDMor15					0,02		31%

Modèles/environnements	Résiduelle	Effet génétique	Interaction GXE	R^2
Simple environnement (G)				
AviS	0,17	0,43	-	72%
AviT	0,19	0,43	-	69%
AgaS	0,25	0,42	-	62%
AgaT	0,18	0,47	-	72%

	G+E	GxE	G+E	GxE	GxE	G+E	GxE
Paire d'environnements							
AviS	10%	0,09	0,47	0,46	0,04	82%	85%
AviT cor 0,93					0,04		85%
AgaS	16%	0,12	0,49	0,47	0,05	75%	81%
AgaT cor 0,87					0,06		81%
Tous les environnements							
AviS	0,15	0,10	0,46	0,46	0,04	75%	84%
AviT					0,03		83%
AgaS					0,09		85%
AgaT					0,05		84%

Annexe 4 : Valeurs estimées des variances résiduelle et génétique par modèle et environnement pour le caractère Leaf dans la population MAGIC (orange) et GWAS (bleu). Le R^2 est estimé à partir du ratio de la variance (effet génétique + interaction GxE) par rapport au total de la variance (résiduelle + effet génétique+ interaction GxE).

Modèles/environnements	Résiduelle	Effet génétique	Interaction GXE	R^2	
Simple environnement (G)					
Avi17	0,62	0,21	-	26%	
HAvi17	0,54	0,26	-	32%	
HSMor16	0,66	0,19	-	22%	
LSMor16	0,66	0,21	-	24%	
Mor15	0,75	0,17	-	19%	
WDMor15	0,66	0,20		24%	
	G+E	**GxE**	**G+E**	**GxE**	**GxE**	**G+E**	**GxE**
Paire d'environnements							
Avi17	0,59	0,46	0,22	0,23	0,07	27%	39%
HAvi17 cor 0,53					0,09		41%
HSMor16	0,61	0,54	0,20	0,21	0,06	25%	33%
LSMor16 cor 0,45					0,05		32%
HSMor16	0,87	0,73	0,07	0,04	0,09	7%	16%
WDMor15 cor 0,04					0,10		16%
Tous les environnements							
Avi17	0,80	0,69	0,10	0,10	0,06	11%	19%
HAvi17					0,08		21%
HSMor16					0,05		18%
LSMor16					0,05		18%
Mor15					0,05		18%
WDMor15					0,07		20%

Modèles/environnements	Résiduelle	Effet génétique	Interaction GXE	R^2	
Simple environnement (G)					
AviS	0,57	0,33	-	37%	
AviT	0,36	0,45	-	56%	
AgaS	0,72	0,26	-	26%	
AgaT	0,67	0,25	-	27%	
	G+E	**GxE**	**G+E**	**GxE**	**GxE**	**G+E**	**GxE**
Paire d'environnements							
AviS	0,48	0,37	0,34	0,37	0,11	42%	57%
AviT cor 0,55					0,08		55%
AviT	0,74	0,58	0,14	0,10	0,17	16%	31%
AgaS cor 0,25					0,12		27%
Tous les environnements							
AviS	0,64	0,51	0,25	0,29	0,06	28%	41%
AviT					0,08		42%
AgaS					0,14		46%
AgaT					0,08		42%

Annexe 5 : Précision de prédiction (*accuracy*) moyenne par environnement suivant différents modèles du caractère flw de la population MAGIC, CV1 et PE=50%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions ; par paire d'environnement et pour tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements				*Accuracy*			
Simple environnement		**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%Change**	
Avi12		0,43		0,42		1%	
Avi17		0,27		0,27		0%	
HAvi17		0,28		0,27		1%	
HSMor16		0,36		0,36		0%	
LSMor16		0,47		0,48		0%	
Mor15		0,49		0,49		0%	
WDMor15		0,48	0,57	0,50	0,57	-2%	PE=50
		G+E	**GxE**			**GxE**	
Paire d'environnements						**Simp_env**	**G+E**
Avi17	Cor	0,36	0,32			6%	-4%
LSMor16	0,59	0,49	0,48			1%	-1%
Mor15	Cor	0,53	0,52			3%	-1%
WDMor15	0,86	0,54	0,53			6%	-1%
Tous les environnements							
Avi12		0,50	0,48			5%	-2%
Avi17		0,35	0,31			5%	-4%
HAvi17		0,33	0,34			6%	1%
HSMor16		0,39	0,39			3%	0%
LSMor16		0,50	0,49			2%	0%
Mor15		0,52	0,52			3%	0%
WDMor15		0,52	0,54			7%	2%	CV1

Modèles/environnements				*Accuracy*			
Simple environnement		**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%Change**	
AviS		0,50		0,52		-2%	
AviT		0,47		0,50		-3%	
AgaS		0,46		0,45		1%	
AgaT		0,43	0,54	0,42	0,55	1%	
		G+E	**GxE**			**GxE**	PE=50%
Paire d'environnements						**Simp_env**	**G+E**
AviS	cor	0,54	0,57			7%	2%
AviT	0,83	0,55	0,52			5%	-4%
AviS	cor	0,53	0,53			3%	0%
AgaT	0,60	0,47	0,46			3%	-1%
AgaS	cor	0,45	0,46			0%	2%
AgaT	0,88	0,45	0,41			-2%	-5%	CV1
Tous les environnements							
AviS		0,53	0,53			4%	0%
AviT		0,54	0,52			5%	-2%
AgaS		0,50	0,49			3%	-1%
AgaT		0,49	0,45			2%	-4%

Annexe 6 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles du caractère fw de la population Magic (orange) et Gwas (bleu), CV1 et PE=50%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions par paire d'environnement et les tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements					***Accuracy***			
Simple environnement	**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%change**			
Avi12	0,43		0,44		-1%			
Avi17	0,25		0,26		-1%			
HAvi17	0,21		0,22		-1%			
HSMor16	0,14		0,14		0%			
LSMor16	0,19		0,17		2%			
Mor15	0,31		0,32		-1%			
WDMor15	0,33	0,45	0,34	0,46	-1%			PE=50%
	G+E	**GxE**			**GxE**			
Paire d'environnements					**Simp_env**	**G+E**		
HAvi17	0,29	0,30			8%	1%		
LSMor16 cor 0,2	0,23	0,23			4%	0%		
Mor15	0,38	0,37			6%	-2%		
WDMor15 cor 0,82	0,39	0,38			5%	-1%		
Tous les environnements								
Avi12	0,47	0,50			7%	4%		
Avi17	0,36	0,36			11%	-1%		
HAvi17	0,28	0,32			10%	4%		
HSMor16	0,14	0,14			0%	0%		
LSMor16	0,28	0,27			8%	-1%		
Mor15	0,37	0,38			7%	1%		
WDMor15	0,39	0,40			7%	2%		CV1

Modèles/environnements					***Accuracy***			
Simple environnement	**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%change**			
AviS	0,69		0,68		1%			
AviT	0,70		0,70		1%			
AgaS	0,65		0,64		1%			
AgaT	0,67	0,73	0,66	0,72	1%			PE=50%
	G+E	**GxE**			**GxE**			
Paire d'environnements					**Simp_env**	**G+E**		
AviS	0,69	0,69			0%	1%		
AviT cor 0,93	0,71	0,71			1%	0%		
AgaS	0,67	0,67			2%	0%		
AgaT cor 0,87	0,68	0,66			0%	-2%		
Tous les environnements								
AviS	0,68	0,70			1%	1%		
AviT	0,70	0,69			-2%	-2%		
AgaS	0,67	0,65			0%	-2%		
AgaT	0,67	0,66			0%	0%		CV1

Annexe 7 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles du caractère leaf de la population Magic (orange) et Gwas (bleu), CV1 et PE=50%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions par paire d'environnement et les tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements				***Accuracy***		
Simple environnement		**BL**		**BayesC**	**%change**	
Avi17		0,22		0,22	0%	
HAvi17		0,24		0,25	-1%	
HSMor16		0,21		0,21	0%	
LSMor16		0,14		0,14	0%	
Mor15		0,08		0,07	1%	
WDMor15		0,21		0,21	0%	PE=50%
		G+E	**GxE**		**GxE**	
Paire d'environnements					**Simp_env**	**G+E**
Avi17		0,26	0,28		7%	2%
HAvi17	53%	0,27	0,29		6%	2%
HSMor16		0,23	0,26		5%	4%
WDMor15	4%	0,22	0,24		3%	2%
Tous les environnements						
Avi17		0,25	0,30		8%	5%
HAvi17		0,22	0,24		0%	2%
HSMor16		0,20	0,26		5%	6%
LSMor16		0,23	0,23		9%	0%
Mor15		0,14	0,14		6%	0%
WDMor15		0,18	0,22		1%	4% CV1

Modèles/environnements				***Accuracy***		
Simple environnement		**BL**		**BayesC**	**%change**	
AviS		0,43		0,43	0%	
AviT		0,49		0,48	1%	
AgaS		0,34		0,32	2%	
AgaT		0,32		0,31	1%	
		G+E	**GxE**		**GxE**	PE=50%
Paire d'environnements					**Simp_env**	**G+E**
AviS		0,44	0,41		-2%	-3%
AviT	cor 0,55	0,48	0,49		-1%	1%
AviT		0,42	0,49		0%	8%
AgaS	cor 0,25	0,30	0,34		0%	4%
Tous les environnements						
AviS		0,40	0,42		-1%	2%
AviT		0,43	0,47		-3%	4%
AgaS		0,28	0,32		-2%	5%
AgaT		0,31	0,33		1%	2% CV1

Annexe 8 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles du caractère flw de la population Magic, CV2 et PE=75%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions par paire d'environnement et les tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements				***Accuracy***			
Simple environnement		**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%Change**	
Avi12		0,49		0,47		1%	
Avi17		0,30		0,31		-1%	
HAvi17		0,33		0,33		0%	
HSMor16		0,41		0,40		2%	
LSMor16		0,52		0,51		1%	
Mor15		0,54		0,53		1%	
WDMor15		0,54	0,61	0,54	0,61	0%	PE=75%
		G+E	**GxE**			**GXE**	
Paire d'environnements						**Simp_env**	**G+E**
Avi17		0,54	0,78			48%	24%
LSMor16	cor 0,59	0,81	0,77			25%	-4%
Mor15		0,95	0,95			42%	0%
WDMor15	cor 0,86	0,93	0,95			41%	2%
Tous les environnements							
Avi12		0,82	0,87			38%	5%
Avi17		0,71	0,78			48%	7%
HAvi17		0,69	0,84			51%	16%
HSMor16		0,62	0,82			40%	20%
LSMor16		0,85	0,82			30%	-3%
Mor15		0,88	0,86			32%	-2%
WDMor15		0,86	0,84			30%	-2% CV2

Modèles/environnements				***Accuracy***			
Simple environnement		**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%Change**	
AviS		0,54		0,55		0%	
AviT		0,54		0,55		-1%	
AgaS		0,48		0,46		2%	
AgaT		0,44	0,58	0,44	0,59	0%	
		G+E	**GxE**			**GxE**	PE=75%
Paire d'environnements						**Simp_env**	**G+E**
AviS	cor	0,90	0,94			39%	4%
AviT	0,83	0,92	0,94			40%	2%
AviS	cor	0,79	0,86			31%	7%
AgaT	0,60	0,83	0,79			35%	-3%
AgaS	cor	0,95	0,92			43%	-3%
AgaT	0,88	0,95	0,92			48%	-3%
Tous les environnements							
AviS		0,84	0,89			34%	5%
AviT		0,83	0,88			35%	5%
AgaS		0,86	0,91			42%	5%
AgaT		0,77	0,89			45%	12% CV2

Annexe 9 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles du caractère fw de la population Magic (orange) et Gwas (bleu), CV2 et PE=75%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions par paire d'environnement et les tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements					***Accuracy***			
Simple environnement	**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%change**			
Avi12	0,47		0,49		-1%			
Avi17	0,31		0,31		0%			
HAvi17	0,27		0,26		1%			
HSMor16	0,17		0,15		1%			
LSMor16	0,22		0,22		0%			
Mor15	0,36		0,38		-1%			
WDMor15	0,41	0,53	0,41	0,53				TP=75%
	G+E	**GxE**			**GxE**			
Paire d'environnements					**Simp_env**	**G+E**		
HAvi17	0,63	0,81			54%	19%		
LSMor16 cor 0,20	0,55	0,67			46%	13%		
Mor15	0,95	0,92			56%	-3%		
WDMor15 cor 0,82	0,95	0,92			52%	-3%		
Tous les environnements								
Avi12	0,84	0,84			37%	0%		
Avi17	0,62	0,60			29%	-2%		
HAvi17	0,50	0,68			41%	18%		
HSMor16	0,52	0,67			50%	15%		
LSMor16	0,45	0,72			51%	28%		
Mor15	0,71	0,76			40%	6%		
WDMor15	0,72	0,82			41%	10%		CV2

Modèles/environnements					***Accuracy***			
Simple environnement	**BL**	**BL_BLUP**	**BayesC**	**BayesC_BLUP**	**%change**			
AviS	0,71		0,70		2%			
AviT	0,74		0,72		2%			
AgaS	0,66		0,67		0%			
AgaT	0,69	0,72	0,68	0,75	2%			PE=75%
	G+E	**GxE**			**GxE**			
Paire d'environnements					**Simp_env**	**G+E**		
AviS	0,94	0,98			27%	4%		
AviT cor 0,93	0,94	0,98			23%	4%		
AgaS	0,93	0,92			25%	-2%		
AgaT cor 0,87	0,96	0,95			26%	-1%		
Tous les environnements								
AviS	0,96	0,95			23%	-2%		
AviT	0,96	0,94			20%	-2%		
AgaS	0,84	0,91			25%	7%		
AgaT	0,95	0,93			24%	-2%		
								CV2

Annexe 10 : Précision de prédiction (*accuracy*) moyenne par environnement suivant les modèles du caractère Leaf de la population Magic (orange), Gwas (bleu), CV2 et PE=75%. %change dans l'approche simple environnement est la différence (en %) entre BL et BayesC, dans les prédictions par paire d'environnement et les tous les environnements confondus, il correspond à la différence (en %) entre le modèle GxE et le modèle BL à gauche et le modèle G+E à droite.

Modèles/environnements				***Accuracy***			
Simple environnement (G)		**BL**		**BayesC**	**%change**		
Avi17		0,27		0,29	-2%		
HAvi17		0,30		0,32	-3%		
HSMor16		0,26		0,23	3%		
LSMor16		0,18		0,19	-1%		
Mor15		0,10		0,10	0%		
WDMor15		0,23	0,33	0,24			PE=75%
		G+E	**GxE**		**GxE**		
Paire d'environnements					**G**	**G+E**	
Avi17		0,77	0,77		50%	-1%	
HAvi17	cor 0,53	0,73	0,84		54%	11%	
HSMor16		0,52	0,73		46%	21%	
WDMor15	cor 0,04	0,55	0,62		39%	7%	
Tous les environnements							
Avi17		0,53	0,69		42%	16%	
HAvi17		0,53	0,68		38%	15%	
HSMor16		0,51	0,69		43%	18%	
LSMor16		0,61	0,70		52%	9%	
Mor15		0,28	0,55		45%	27%	
WDMor15		0,31	0,66		43%	35%	CV2

Modèles/environnements				***Accuracy***			
Simple environnement (G)		**BL**		**BayesC**	**%change**		
AviS		0,44		0,44	0%		
AviT		0,53		0,51	2%		
AgaS		0,36		0,36	-1%		
AgaT		0,35		0,31	4%		
		G+E	**GxE**		**GxE**		PE=75%
Paire d'environnements					**G**	**G+E**	
AviS		0,72	0,77		33%	5%	
AviT	cor 0,55	0,82	0,83		30%	1%	
AviT		0,51	0,73		21%	22%	
AgaS	cor 0,25	0,45	0,71		36%	26%	
Tous les environnements							
AviS		0,75	0,80		36%	5%	
AviT		0,67	0,78		25%	11%	
AgaS		0,45	0,78		43%	34%	
AgaT		0,63	0,78		43%	15%	CV2

Annexe 11 : Comparaison des marqueurs à fort effet par rapport aux QTLs détectés.

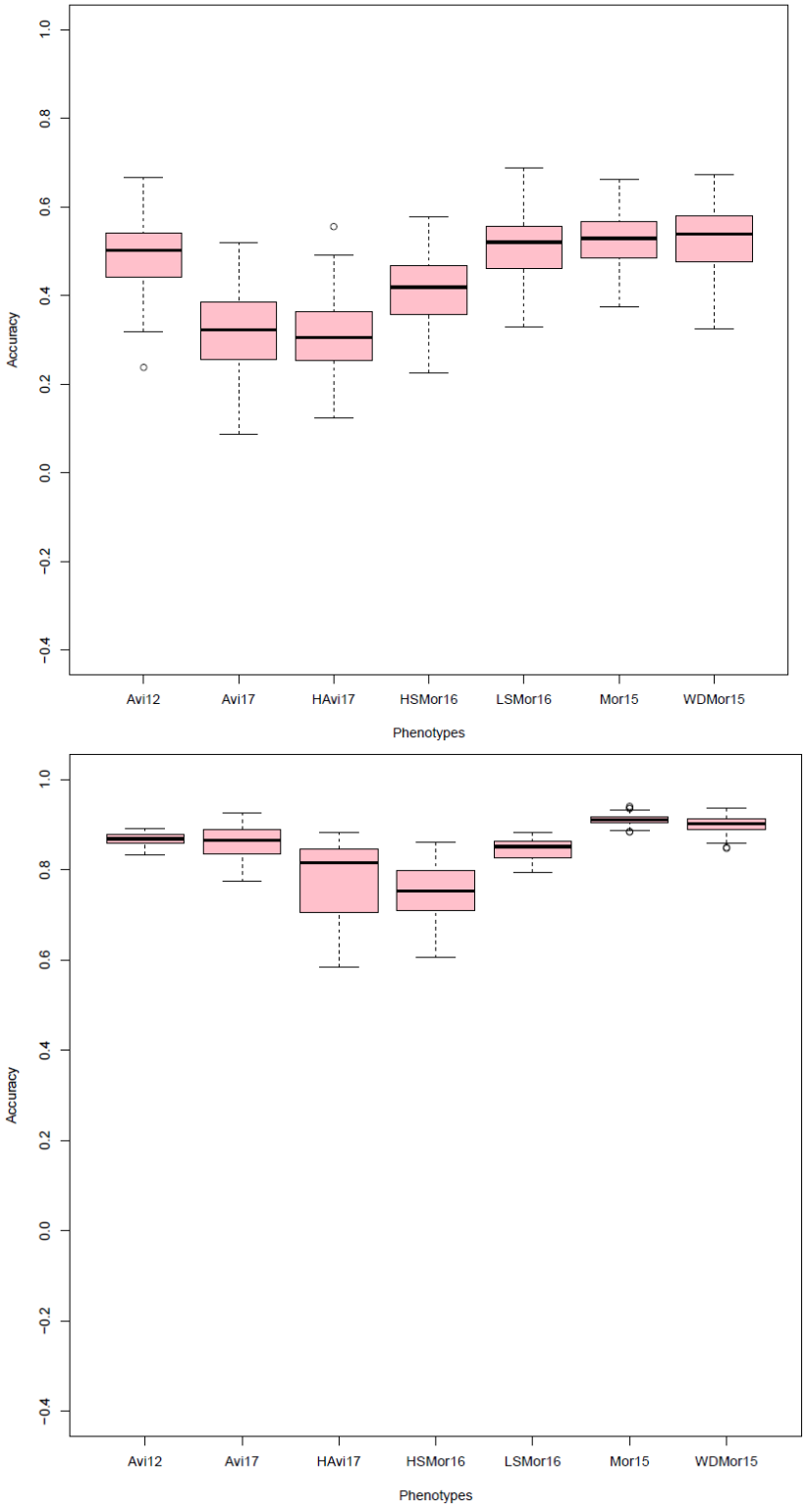
La recherche de correspondance entre les marqueurs à forts effets, issus du modèle GxE de la population GWAS, et les QTLs détectés dans des études précédentes a donné les résultats suivants : En ce qui concerne le caractère flw, 3 QTLs ont été détectés par Albert *et al.,* (2016), dont 3 à Avignon et 2 à Agadir. A Agadir tous les QTLs ont été détectés en présence de stress hydrique, tandis qu'à Avignon, 2 ont été obtenus en stress hydrique et le 3^{e} en condition de contrôle. Pour un seuil fixé à 2, aucun marqueur à effet fort ne se trouve dans les QTLs d'Avignon. Pour flw à Agadir, 9 marqueurs (Tableau ci-dessous) ayant des seuils de LOD supérieurs à 2 se trouvent dans le QTL délimité par les marqueurs S05_03454610 - S05_63392432. Pour le caractère fw, 3 QTLs ont été détectés à Avignon, dont 1 en stress hydrique et les 2 autres en stress hydrique et condition contrôle. A Agadir, 2 QTLS ont été détectés en condition de contrôle. En tout, 9 marqueurs à effets forts se situent dans les QTLs d'Avignon, dont 4 dans le QTL délimité par les marqueurs S02_02553179 - S02_33883746 (stress) et 5 dans le QTL délimité par S02_41013188 - S02_42883046 (stress et contrôle). Cinq marqueurs à effet supérieur à 2 se situent dans le QTL délimité par S02_41013188 - S02_42883046, à Agadir en condition contrôle. En somme, On ne trouve pas de marqueurs à effet fort dans toutes les régions de QTL.

Tableau résumé des marqueurs à effet supérieur à 2 qui se situent dans les QTLs détecté par Albert *et al.,* (2016) pour les différents caractères de la population GWAS.

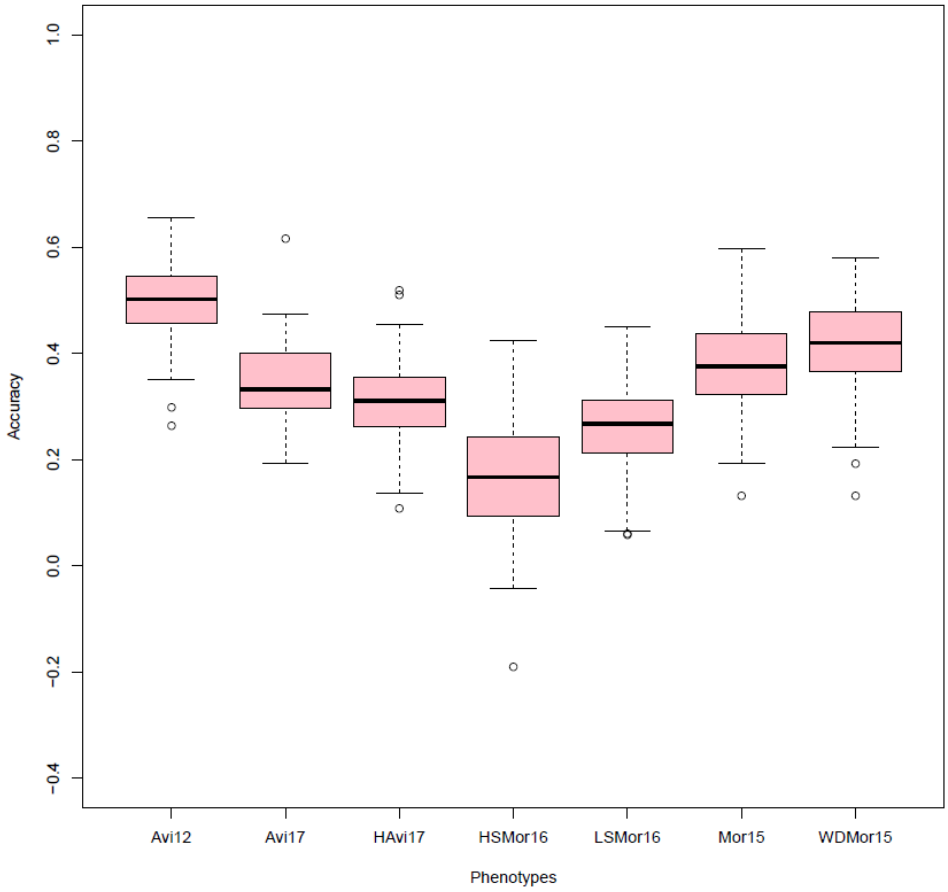
Traits	Conditions	QTLs	Marqueurs (Effet>2)
Flw.Aga	AgaS	S05_03454610-S05_63392432	S05_03764168 ; S05_04165025 ; S05_11527220 ; S05_15172425 ; S05_61072064 ; S05_03670958 ; S05_04066230 ; S05_60282827 ; S05_63309656
Fw.Avi	AviS	S02_02553179 - S02_33883746	S02_22214388 ; S02_24276046 ; S02_29463823 ; S02_30479427
	AviS et AviT	S02_41013188-S02_42883046	S02_41134765 ; S02_41336320 ; S02_41353569 ; S02_42337677 ; S02_42432151
Fw.Aga	AgaT	S02_41013188-S02_42883046	S02_41353569 ; S02_42337677 ; S02_42432151 ; S02_41134765 ; S02_41336320

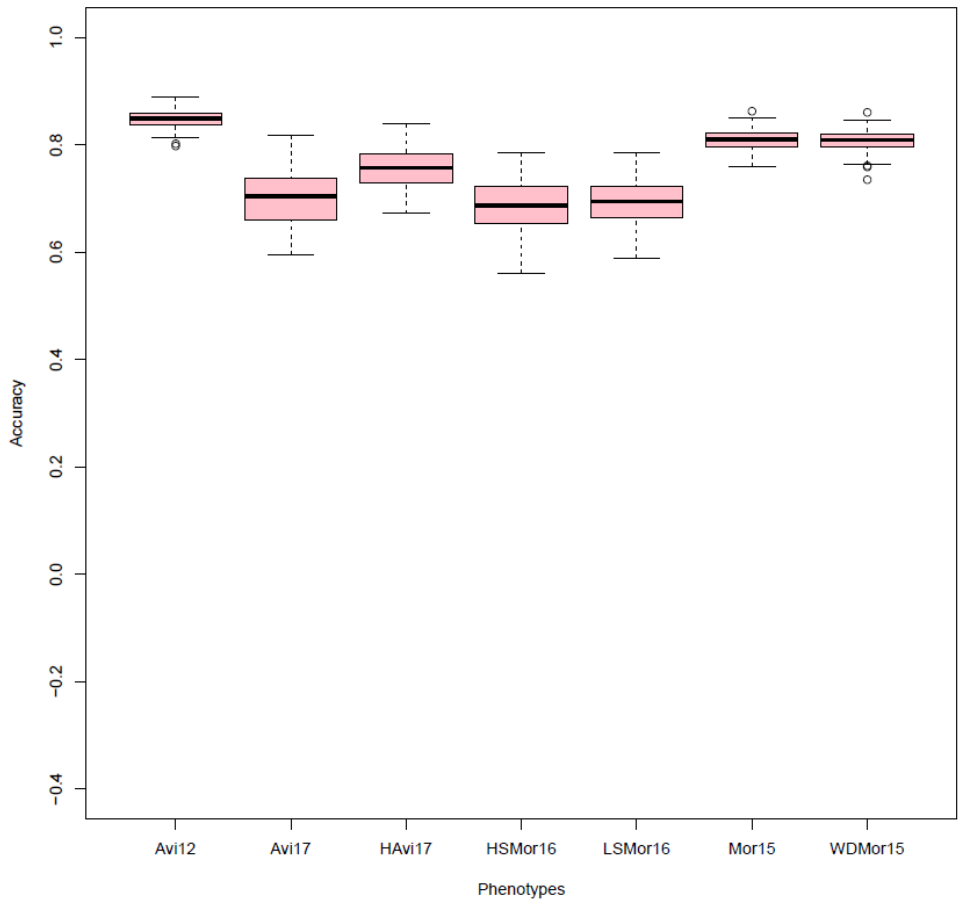
Annexe 12 : Boxplot de la précision de la prédiction (*Accuracy*) des caractères flw, fw et leaf de la population Magic, obtenue avec les modèles GxE + cofacteurs dans CV1 et CV2. Ils ont rangé comme suit : A gauche, précision de prédiction des phénotypes en CV1, modèle = RKHS et à droite, précision de prédiction des phénotypes en CV2, modèle = RKHS (entête). En abscisses, phénotypes dans les différents environnements et *Accuracy* en ordonnées.

FLW

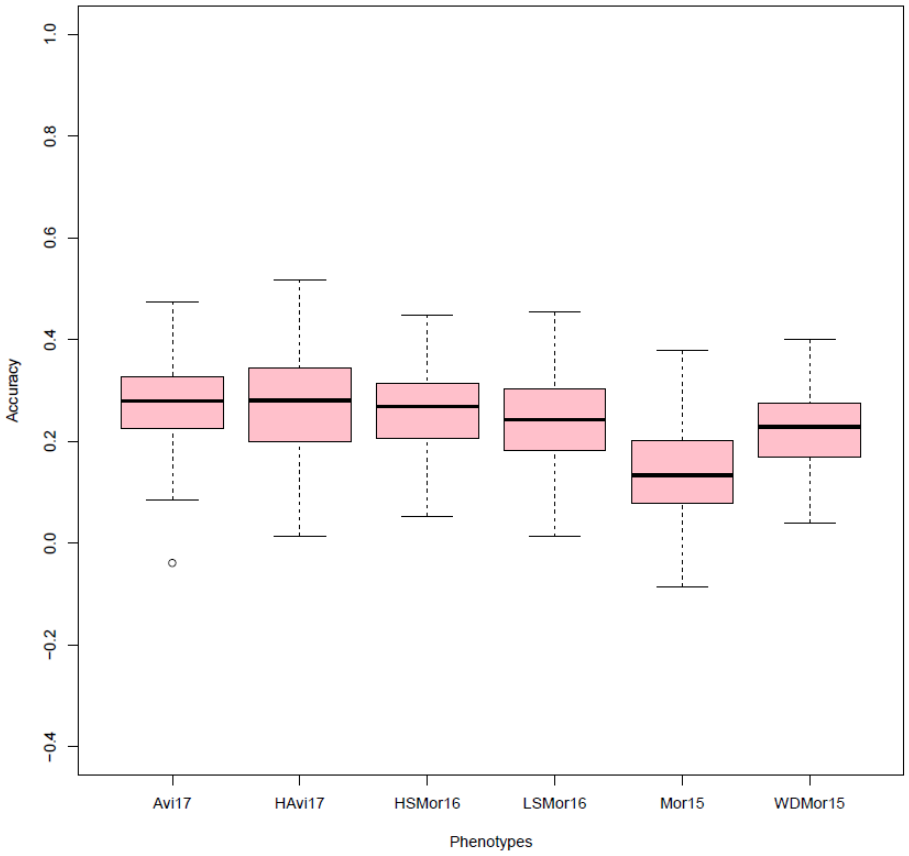


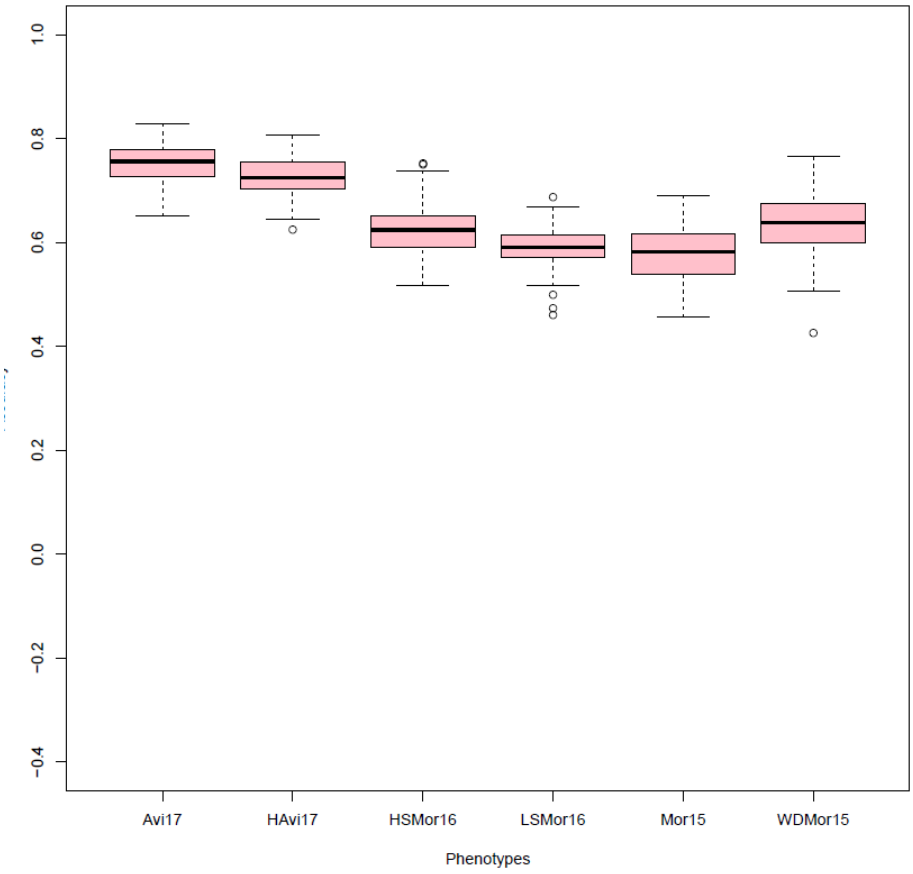
FW


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Avi12
Avi17
HAvi17
HSMor16
LSMor16
Mor15
WDMor15
Phenotypes


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Avi12
Avi17
HAvi17
HSMor16
LSMor16
Mor15
WDMor15
Phenotypes

LEAF


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Avi17
HAvi17
HSMor16
LSMor16
Mor15
WDMor15
Phenotypes


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Avi17
HAvi17
HSMor16
LSMor16
Mor15
WDMor15
Phenotypes

Annexe 13 : Script utilisé pour la prédiction du caractère flw de la population Magic avec le modèle Simple environnement (modèle BL, PE=75%).

```
## Packages
install.packages("pastecs",lib="~/local/R_libs/")
install.packages("BGLR",lib="~/local/R_libs/")
install.packages("corrplot",lib="~/local/R_libs/")

pheno<-read.table(file="Magic_flw_pheno_GxE_1805.csv",sep=";",dec = ".",header=T)
str(pheno)
dim(pheno)
head(pheno)

#Analyse descriptive
pheno2<-data.frame(pheno[,-1],row.names = NULL,check.rows = FALSE, check.names = TRUE)
dim(pheno2)
str(pheno2)

library(pastecs,lib="~/local/R_libs/")
stat.desc(pheno2, norm = TRUE)

#Representation graphique
variables <- colnames(pheno2)[c(1:7):ncol(pheno2)]
variables
par(mfrow = c(2,4))
for (i in 1:length(variables)) {
boxplot(pheno2[, variables[i]], las = 2, cex.axis = 0.2, main = variables[i])
}
par(mfrow = c(1,1))
boxplot(pheno2)

#### Mettre les matrices phéno en numéric
matrice1<-matrix(nrow=250, ncol=1)
for (i in c(2:8)){
a<-as.numeric(pheno[,i])
matrice1<-cbind(matrice1,a)
}
dim(matrice1)
head(matrice1)

#### Enlever les NA de la matrice des phéno : remplacer NA par la valeur de
#la moyenne du phénotype
matrice<-matrix(nrow=250, ncol=1)
for (i in c(2:8)){
mean<-mean(na.omit(matrice1[,i]))
remplace<-replace(matrice1[,i],is.na(matrice1[,i]),mean)
matrice<-cbind(matrice,remplace)
}
#Renommer les colonnes avec le nom des marqueurs
noms<-colnames(pheno)
colnames(matrice)<-noms
pheno3<-matrice[,2:8]

#Renommer les lignes avec le nom des individus en rownames
rownames(pheno)<-pheno[,1]
```

```
rownames(pheno3)<-rownames(pheno)
dim(pheno3)
head(pheno3)

##correlation
cormat<-signif(cor(pheno3),2)
head(cormat)
library(corrplot,lib="~/local/R_libs/")
corrplot(cormat, type="upper", order="hclust", tl.col="black", tl.srt=45)

#Mettre les donnÃ©es dans le format de fichier pour le package BGLR

geno <- as.matrix(read.csv("magic_geno_1805.csv",header=T,sep=";"))
dim(geno)
str(geno)
head(geno)

#### Mettre les matrices geno en numéric
matrice1<-matrix(nrow=250, ncol=1)
for (i in c(2:1346)){
a<-as.numeric(geno[,i])
matrice1<-cbind(matrice1,a)
}
dim(matrice1)

#### Enlever les NA de la matrice des marqueurs : remplacer NA par la valeur de
#l'allèle le plus fréquent
matrice<-matrix(nrow=250, ncol=1)
for (i in c(2:1346)){
table<-table(matrice1[,i])
table<-as.matrix(table)
b<-as.vector(table[1,])
c<-as.vector(table[2,])
if (b>c) a=0
if (b==c) a=0
if (c>b) a=2
alelle_le_plus_frequent<-a
remplace<-replace(matrice1[,i],is.na(matrice1[,i]),alelle_le_plus_frequent)
matrice<-cbind(matrice,remplace)
}
#Renommer les colonnes avec le nom des marqueurs
noms<-colnames(geno)
colnames(matrice)<-noms
geno2<-matrice[,2:1346]

#Renommer les lignes avec le nom des individus en rownames
rownames(geno)<-geno[,1]
rownames(geno2)<-rownames(geno)
dim(geno2)

##############################################################################
########################  Package BGLR ###################################
##############################################################################

library("BGLR",lib="~/local/R_libs/")
```

```
pheno3<-scale(pheno3,center = TRUE,scale = TRUE)
rownames(geno2)
rownames(pheno3)
dim(geno2)
dim(pheno3)

################################## TP = 75 % ##############################
traits=7
cycles=100
accuracy75 = matrix(nrow=cycles, ncol=traits)
effectsmarkers = matrix(nrow=1345, ncol=traits)

for(n in c(1:traits)){

bHat_pheno<-c(1:1345)
for(r in c(1: cycles)){

#1# Choose a Testing set
y<-pheno3[,n]
yNA<-y
tst<-sample(1:250,size=63,replace=FALSE)
yNA[tst]<-NA

#2# Setting the linear predictor
ETA<-list(list(X=geno2[c(1:250),], model='BL'))

#3# Fitting the model
fm<-BGLR(y=yNA,ETA=ETA, nIter=1200, burnIn=200, verbose=FALSE)
yHat<-fm$yHat[tst]

#4#Accuracy
accuracy75[r, n] <- cor(fm$yHat[tst],y[tst])

#5# Predictions markers effect
bHat<- fm$ETA[[1]]$b
bHat_pheno<-cbind(bHat_pheno,bHat)
dim(bHat_pheno)
head(bHat_pheno)
}
bHat_pheno<-bHat_pheno[,-1]
for(i in c(1:1345)){
mean<-mean(bHat_pheno[i,])
effectsmarkers [i, n] <- mean
}
}

colnames(accuracy75)<-colnames(pheno3)

# Effet des marqueurs (M en lignes ; phénos en colonne)
rownames(effectsmarkers)<-colnames(geno2)
head(effectsmarkers)
dim(effectsmarkers)
colnames(effectsmarkers)<-colnames(pheno3)
phenotypes<-colnames(effectsmarkers)
```

```
#creer un fichier de donnees des effets des marqueurs et l'accuracy

write.table(effectsmarkers, "effectsmarkers_TP75_flw.csv", row.names=TRUE, sep=";",dec=".",
na=" ")
write.table(accuracy75, "accuracy75_BGLR_BL_TP=75_flw.csv",row.names=TRUE,
sep=";",dec=".", na=" ")

##Representation graphique

# Boucle pour obtenir les graphiques des effets des marqueurs par phenotype
pdf("effets_marqueurs_TP75_flw.pdf", height=10,width=10)
for(i in c(1:traits)){
plot(effectsmarkers[,i]^2,ylab='Estimated Squared-Marker Effect',
type='o',cex=.5,col=4,main=phenotypes[i])
}
dev.off()

#Boxplot en pdf
pdf("graphe_TP=75_flw_BGLR_BL.pdf", height=10,width=10)
boxplot(accuracy75, xlab="Phenotypes", ylab="Accuracy", col="pink", ylim=c(-0.4,1),
main ="Precision de la prediction des Phenotypes
(TP=75%, modele=BL, BGLR package)")
dev.off()

# Effet estimé des marqueurs
bHat<- fm$ETA[[1]]$b
SD.bHat<- fm$ETA[[1]]$SD.b
plot(bHat^2, ylab='Estimated Squared-Marker Effect',
type='o',cex=.5,col=4,main='Marker Effects')

#2# Predictions
# Total prediction
yHat<-fm$yHat
tmp<-range(c(y,yHat))
plot(yHat~y,xlab='Observed',ylab='Predicted',col=2, xlim=tmp,ylim=tmp, main="TP=75%");
abline(a=0,b=1,col=4,lwd=2)
```

Annexe 14 : Script utilisé pour la prédiction du caractère flw de la population Magic avec le modèle GxE (modèle RKHS, CV1 et CV2).

```
## Packages
install.packages("BGLR",lib="~/local/R_libs/")
install.packages("apercu",lib="~/local/R_libs/")
install.packages("corpcor",lib="~/local/R_libs/")
library("apercu",lib="~/local/R_libs/")
library("corpcor",lib="~/local/R_libs/")

pheno<-read.table(file="Magic_flw_pheno_GxE_1805.csv",sep=";",dec = ".",header=T)
str(pheno)
dim(pheno)
head(pheno)

#### Mettre les matrices phéno en numéric
matrice1<-matrix(nrow=250, ncol=1)
for (i in c(2:8)){
a<-as.numeric(pheno[,i])
matrice1<-cbind(matrice1,a)
}
dim(matrice1)
head(matrice1)

#### Enlever les NA de la matrice des phéno : remplacer NA par la valeur de
#la moyenne du phénotype
matrice<-matrix(nrow=250, ncol=1)
for (i in c(2:8)){
mean<-mean(na.omit(matrice1[,i]))
remplace<-replace(matrice1[,i],is.na(matrice1[,i]),mean)
matrice<-cbind(matrice,remplace)
}
#Renommer les colonnes avec le nom des marqueurs
noms<-colnames(pheno)
colnames(matrice)<-noms
pheno2<-matrice[,2:8]

#Renommer les lignes avec le nom des individus en rownames
rownames(pheno)<-pheno[,1]
rownames(pheno2)<-rownames(pheno)
dim(pheno2)
head(pheno2)

#Mettre les donnÃ©es dans le format de fichier pour le package BGLR

geno <- as.matrix(read.csv("magic_geno_1805.csv",header=T,sep=";"))
dim(geno)
str(geno)
head(geno)

#### Mettre les matrices geno en numéric
matrice1<-matrix(nrow=250, ncol=1)
for (i in c(2:1346)){
a<-as.numeric(geno[,i])
matrice1<-cbind(matrice1,a)
```

```
}
dim(matrice1)

#### Enlever les NA de la matrice des marqueurs : remplacer NA par la valeur de
#l'allèle le plus fréquent
matrice<-matrix(nrow=250, ncol=1)
for (i in c(2:1346)){
table<-table(matrice1[,i])
table<-as.matrix(table)
b<-as.vector(table[1,])
c<-as.vector(table[2,])
if (b>c) a=0
if (b==c) a=0
if (c>b) a=2
alelle_le_plus_frequent<-a
remplace<-replace(matrice1[,i],is.na(matrice1[,i]),alelle_le_plus_frequent)
matrice<-cbind(matrice,remplace)
}
#Renommer les colonnes avec le nom des marqueurs
noms<-colnames(geno)
colnames(matrice)<-noms
geno2<-matrice[,2:1346]

#Renommer les lignes avec le nom des individus en rownames
rownames(geno)<-geno[,1]
rownames(geno2)<-rownames(geno)
dim(geno2)

###KINSHIP####
#standardisation (centage-réduction) de la matrice de génotypage

p <- colMeans(geno2) / 2
q <- 1 - p
geno3 <- scale(geno2, center = 2 * p, scale = sqrt(2 * p * q))
ap(geno3)

#calcul de la matrice d'apparentement

K <- tcrossprod(geno3) / ncol(geno3)
ap(K)

#vérification que la matrice est définie positive

is.positive.definite(K)

# si FALSE
G <- make.positive.definite(K)
is.positive.definite(G)
ap(G)

##############################################################################
########################  Package BGLR ######################################
##############################################################################

library("BGLR",lib="~/local/R_libs/")
```

```
pheno3<-scale(pheno2,center = TRUE,scale = TRUE)
rownames(G)
rownames(pheno3)
dim(G)
dim(pheno3)

############################################################
########### CV1 ############################################
############################################################
env <- c(1:7) # choose any set of environments from 1:ncol(Y)
nEnv <- length(env)
cycles=100
accuracycv1 = matrix(nrow=cycles, ncol=nEnv,NA)
effectsmarkers = matrix(nrow=1345, ncol=nEnv)

for(m in c(1:nEnv)){

bHat_pheno<-c(1:1345)
for(r in c(1: cycles)){

 #1# choose testing set
 Y <- pheno3[,env]
 n <- nrow(Y)
 percTST<-0.3
 nTST <- round(percTST*n)
 tst<-sample(1:n,size=nTST,replace=FALSE)
 YNA <- Y
 YNA[tst,]<-NA

 YNA <- as.vector(YNA)

  #2# Fixed effect
 envID <- rep(env,each=nrow(Y))
 ETA <- list(list(~factor(envID)-1,model="FIXED"))

  #3# Main effects of markers
 G0 <- kronecker(matrix(nrow=nEnv,ncol=nEnv,1),G)
 ETA[[2]] <- list(K=G0,model='RKHS')

 #4# Adding interaction terms
 for(i in 1:nEnv){
   tmp <- rep(0,nEnv) ; tmp[i] <- 1
   G1 <- kronecker(diag(tmp),G)
   ETA[[(i+2)]] <- list(K=G1, model='RKHS')
 }
 #5# Model Fitting
 fm <- BGLR(y=YNA,ETA=ETA,nIter=1200,burnIn=200,verbose = FALSE)
 YHatInt <- matrix(fm$yHat,ncol=nEnv)
 #6#Accuracy
 accuracycv1[r, m] <- cor(Y[tst,m],YHatInt[tst,m])

 #7# Predictions markers effect
 bHat<- fm$ETA[[2]]$u
 bHat_pheno<-cbind(bHat_pheno,bHat)
```

```
  dim(bHat_pheno)
  head(bHat_pheno)
  }
  bHat_pheno<-bHat_pheno[,-1]
  for(i in c(1:1345)){
  mean<-mean(bHat_pheno[i,])
  effectsmarkers [i, m] <- mean
  }
}

colnames(accuracycv1)<-colnames(Y)

# Effet des marqueurs (M en lignes ; phénos en colonne)
rownames(effectsmarkers)<-colnames(geno3)
head(effectsmarkers)
dim(effectsmarkers)
colnames(effectsmarkers)<-colnames(Y)
phenotypes<-colnames(effectsmarkers)

#créer un fichier de données des effets des marqueurs et l'accuracy

write.table(effectsmarkers, "effectsmarkers_cv1_flw.csv", row.names=TRUE, sep=";",dec=".",
na=" ")
write.table(accuracycv1, "accuracyCV1_BGLR_GXE_flw.csv",row.names=TRUE,
sep=";",dec=".", na=" ")


##Représentation graphique

# Boucle pour obtenir les graphiques des effets des marqueurs moyens par phenotype
pdf("effets_marqueurs_CV1_flw.pdf", height=10,width=10)
for(i in c(1:nEnv)){
plot(effectsmarkers[,i]^2,ylab='Estimated Squared-Marker Effect',
type='o',cex=.5,col=4,main=phenotypes[i])
}
dev.off()

#Boxplot en pdf
pdf("graphe_cv1_flw_BGLR_GxE.pdf", height=10,width=10)
boxplot(accuracycv1, xlab="Phenotypes", ylab="Accuracy", col="pink", ylim=c(-0.4,1),
main ="Precision de la prediction des Phenotypes
(CV1, modele=RKHS, BGLR package)")
dev.off()

# Effet estimé des marqueurs
bHat<- fm$ETA[[2]]$u
SD.bHat<- fm$ETA[[2]]$SD.u
plot(bHat^2, ylab='Estimated Squared-Marker Effect',
type='o',cex=.5,col=4,main='Marker Effects')

# Total prediction
tmp<-range(c(Y,YHatInt))
plot(YHatInt~Y,xlab='Observed',ylab='Predicted',col=2, xlim=tmp,ylim=tmp, main="CV1");
abline(a=0,b=1,col=4,lwd=2)
```

```
##############################################################
########### CV2  ###############################################
##############################################################
env <- c(1:7) # choose any set of environments from 1:ncol(Y)
nEnv <- length(env)
cycles=100
accuracycv2 = matrix(nrow=cycles, ncol=nEnv,NA)
effectsmarkers = matrix(nrow=1345, ncol=nEnv)

for(m in c(1:nEnv)){

 bHat_pheno<-c(1:1345)
 for(r in c(1: cycles)){

  #1# choose testing set
  Y <- pheno3[,env]
  n <- nrow(Y)
  percTST<-0.3
  nTST <- round(percTST*n)
  nNA <- nEnv*nTST
  if(nNA<n){ indexNA <- sample(1:n,nNA,replace=FALSE) }
  if(nNA>=n){
   nRep <- floor(nNA/n)
   remain <- sample(1:n,nNA%%n,replace=FALSE)
   a0 <- sample(1:n,n,replace=FALSE)
   indexNA <- rep(a0,nRep)
   if(length(remain)>0){
    a1 <- floor(length(indexNA)/nTST)*nTST
    a2 <- nNA - a1 - length(remain)
    bb <- sample(a0[!a0%in%remain],a2,replace=FALSE)
    noInIndexNA <- c(rep(a0,nRep-1),a0[!a0%in%bb])
    indexNA <- c(noInIndexNA,bb,remain)
   }
  }
  indexEnv <- rep(1:nEnv,each=nTST)
  YNA <- Y
  for(j in 1:nEnv) YNA[indexNA[indexEnv==j],j] <- NA

  YNA <- as.vector(YNA)

  #2# Fixed effect
  envID <- rep(env,each=nrow(Y))
  ETA <- list(list(~factor(envID)-1,model="FIXED"))

  #3# Main effects of markers
  G0 <- kronecker(matrix(nrow=nEnv,ncol=nEnv,1),G)
  ETA[[2]] <- list(K=G0,model='RKHS')

  #4# Adding interaction terms
  for(i in 1:nEnv){
   tmp <- rep(0,nEnv) ; tmp[i] <- 1
   G1 <- kronecker(diag(tmp),G)
   ETA[[(i+2)]] <- list(K=G1, model='RKHS')
  }
  #5# Model Fitting
```

```
    fm <- BGLR(y=YNA,ETA=ETA,nIter=1200,burnIn=200,verbose = FALSE)
    YHatInt <- matrix(fm$yHat,ncol=nEnv)

    #6# Accuracy
    accuracycv2[r, m] <- cor(Y[tst,m],YHatInt[tst,m])

    #7# Predictions markers effect
    bHat<- fm$ETA[[2]]$u
    bHat_pheno<-cbind(bHat_pheno,bHat)
    dim(bHat_pheno)
    head(bHat_pheno)

  }
  bHat_pheno<-bHat_pheno[,-1]
  for(i in c(1:1345)){
    mean<-mean(bHat_pheno[i,])
    effectsmarkers [i, m] <- mean
  }
}

colnames(accuracycv2)<-colnames(Y)

# Effet des marqueurs (M en lignes ; phénos en colonne)
rownames(effectsmarkers)<-colnames(geno3)
head(effectsmarkers)
dim(effectsmarkers)
colnames(effectsmarkers)<-colnames(Y)
phenotypes<-colnames(effectsmarkers)

#creer un fichier de donnees des effets des marqueurs et l'accuracy

write.table(effectsmarkers, "effectsmarkers_cv2_flw.csv", row.names=TRUE, sep=";",dec=".",
na=" ")

write.table(accuracycv2, "accuracyCV2_BGLR_GXE_flw.csv",row.names=TRUE,
sep=";",dec=".", na=" ")


##Representation graphique

# Boucle pour obtenir les graphiques des effets des marqueurs moyens par phenotype
pdf("effets_marqueurs_CV2_flw.pdf", height=10,width=10)
for(i in c(1:nEnv)){
  plot(effectsmarkers[,i]^2,ylab='Estimated Squared-Marker Effect',
       type='o',cex=.5,col=4,main=phenotypes[i])
}
dev.off()

#Boxplot en pdf
pdf("graphe_CV2_flw_BGLR_GxE.pdf", height=10,width=10)
boxplot(accuracycv2, xlab="Phenotypes", ylab="Accuracy", col="pink", ylim=c(-0.4,1),
        main ="Precision de la prediction des Phenotypes
(CV2, modele=RKHS, BGLR package)")
dev.off()
```

```
# Effet estimé des marqueurs
bHat<- fm$ETA[[2]]$u
SD.bHat<- fm$ETA[[2]]$SD.u
plot(bHat^2, ylab='Estimated Squared-Marker Effect',
    type='o',cex=.5,col=4,main='Marker Effects')

# Total prediction
tmp<-range(c(Y,YHatInt))
plot(YHatInt~Y,xlab='Observed',ylab='Predicted',col=2, xlim=tmp,ylim=tmp, main="CV2");
abline(a=0,b=1,col=4,lwd=2)
```

RESUME

La sélection génomique est une alternative permettant la réduction des coûts de la sélection et un progrès génétique rapide. Les études précédentes de sélection génomique sur la tomate ont permis d'évaluer l'impact des facteurs de la prédiction génomique tels que le déséquilibre de liaison et la taille de la population d'entrainement sur la précision de la prédiction. Ces évaluations ont été faites à l'aide de modèles simples environnements qui n'intègrent pas les interactions GxE.

Dans notre étude, nous avons évalué l'impact de l'intégration de l'interaction GxE et de cofacteurs environnementaux sur la précision de la prédiction. Trois modèles ont été comparés, simple environnement, G+E et GxE, pour évaluer la précision de prédiction des jeux de données de deux populations, MAGIC et GWAS, génotypées et phénotypées pour plusieurs caractères dans différents environnements plus ou moins stressés.

Cette étude a révèlé que la précision de prédiction des modèles multi-environnements est meilleure à celle du modèle simple environnement. Le modèle GxE a une précision de prédiction aussi élevée, voire meilleure que celle de G+E, mais avec des différences suivant les environnements ou la composition de la population d'entrainement. Ces deux modèles ont une précision de prédiction très supérieure à celle du modèle simple environnement (G) lorsque les individus sont testés dans au moins un environnement (CV2). Le gain est plus faible pour prédire des individus testés dans aucun environnement (CV1). La précision de prédiction de ces différents modèles est fortement corrélée à l'héritabilité du caractère. Elle augmente quand les environnements sont corrélés entre eux. L'intégration des cofacteurs environnementaux au modèle GxE, améliore la précision de prédiction. En condition limitée, l'échantillonnage CV2 est le plus approprié pour faire la prédiction génomique avec des modèles multi-environnements.

Mots clés: *Solanum* lycopersicum, Prédiction génomique, interaction GxE, cofacteurs environnementaux, validation croisée.

ABSTRACT

Genomic selection is an alternative for the reduction of selection costs and rapid genetic progress in plant breeding. Previous genomic selection studies on tomato have evaluated the impact of several factors such as linkage disequilibrium and training population size on prediction *accuracy*. These evaluations were done using simple environment models that do not consider GxE interactions.

In our study, we have evaluated the impact of the integration of GxE interaction and environmental cofactors on prediction *accuracy*. The simple environment, G+E and GxE models were used to evaluate the prediction *accuracy* in two population datasets (MAGIC and GWAS) that have been genotyped and phenotyped for several traits in different environments.

This study revealed that the prediction *accuracy* of multi-environment models is better than that of

the single environment model. The GxE model has a prediction *accuracy* as high as that of G+E, and even predicted better than G+E depending on the environment or the composition of the training population. Both models have a higher prediction *accuracy* than the single environment model (G) when accessions are tested in at least one environment (CV2 sampling) compared to models where some accessions are tested in none of the environments (CV1). The prediction *accuracy* of these different models was strongly correlated to the heritability of the traits. It increased when the environments were correlated with each other. The integration of environmental co-factors in the GxE model improved prediction *accuracy*. In limited conditions, CV2 sampling is the most appropriate to make genomic predictions with multi-environment models.
Key words: *Solanum lycopersicum*, Genomic prediction, GxE interaction, environmental cofactors, cross validation.